Michal Kebísek

Data Mining in the Industry

Scientific Monographs in Automation and Computer Science

Edited by
Prof. Dr. Peter Husar (Ilmenau University of Technology) and
Dr. Kvetoslava Resetova (Slovak University of Technology in Bratislava)

Vol. 6

DATA MINING IN THE INDUSTRY

Michal Kebísek

Universitätsverlag Ilmenau
2012

Impressum

Bibliographic information of the German National Library
The German National Library lists this publication in the German national bibliography, with detailed bibliographic information on the Internet at http://dnb.d-nb.de.

Author’s acknowledgement to Gabriela Chmelíková for translation.

This scientific monograph originated from the author's dissertation thesis defended at the Slovak University of Technology in Bratislava, Faculty of Materials Science and Technology in Trnava.

Reviewers:
Prof. Ing. Juraj Spalek, PhD.
Prof. Dr. Peter Husar
Doc. Ing. Anton Kachaňák, CSc.

Author’s contact address:
Ing. Michal Kebísek, PhD.
Slovak University of Technology in Bratislava
Faculty of Materials Science and Technology in Trnava

Ilmenau Technical University / University Library
Universitätsverlag Ilmenau
Postfach 10 05 65
98684 Ilmenau
www.tu-ilmenau.de/universitaetsverlag

Production and delivery
Verlagshaus Monsenstein und Vannerdat OHG
Am Hawerkamp 31
48155 Münster
www.mv-verlag.de

ISSN 2193-6439 (Print)
ISBN 978-3-86360-048-8 (Print)
URN urn:nbn:de:gbv:ilm1-2012100209

Titelfoto: photocase.com

Abstract

The monograph proposes a suitable process application for a knowledge discovery process in industry databases. The entire process was divided into distinct stages. In Stage 1 the subject matter to be resolved by employing the knowledge discovery process was identified. Then the basic problems encountered in using the knowledge discovery process application in industry databases were identified. In Stage 2 the data of the production system is analysed using the STATISTICA Data Miner KDD tool. Several mining models, in which various methods and techniques of data mining in dependence on analyzed data and subject matter investigated, were developed. In order to examine how interesting and useful the knowledge discovered was, it was applied to a production system, whose data operated as input data to the process of knowledge discovery in databases. The knowledge discovered was applied to a simulation model of the production system. Then the application of the knowledge discovered was compared to the results achieved by the production system before the knowledge application was conducted. The results achieved proved that the knowledge discovered was useful and a modified simulation model achieved the predicted behaviour.

Finally, the proposal of the process application methodology of knowledge discovery in industry databases is discussed. This methodology describes the particular steps of implementing the process of knowledge discovery in databases. The proposed methodology can help identify specific requirements and potential problems in the process stages that might be encountered in the course of its application in the industry.

The main contributions of the monograph can be summarized as follows:

- The identification of basic resolvable issues via knowledge discovery in industry databases.
- The application of the process of knowledge discovery in databases to production processes and new knowledge discovery in the process analyzed, and the evaluation and verification of the knowledge discovered and its subsequent application to the production process.
- The proposal of data mining process methodology in industry to improve its control.

Key words

Knowledge Discovery in Databases, Data Mining, Relational Database, Data Warehouse, Production System, Simulation model

LIST OF ABBREVIATIONS

CIM	Computer Integrated Manufacturing
DM	Data Mining
DW	Data Warehousing
GUHA	General Unary Hypotheses Automaton
KDD	Knowledge Discovery in Databases
NC/CNC	Numerical Control / Computer Numerical Controlled
MES	Manufacturing Execution System
MRP	Manufacturing Resource Planning
ODBC	Open Database Connectivity
OPC	OLE for Process Control
PLC	Programmable Logic Controller
PMML	Predictive Model Markup Language
RDBMS	Relational Database Management System
RFID	Radio Frequency Identification
SCADA	Supervisor Control And Data Acquisition
SQL	Structured Query Language
SVM	Support Vector Machines
UML	Unified Modelling Language

INTRODUCTION

As information technology is applied to more and more aspects of human life, produced and saved data grows proportionally. In the day-to-day operation of state administrations, offices, schools, hospitals, retail outlets, data is generated and saved. Similarly, industrial organizations increasingly save customer and supplier data. They administer orders, receipt cards, invoices and attempt to save as much data on the production process as is possible. Most organizations store data in their databases. However, they need access to useful information. The idea of an information society, or at the very least, the utilization of its strategic power found in data sources, requires not only new tools but also new way of thinking. The subject matter lies not only in the elaboration of new models. It is also about the acquisition of information on objects, their behaviour, needs, covered relations, etc.

Market analysis, company analyses, risk management and uncovering fraud, are other examples of the fields of control that necessitate complete knowledge of the people carrying out the activities. Of relevance here is knowledge control or about knowledge systems control and the process whereby responsible individuals become knowledge operators. The process of knowledge discovery in databases, often referred to as data mining, represents the first important step in knowledge control technology.

Company data operation can be divided into several fields that are also related to the possibility of specific technology utilization. In Stage 1 it is necessary to manage the administration of company data, as well as the efficiency and speed of its processing. Then it is necessary to convert the data into a useable format for analyses and presentations, i.e. gather a lot of information, clean and transform it. Of importance is here is the decision of the location of the data and whether we have to prepare, for example, a data warehouse. The final layer needs to make the data available through developing derived information. In this layer, called data mining, data analysis occurs and the information is presented in the most suitable format for the end user.

The discovery of Keppler's laws– the laws of planetary motion – is a classical example of utilizing the process of knowledge discovery in data. Johannes Keppler processed data on planets motion collected for years into a simple model that defined their orbits and thus, he derived the laws of planetary motion (16).

This monograph deals with potential applications of knowledge discovery in databases within a production process. It firstly contains a proposal of a production line and relational database designed to save production process data. Secondly it discusses the application of the process of knowledge discovery in databases to the aforementioned proposed production process. Next, it focuses on the production system problems and their solution via the data mining and data analysis that was obtained from the proposed production system via a selected KDD (Knowledge Discovery in Databases) tool, and on the application of discovered knowledge to the proposed production system. Subsequently, the knowledge is generalized to a methodology that provides a complex description of data mining application in industry.

1. OVERVIEW OF CURRENT STATE IN THE FIELD

1.1 Knowledge Discovery in Databases

Nowadays, powerful hardware combined with database systems permits the saving a lot of data. These systems themselves provide no means to describe the data. Common relational databases assume an ad hoc approach to acquiring information in databases, and therefore, they are not suitable for the analysis of large amounts of automatically obtained data. It is not possible to process such accumulated data by analytical methods like selection query or table processors. Statistical programs can provide a general overview of the database content but only with small scale information, e.g. medium values, variances, linear relations among variables (linear regression), etc.

Information saved in database systems can be classified in two categories. The first category comprises the information for which the database has been constructed and operated. Useful information is found in the specific values of attributes and their combinations because of the connections determined by the database arrangement. The user manipulates the queries via query language, e.g. via SQL. The query results in a set of database cases, each of which is a representation of a real case from the external world. If the database is a true representation of reality, then the query results are true. Through these deductive approach methods to databases it is only possible to derive the logical consequences of information saved in a specific database. The second category describes the information that initially is not obvious from its attribute values or from the database structure. The information is concealed in a large amount of database records and can be discovered using the method of inductive derivation from the attribute's values. There are searchable regularities in databases, i.e. values combinations of certain attributes that are common and typical for the subsets (classes) of facts saved in the database. The results are mostly presented in the form of logical formulations that express the decomposition conditions of saved data into particular classes. The information does not only have a descriptive character, it also expresses specific knowledge in related domains and can comprise elements of prognosis. The results obtained from current data can only predict something about future data by being entered into the database. Derived knowledge is not necessarily true in the real world. Indeed, each result formulation has a probability of mostly less than 1 (32).

Methods of deductive database utilization are an aspect of every database control system. In contrast, the producers of database control systems do not supply the means for inductive operations with data. There are exceptions to this, e.g. Oracle and Microsoft SQL Server.

With respect to the requisite knowledge, in cases the following triangle scheme is introduced. Data are understood as random groups of simple facts or events with more complex facts originating in the aggregation of simple factual knowledge that represents the information. The knowledge is the result of our perception processes and it is organized so that the conclusions can be derived from them. Common sense or good judgements are generally accepted as wisdom (50).

The need to find all relevant information in a database and the insufficiency of existing methods led in late 1980s to the development of a brand new discipline – knowledge discovery in databases. Besides this terminology, several other terms are used: information harvesting, data archaeology, knowledge extraction or data mining (13), (31).

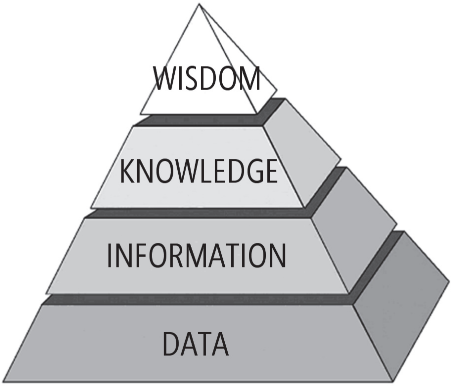


***Fig. 1** From data to wisdom*

The term *knowledge discovery in databases* (KDD) was first used at the 1st Conference on KDD in 1989 to emphasize that knowledge is the final product. In cases we can use the term *data mining* (DM) as a synonym. In compliance with the 1st International Conference on KDD in Montreal in 1995 it is recommended that the term is only used in the stage of discovery within the knowledge discovery in databases (8).

The term data mining originated in the database community and indicates for the most part, the techniques of discovering various patterns in databases, whereas KDD was developed in the field of artificial intelligence. It is thus a more general term that includes the following: preparation stages of data selection and pre-processing; data mining itself; and the final stage of describing the results of data mining algorithms, as well as information representation.

The author utilizes English abbreviations in this monograph, i.e. KDD for *knowledge discovery in databases* and DM for *data mining.*

1.2 Process of Knowledge discovery in databases

The process of knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (13).

Data is understood as the set of arranged facts (e.g. cases in a database) and knowledge is understood in terms of some language describing the sub-set of data or the model applicable to the sub-set. Therefore, the term knowledge discovery can be also understood as the proposal and specification of a researched reality model, so that it can efficiently describe the data, search for data structure description, and finally, prepare data description on a higher level (meta-knowledge).

KDD is defined as a broad-ranging process that comprises a series of specific steps including data preparation, searches for regularities, verification, testing and the refinement of discovered knowledge, where everything is repeated in numerous iterations. Non-triviality requires that the aforementioned steps include further partial tasks. It is not the direct calculation of advance known quantities. Legitimacy considers some rate of certainty and is not considered an absolute. The novelty and usefulness of the process relates to the user. It is aimed at mined data contributions and their possible transformation into knowledge. Comprehensibility often lacks and is subject to further processing, e.g. via data visualization. The usefulness relates directly to interest and is taken as a total rate of pattern value combining its legitimacy, novelty, usefulness and simplicity.

Functions of usefulness can be defined explicitly, or as the patterns that produce the answer to the requirements arranged due to the interest directed by KDD system.

The entire process is commonly divided into the following stages:

1. Problem definition – determination of the problem to be dealt with via KDD process.
2. Data selection – data are selected or segmented due to some criterion to choose a relevant data set for further use. For some DM algorithms it is enough to choose only data samples and it is not necessary to take out all of the data of relational databases or data warehouse.
3. Data cleaning and pre-processing – i.e. format modification and data cleaning, when some data are eliminated, since they are not necessary or they could prevent an efficient query evaluation. Data format modification is part of data cleaning, e.g. code of sex is unified to a binary attribute with values 0 and 1, etc.
4. Data transformation – the transformation of modified and cleaned data which can be enhanced by other attributes, for example, from external sources, and hence extend the usability not only of data but of results obtained as well.
5. Data mining – the search for interesting patterns, whose form depends on the selected DM method and can also be from classification rules, decision trees, function dependences, logical rules, etc. The results of this step significantly depend on the previous steps.
6. Interpretation and reporting – the patterns identified by the system are described as knowledge applicable for decision support (of human, program, etc.). The tasks related to prediction and classification for

deciding how the database content is summarized or the monitored phenomena are explained.

7. Discovered knowledge application – the application of discovered knowledge to the problem resolved (11), (13), (25), (28), (30), (50).

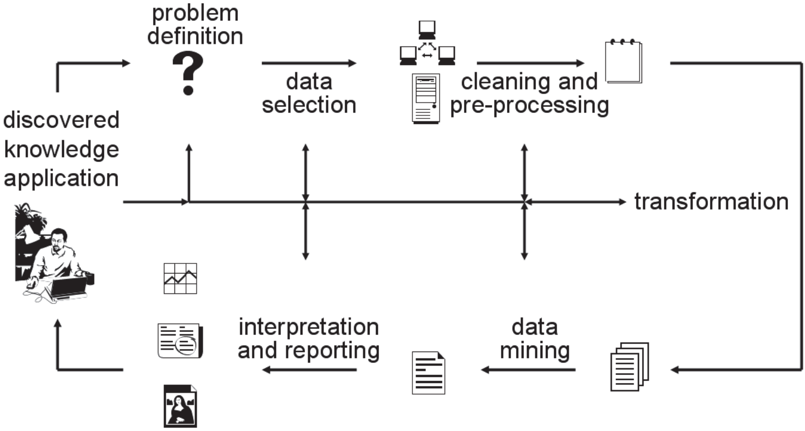

problem definition
data selection
cleaning and pre-processing
discovered knowledge application
transformation
interpretation and reporting
data mining

***Fig. 2** Process of Knowledge discovery in databases*

The process of knowledge discovery in databases is not the process in which its particular stages are executed in stated order from the beginning to the end. In the course of a random stage in the process it can result in the modifications of some of its setups. It can also result in the modification or change of the initial problem. The set of processed data can change transformation rules and can be modified, etc. When these changes occur, it is possible to go back in the process by the necessary number of stages and hence, suitably modify the necessary stage so that the process can meet the new requirements.

It frequently occurs that the achieved results do not correspond to the anticipated results, or another interesting result is achieved. This commonly leads to new iterations of the process in order to verify or extend the results achieved, so that the anomalies, which can occur e.g. by incorrect data set selection or by an unsuitable selected DM method, are eliminated.

1.3 Data mining

In relation to the subject matter there are two key perspectives on data mining:

Within the term of data mining we can associate all „more complex“ activities over the database, or possibly data warehouse. Precisely, we can determine data mining as a specific process of acquiring in advance unknown information. Here advanced tools prepared by specialists have to be available and the end user is in a better position, when he/she obtains new and often in advance unknown information.

There are many definitions of data mining. The following are some of the key definitions:

Fayyad: „Data mining is a single step in the process of knowledge discovery in databases that involves finding patterns in the data.” (12)

John: „Data mining is a specific process of data acquisition from quite large data warehouses via the extraction of relevant in advance of unknown information.“ (26)

Aaron Zornes, The META Group: „Data mining is the process of knowledge discovery via discovering in advance unknown action information in very large databases.“ (56)

Business definition: „Data mining is a process of selection, search and modelling in large bulks of data that allows for the revealing in advance unknown relations among data to win the business advantage.“ (26)

Statsoft: „Data mining is an analytical process proposed to search large bulks of data in order to find and verify consistent signs or systematic relations among variables.“ (57)

Gartner Group: „Data mining is a process of discovering meaningful new correlations, structures and trends by intrusion of an amount of saved data using the model recognition technologies that employ statistical and mathematical techniques.“ (26)

As seen, there are several different data analysis technologies definitions that are called data mining. Most of them include the approach of the search and discovery of useful relations within large databases. Similarly, as personal computers have become efficient and user friendly, new tools for data mining have been developed to gain an advantage from the growth of computer technology. Approaches in data mining developed as an answer to the new and more demanding requirements of making managerial decisions. Some definitions of data mining are anchored in specific analytical approaches such as neural networks, genetic algorithms, etc. Yet other definitions of data mining are sometimes changed by definitions concerning data saving in data warehouses. The preparation of data warehouses and data mining are complementary. The data warehouse provides data saving; however, it does not provide the transformation of data into information. Data mining changes data to information and information to knowledge (13), (17), (32).

The process of data mining is divided into three stages:

- data investigation,
- finding patterns or dependencies,
- verification of assembled model.

In an ideal case, the approach iteratively repeats until a sufficiently robust model is found. The final model is usually the result of previous knowledge combinations and newly discovered information. In practice the means for verification of the model are commonly limited. Therefore, in many cases it is necessary to use the heuristically derived results as well.

Data mining uses data to develop models of the real world. The description of patterns and relations in data are the result of the model.

Models can be divided into two basic types:

- models based on theory,
- models based on data.

Modelling based in theory, often called hypothesis testing, tries to prove or disprove initial ideas. User based modelling is a specific model founded mostly on prior knowledge that is further tested to learn whether the model is a correct one. On the other hand, data generated modelling tools automatically develop a model based on schemes found in data. It is also necessary to test such models until we can claim it a correct one.

The aims of the data mining process come out of their assumed use. There are two types of aims:

- verification,
- exploration.

In verification the system has to verify the user hypothesis and in exploration the system automatically looks for new knowledge.

The aims of exploration can be further classified as follows:

- prediction activity – the system acquires knowledge to predict behaviour of some entity,
- description – the system acquires knowledge to present results to the user in a transparent and comprehensible form.

1.4 Basic methods and data mining techniques

In the subject matter investigated several types of methods and data mining techniques were proposed. However, it is not easy to find a specific approach in this mixture of methods and techniques. According to the theoretical literature, the methods can be divided into two large groups:

- statistical,
- logical.

The literature also argues that the methods can be divided into:

- summarization and generalization,
- search for dependencies,
- classification and segmentation,
- statistical analysis (mainly regression),
- detection of changes and deviations,
- search for similarities in time, or time-space databases, etc.

From the many methods of data mining used in practice we can introduce the following:

- classification,
- regression,
- segmentation,
- summarization and generalization,
- modelling of associations,
- detection of changes and deviations,
- methods based on examples,
- prediction according to time orders,
- symbolic methods,
- discovery of sequences,
- search for similarities.

The methods stated are based on the techniques derived from statistics and inductive learning, e.g.:

- decision trees,
- association rules,
- neural networks,
- genetic algorithms,
- segmentation analysis,
- regression analysis.

A detailed description of basic methods and techniques, as well as other data mining methods and techniques, can found in the following references (4), (6), (11), (13), (15), (28), (32), (51), (66).

1.5 Problems with the application of Knowledge discovery in databases in practice

There can be problems in the utilization of knowledge discovery in database processes. Criteria for setting KDD, e.g. in industry or by scientific data analysis, can be classified according to practical and technical categories. Practical criteria for KDD projects are similar to other applications of advanced information technologies and have a potential influence on further applications. In their implementation it is necessary to consider the possible absence of simpler alternative solutions and to ensure strong organizational support for the use of the technology proposed.

When considering the technical criteria for using KDD methods in practice, it is essential to take into account the existence and applicability of appropriate data. In general, the data included both a large amount of items and a large amount of complex knowledge. The more complex data structure and the more complex dependencies they comprise, the larger amount of data is needed for the efficient application of data mining methods, e.g. a quality set of initial knowledge can significantly reduce necessary cases.

The other issue refers to the relevance or significance of the attributes. It is crucial to have data attributes that are significant for the given task. Data exploration based on attributes that do not comprise necessary information will not be successful, regardless of the amount of data used.

A low noise level is also very important, since high noise can impede knowledge identification. The quality of the initial background is another important aspect. The background details a given problem field – what the important items are, what the probabilistic relations are, what the user functionality will be, what knowledge is already known, etc. (13), (18), (21), (26), (32), (47), (68).

The main problems of applying the process of knowledge discovery in databases in practice are (13), (32):

1. Databases increase: Databases with thousands of attributes, sub tables with millions of records and sizes in terms of gigabytes or terabytes occur more frequently. Therefore, the methods for operating with large data volumes should comprise very efficient algorithms for sampling, approximation and massive parallel processing.
2. Bulkiness: It is not only about the large amount of records in the database, but also about the large amount of attributes or parameters, or the issue of dimensionality. Bulky data causes problems in the case of the increasing size of the searched state space for deriving the model according to combinatory explosion principles. Then the chance that DM algorithms find false knowledge, that is not further applicable, increases. The approach to this problem comprises the methods for efficient dimensionality reduction and the use of previous knowledge to identify the irrelevant variables.
3. Over-fitting: This concerns algorithm searches for the best parameters for one particular model based on a limited set of data. Hence, it can model not only general knowledge in data, but also any noise specific to the

particular data set, which causes poor performance and reduces the quality of the method employed.

4. Stating the statistical significance: The issue occurs (similarly as over-fitting), when the system tests the amount of acceptable possible models. For instance, if the system tests the models at the level of 0.001 significance on average, with pure random data, then N/1,000 models are considered to be important. This consideration is commonly missing in many initial experiments with KDD. One way to solve the issue, is to employ methods that utilize random testing or to adjust the statistical test as a function of searching.
5. Transformability of data and knowledge: Thanks to quickly changing non-stationary data, the previously discovered knowledge can be rendered invalid. The cases taken from a given application database can also be modified, deleted or they can change size for a certain time. To solve this we can use incrementation methods to update knowledge, regarding the current data change and then analyze the previous knowledge and compare them to the newly measured results.
6. Missing or noised data: This problem often occurs in business or company databases, e.g., the data from the US nose count shows an error of the scale of more than 20%. If the database is not considered for knowledge discovery in the future, some important attributes can be missing. A possible solution lies in a sophisticated statistical strategy to identify the covered variables and dependences.
7. Complex relations between items: Hierarchically structured attributes or values, relations between attributes and very complex ways of expressing the knowledge in the context of the entire database content require

algorithms, which can utilize such information efficiently. Historically, the data mining algorithms have been developed for simple record's attributes-value, and therefore it is still necessary to develop new methods to derive the relationships between variables.

8. Comprehensibility and simple clarity of discovered knowledge: It is important that the discovered knowledge is clear in the applications. Graphical representations, structured representations, structured rules, generating of a natural language and method for data and knowledge visualization are useful here. The implicit or explicit redundancy of discovered knowledge is a problem as well. In this case it is necessary to apply various strategies of „cleaning“ the acquired rules, which leads to redundancy elimination, and hence an increase in the base transparency.
9. Interaction of a user and previous knowledge: Current KDD tools and methods are not interactive enough and they cannot implement previous knowledge on the issue in the model (base) in any simple way. The use of domain (particular) knowledge is, however, very important in all KDD process steps. Bayesian's approach uses, for example, the previous conditioned probabilities as one of the forms of previous knowledge expression. Other advantages of using previous knowledge lie in the approach of searching the model or database state space model correctly.
10. Interaction with other systems: An independent knowledge system does not have to be very useful. Typical integration problems are, for example, integration with the base control system (e.g. via query interface), integration with table calculators and visualization tools or ensuring the reading from sensors in real time.

It is not possible to eliminate many of these problems in the course of knowledge discovery in databases application process completely. Therefore, it is necessary to be aware of them and try maximize the effort to eliminate them.

1.6 KDD process use in practical applications

Fields with a high added value of knowledge discovery in databases application process include the banking sector, telecommunications, insurance, marketing, heavy engineering, state administration, health care, astronomy, direct marketing, electronic businesses, criminal justice and the fight against terrorism. However, it is also used for political and pre-election campaigns and „micro-targeting“.

In banking it chiefly includes the archiving of banking transactions, applications for credit, bank repayments, revealing fraud in bank transfers and credit cards. In telecommunications it mainly includes information on telephone operations and payments, and in the case of mobile phones, the information on location etc. (46).

In the millions of insurance incidents, income tax returns, applications for grants there are relatively few incidents of fraud; yet, they can result in high losses. Therefore, the possibility of defining a so-called fraud model to serve to classify particular cases into homogenous groups is utilized so as to reveal fraud. Cases not falling into any of the large groups are suspicious thus there is a high probability that they represent the frauds (46), (70).

In supermarkets or in electronic businesses where this the purchasing of thousands of products, it is a distinct advantage to know which products are usually purchased together, so that it is possible to assist clients in their selection. The Market Basket Analysis is commonly utilized for analyzing the clients' previous purchases and searches for purchasing patterns. These patterns are then utilized for targeted offers in the form of personal or telephone offers, by sending an offer via post or e-mail, or possibly by the automated display in a browser of the website of the shop. They can also be utilized for the optimization of goods arrangement in the shop (1), (46), (48), (67), (69).

In health care large volumes of data are collected and saved (sex, age, temperature, blood pressure, check-up results, X-ray, and patient statements) which, however, are not sufficiently utilized in deciding upon treatment. The diagnostic model allows fast and precise diagnosis that uses all the relevant information available. Similarly, it is possible to identify those patients in the database, whose health conditions could worsen in the future (5).

Via telescopes we can obtain a lot of different data, which cannot be manually processed. For example, within the Palomar Observatory Sky Survey Project 3 TB of image data comprising approximately two billions of astronomically relevant objects have been obtained. The SKICAT System (Sky Image Cataloguing and Analysis Tool) first segments the images and then determines 40 different attributes for each object found. The objects are next automatically classified into groups via a decision tree (e.g. various star types or galaxies), which is the basis for further (manual) analysis. On the one hand, the SKICAT System is much faster than manual classification. On the other hand, it allows classifying of very distant objects, which are not possible to manually classify (12), (14).

The process of knowledge discovery in databases is also being gradually utilized in industry, e.g. in the detection of defects during the production process, then for quality control of the production process, and for production planning, etc. (10), (29), (40).

From the many other incidences of utilizing the process of knowledge discovery in databases we can introduce one more interesting example. Barack Obama, the US President, used these process possibilities in his election campaign to search for indecisive voters and sponsors (38).

1.7 KDD tools

Tools appropriate for the operation with knowledge discovery in databases can be classified according to different criteria. They can be firstly divided according to the numbers of provided methods and techniques for mining. Secondly, according to the means of achieved result presentations and according to interactivity in the course of data mining, etc. The criterion of the tool availability or the means of its distribution can be included in the criteria. As a consequence they are divided into commercial and non-commercial tools.

1.7.1 Commercial tools

Oracle Data Mining

The Oracle Company includes in its database server a complementary application determined for data mining. It is based on Oracle Data Mining SQL and Java API and ensures the possibility of data mining execution across company databases in real time. Because the tool is primarily designed to operate with Oracle databases, the time and system means necessary to transfer data from databases and transform them into a format determined for data mining are minimized. Due to the fact that the data remains saved in an Oracle database, it is possible to include them in the security policy provided by the database. The tool also operates within standard relational database as well as within the data warehouse (19). Models from Oracle Data Mining can be transferred into standard SQL language and transposed to business intelligence applications for further analysis.

Oracle Data Mining provides a rich graphical user environment, in which it is possible to set up specific parameters of data mining and employ these parameters even further when needs be. It is possible to utilize more methods and techniques, such as regression analysis, cluster analysis, classification methods, association rules, anomalies detection, patterns extraction, etc. and resolve a wider scope of tasks via them. Oracle Data Mining comprises rich possibilities in the visualization of results and their further analysis (35), (44).

SAS Enterprise Miner

The SAS Enterprise Miner simplifies the process of knowledge discovery in databases and allows the development of highly accurate, predictive and descriptive models based on the large amount of company data. It is suitable for fraud discovery, credit risk minimization, the estimation of source dependency. It also increases answer probability rate in the targeted marketing campaigns and can predicts customer behaviour, etc.

The tool provides a well-arranged graphical user interface in which it is possible to interactively set-up specific parameters of the process of knowledge discovery in databases. It incorporates various methods and techniques of data mining, e.g. cluster analysis, self-organizing maps, linear and logical regressions, decision trees, neural networks, etc. It is possible to integrate the tool with other SAS Institute Company products (7), (42), (53).

STATISTICA Data Miner

The STATISTICA Data Miner represents one of the most extensive tools for data mining based on a very simple and well-arranged user environment. It provides a large selection of integrated, immediately usable solutions systems to specific data mining issues. They can also be modified to solve non-typical problems. The techniques of data mining are based on efficient tools that are included in five modules, which can be used interactively or for building, testing and implementing new solution tools.

One of the modules is called STATISTICA QC Miner and it is used for company purposes, process monitoring, identifying the problems related to check and increase quality as well as for preventing such problems. It comprises functions for quality control, regulation diagrams, process qualification analysis, experiment proposal procedures and Six Sigma methodology implementation that is connected with advanced exploration and support data mining procedures.

It also includes other modules for operation with neural networks, statistical data evaluation, Text Mining and Web Mining.

The tool permits connection to various types of relational databases (e.g. Oracle, Microsoft SQL Server, MySQL, etc.), as well as various data warehouses (57), (58).

Microsoft SQL Server 2008 Data Mining

The Microsoft SQL Server 2008 allows for the execution of qualified decisions and predictive analyses via intuitive data mining that is perfectly integrated in the Microsoft Business Intelligence platforms in company applications.

It affords the user the possibility to examine market basket analysis, a customer's susceptibility to leave and the reasons for this, market analyses and prediction of customer's behaviour or the process monitored. Furthermore, the user can predict process weaknesses, construct environment analyses so as to conduct targeted marketing campaigns, etc. The tool includes modules for Web Mining and Text Mining as well.

It also provides the user with a graphical interface in which he/she can define particular steps of the process of knowledge discovery in databases, or utilize the DM Wizard to help with processes of definition.

It can connect to standard programs in the Microsoft Office 2007 package. This allows cooperation between Microsoft SQL Server 2008 and Microsoft Excel 2007 program, or Microsoft Visio 2007 during the whole process of knowledge discovery in databases (33), (39), (43), (62).

IBM InfoSphere Warehouse

The IBM InfoShare Warehouse provides a primary solution for the development and administration of a data warehouse. It includes a module for data mining. The module allows simple data mining within company data warehouses. It is mainly used for fraud detections, exposure of a customer's susceptibility to leave, identifying customer segmentation and simple analysis of a market basket, etc. The tool's structure allows the integration of the module of data mining to existing systems. This provides wide flexibility and highly efficient and prediction analyses without the necessity of transferring them into their own special database.

It also provides basic methods such as cluster analysis, associations, classification and prediction and uses them in a graphical Design Studio. The models developed can be transformed into the industrial standard Predictive Model Markup Language (PMML), thus allowing for easier portability among various applications. Models for data mining can be activated in the production environment so as to carry out data analyses in real time. The tool also provides rich visual presentations of results (13).

PASW Modeller

This tool is a derivation of the well-known tool called Clementine. It comprises a well-arranged graphical user interface in which it is possible to define precisely the particular steps of the process of knowledge discovery in databases. Some parts of the process can be carried out automatically since the PASW Modeller includes tools such as automated data modification and transformation. It is also possible to incorporate data from the databases developed by various producers (e.g. IBM, MySQL, Oracle, Sybase IQ, etc.).

The PASW Modeller comprises various modules for the execution of specific analyses, e.g. cluster and regression analyses, a module for dependences analysis, a module for operations with neural networks, etc.

The tool also comprises special modules for Web Mining and Text Mining (56).

1.7.2 Non-commercial tools

Weka

The Weka System (Waikato Environment for Knowledge Analysis) was developed and implemented at University of Waikato, Hamilton, New Zealand. It uses the Java language, and therefore it is possible to easily install and use in MS Windows and Linux operation systems. Weka is continuously being developed and enhanced by including new algorithms of machine learning and data mining methods and by incorporating some support tools for data preparation and results processing. Achieving the best possible applicability is a key aim and therefore it is not (in contrast to many commercial systems) only oriented towards special data formats and assumed utilization. Sample data files are in a text format making it easy to try the many possibilities of Weka system tools during installation. It then includes its own text data editor and a conversion function for some commonly used text and binary formats.

Even if it is a non-commercial, freely distributed program, in many respects it can be compared to commercial systems. Weka is an experimental system that offers a wide scope of algorithms for learning and pre-processing that are known by academics. It further provides the possibilities of utilizing the regression and linear analyses, classification, time oriented models, decision trees, neural networks, etc. Finally, it has rich possibilities for result visualization and means for proposed model combination at its disposal (65).

LISp–Miner

Since 1996 at the University of Economics in Prague (VŠE), Faculty of Informatics and Statistics have been developing an open academic project LISp–Miner that is designed for research and teaching the process of knowledge discovery in databases. It is used mainly for teaching students, however, it can be used also for small or medium-sized KDD projects.

The project provides algorithms for searching for interesting relationships in data (it thus follows the General Unary Hypotheses Automation (GUHA) method), as well as algorithms for decision rules development. The core is produced by several GUHA procedures. GUHA is an original Czech-Slovak method of data exploration analysis. Its development began in the Czech-Slovak Academy of Sciences in the 1960s.

The LISp–Miner project is developed by the module for data preparation and pre-processing and seven analytical procedures: 4FT-Miner (the latest implementation of the original GUHA procedure ASSOC enhanced by several new elements), KL-Miner, CF-Miner, SD4FT-Miner, SDKLMiner, SDCF-Miner and KEX. With the exception of the KEX procedure determined to develop the decision rules, and hence for the task of classifying type, all other procedures are focused on searching for different types of rules for describing the given data (in compliance with GUHA method called hypotheses). All of these procedures are the result of author’s team of VŠE. The project utilizes the original technique of bit chains in generating and testing. This significantly increases the speed of the calculation. Several student software projects follow in LISp–Miner.

LISp–Miner was applied, for example, on the data analysis of risks factors of atherosclerosis in middle-aged men (within the EuroMISE research centre), on data analysis of reasons for traffic accidents in the Great Britain (within European research project Sol-Eu-Net), on data analysis of Prague residents' opinions of the quality of life, and on the analysis of event descriptions of football matches (within European research project K-Space) (5), (37).

Ferda Data Miner

The Ferda Data Miner continues on from the LISp–Miner project, and serves as a reliable basis for the utilization of GUHA methods for data mining. Ferda has been developed since 2004 and was first introduced in 2005. The first version of Ferda system still used LISp–Miner modules, nevertheless, the newer versions 2.xx onwards are independent. The new versions bring improved process setup and possibilities for its adjustment. Via new system modules it is now possible to implement knowledge domains directly into the process. It represents a highly modular system, which can utilize various complementary modules programmed in five programming languages and also provides its own partially recursive programming language.

It comprises seven relational GUHA methods, two multi-relational procedures, new quantifiers, supports ontology, user deciding, etc. (75).

TANAGRA

TANAGRA is another project developed for academic and research purposes. It utilizes several methods of data mining resulting from data analysis, statistical learning, machine learning and database processing.

The project continues in the SIPINA Project that is primarily based on the utilization of various machine-learning algorithms, especially on the interactive and visual construction of decision trees (55). TANAGRA enhances the possibilities of its predecessor in the field of supported methods; it comprises several controlled learning, cluster and factorial analyses, parametric and non-parametric statistical methods, association rules, etc. TANAGRA is a freely distributed project together with the source application code, which in compliance with the license contract allows the user to add their own algorithms in the project.

The main project objective is to provide researchers and students with a simple and well-arranged tool for data mining that corresponds with current standards for software development in the field (chiefly by the proposal of a graphical user interface and its utilization) and allows for the analysis of real or synthetic data.

The second TANAGRA objective is to provide a tool, which can be complemented by its own methods of data mining and hence compare the achieved results. Tanagra also acts as an experimental platform that allows the user to examine new methods of data mining without the necessity to program the tool for their implementation.

The third and last project objective is to provide the beginner programmer with the possibility of knowledge enhancement in the field of the proposal and programming of data mining applications. There is free access to the source application codes that provide information on project structure analysis, its particular libraries and functional parts. Therefore, TANAGRA can be considered to be a pedagogical sound means for teaching programming techniques as well.

TANAGRA in its current version includes direct access to data warehouses and to databases of various producers, data pre-processing and cleaning, higher process interactivity, etc., which are typically a strength of commercial products (52), (61).

2. IDENTIFICATION OF PRODUCTION SYSTEM PROBLEMS RESOLVABLE VIA DATA MINING

An industrial company with an implemented Computer Integrated Manufacturing system (CIM) is the suitable environment for data mining application. In such a company there is an integrated information system called information and control system. The following figure shows its standard structure (28), (59), (60).

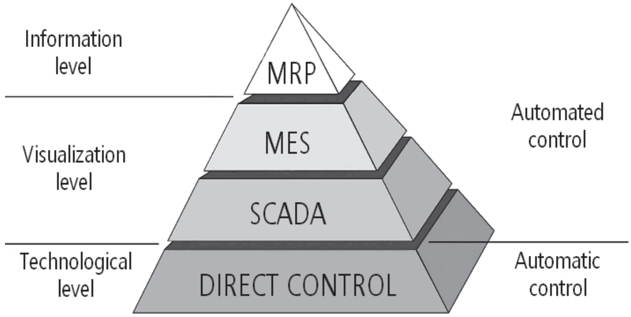


***Fig. 3** Structure of information and control system in industrial company*

MRP (Manufacturing Resource Planning)

Managerial level – functions of non-operation business (investments, financial and personnel planning, development strategy processing, etc.). Knowledge discovered in databases on this level can be used in the evaluation and analysis of demand, purchase, costs and customers behaviour in terms of consumption and production of the company concerned.

MES (Manufacturing Execution System)

Management of production operation – this refers to functions of technical control management (production process monitoring, power balancing, quality assurance) and disposal functions (maintenance, connection of manufacturing operation to its surroundings and environment). Knowledge discovered in databases at this level can be utilized in proposing and controlling the human-machine interaction, decision making in critical situations, etc.

SCADA (Supervisor Control and Data Acquisition)

Block level – collection and processing of process information (measurement, control and regulation) as well as service and monitoring of technological processes. Knowledge discovered in databases on this level can be used for current state description, prediction of emergency situations, production anomalies and quality control.

Direct control

Process level – this refers to effect on processes and information acquisition from processes. Knowledge discovered in databases at this level can be used for behaviour prediction of manufacturing devices and also their control.

In the production process a sufficient amount of data collected and processed at block level (SCADA) origins, and the data are further utilized at higher levels. It is appropriate to apply the data mining process to this kind discovered data. On the basis of previous information the data can predict potential malfunctions and states that endanger manufacturing process continuity on the manufacturing process, progress of production batches, production devices setup, sequence of operations, etc. This knowledge can further help production quality management.

The following problems typically arise in the process of knowledge discovery in databases for industries:

- identification of production parameters influence on production process,
- identification of defective cases in production,
- identification of production time series,
- detection of deviations during the course of production,
- detection of defective states of production devices,
- production quality assurance,
- optimization of the production process,
- optimization of production device arrangement,
- optimization of warehouse administration,
- prediction of preventive checks for production devices,
- prediction of defects in the production process,
- prediction of customers' behaviour,
- prediction of defects in assembly.

In the production process there are many problems that can be resolved by the process of knowledge discovery in databases. Correct problem identification and a corresponding solution are important. The process of knowledge discovery in databases provides sufficient means to deal with the problems that arise in the field of production systems.

3. PRODUCTION SYSTEM SIMULATION MODEL PROPOSAL AND RELATIONAL DATABASE STRUCTURE

3.1 Production system simulation model proposal

The Witness software tool developed by Lanner Group Ltd was used for the production system proposal (22), (36), (49). This tool is one of the most commonly used software for the simulation and optimization of production, service and logistical systems. It permits the development of the production system simulation model from required production objects (73). All of the production system objects can have their parameters set up and then it is possible to monitor how these parameters change in the course of the production process. In Witness models the material or system customer's motions, states of the particular objects, operations carried out, the current utilization of resources are displayed dynamically. Simultaneously, all events occurring in the production process are recorded (71). The data can be collected and used for more detailed analyses of the proposed production process model. The Witness software tool is supported by direct cooperation with selected database systems, table editors and other programs using ActiveX or ODBC standard (72), (74).

The proposed model of the production system represents the production line that will produce of three types of technologically similar products. The products are sent into the production process in individual production batches with a dynamically changing size – from one to six products in one the production batch.

The production line consists of ten individual production devices and each device can carry out more production operations. Amongst the various production operations on one production device it is necessary to execute the arrangement of the device. The time of the arrangement is defined individually for each of the devices as well as for each specific operation to be carried out.

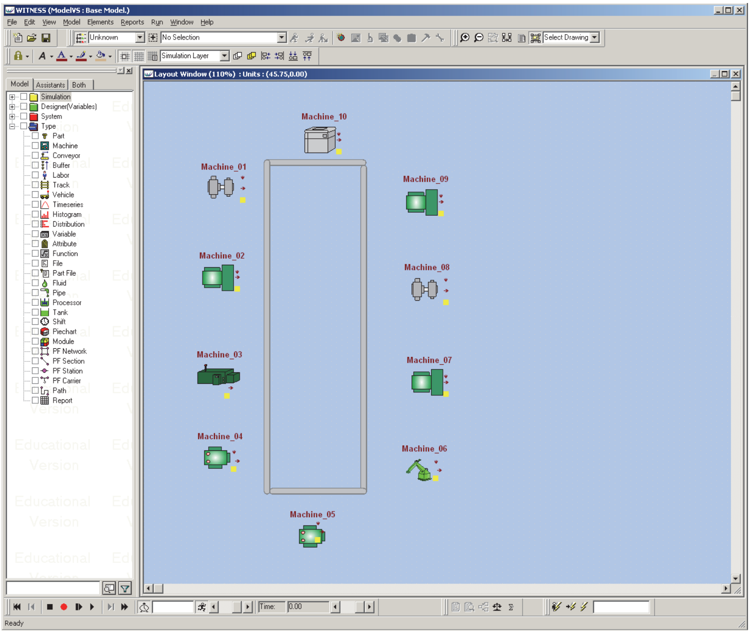


Fig. 4 Production system model in Witness software tool

The P\production process of specific products is carried out on various production devices:

Production batch of the Product 1 comes through Machine_01 → Machine_03 → Machine_04 → Machine_07 → Machine_09. After operations end on the machine No. 9 the production batch of Product 1 leaves the production system.

Production batch of the Product 2 comes through Machine_01 → Machine_02 → Machine_05 → Machine_06 → Machine_08 → Machine_10. After the operation ends on the machine No. 10 the production batch leaves the production.

Production batch of the Product 3 comes through Machine_02 → Machine_03 → Machine_04 → Machine_06 → Machine_07 → Machine_08 → Machine_10. After operations end on the machine No. 10 the production batch leaves the production.

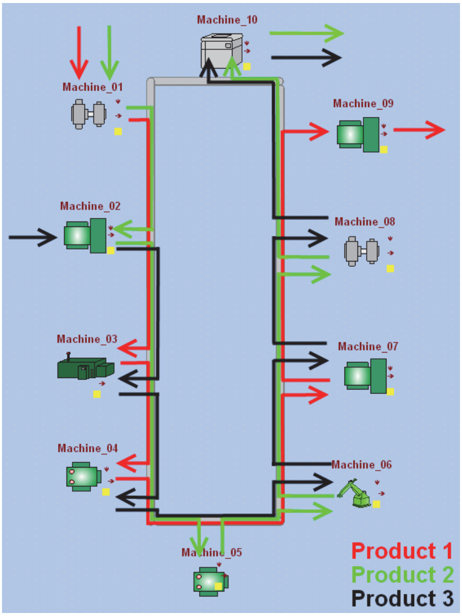
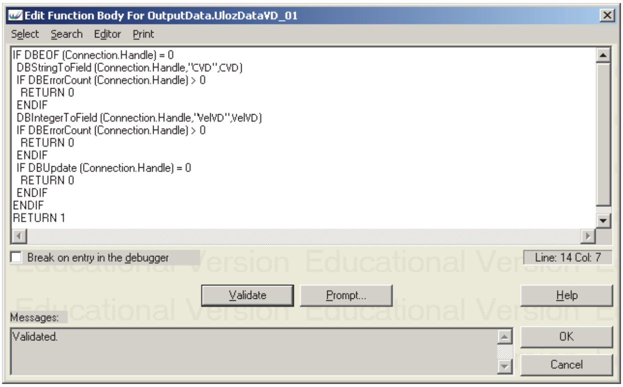


Fig. 5 Material flow in production system

Individual production batches entering the production system have variable input times as well as variable production batch sizes.

Production devices are connected to one another on a closed transport system of three conveyor belts. The transport capacity corresponds to the maximum size of one production batch, i.e. for the transport of the entire production batch one transport device is sufficient. Each production device comprises its own manipulation spot that is determined for the loading and unloading of production batches.

Data generated by the production system in the course of simulation is saved in the proposed information system. The following are the examples of data saved in the information system: production batch identification of particular products, the number of products in one production batch, the production device employed, the start and end of production operation, the production operation type, the production device condition, etc.



Edit Function Body For OutputData.UlozDataVD_01

```
IF DBEOF (Connection.Handle) = 0
 DBStringToField (Connection.Handle,"CVD",CVD)
 IF DBErrorCount (Connection.Handle) > 0
  RETURN 0
 ENDIF
 DBIntegerToField (Connection.Handle,"VelVD",VelVD)
 IF DBErrorCount (Connection.Handle) > 0
  RETURN 0
 ENDIF
 IF DBUpdate (Connection.Handle) = 0
  RETURN 0
 ENDIF
ENDIF
RETURN 1
```

□ Break on entry in the debugger

Line: 14 Col: 7

Messages:

Validated.

Fig. 6 *Example of function definition for database operation*

Production devices have a defined probability of device failure as well as the necessary time for its elimination. If the device fails, it is not possible to carry out the production operations and the production batches coming through the device are stopped. The time of origin and removal of the device failure as well as the identification of the failed production device are recorded in the proposed information system.

The simulation time was set to 2 592 000 seconds or 30 days. The simulation time converted to data and time starts at 0:00:00 on 1 Jan 2012 and ends at 0:00:00 on 31 Jan 2012. In the course of the simulation 23,726 database records of the production process were generated.

3.2 Proposal of relational database structure

During the simulation the proposed model of production system generates data. It is necessary to appropriately collect and save this for further analysis. An information system with a relational database was chosen for production process data saving. Relational databases allow the saving of generated data in the course of a production process simulation and at the same time it provides data availability for analysis 9.

Database server Oracle 10g release 2 was used as RDBMS. Oracle provided sufficient tools for the operation and administration of the proposed relational database and the database server in question can be used as an input data source for further analysis in STATISTICA Data Miner.

3.2.1 Development of UML information system model

The UML methodology was employed for the proposal of the information system (2), (3), (27), (45), (54), (63), (64). For the proposed model of information system we utilized the development tool IBM Rational Software Architect 7.0. This allowed the modelling of the static and dynamic structure of the proposed model. It is thus possible to build a physical database model in it and generate SQL script for its development with respect to the selected database server.

In Stage 1 of the production information system proposal the catalogue of user requirements was elaborated. The catalogue comprises the basic requirements for the information system proposed.

Further, the particular participants and their actions in the proposed information system were identified.

To utilize the process of knowledge discovery in databases it is crucial that all the necessary data is saved in the database. This refers to the data related to the production devices and the production process itself. The participant „DM Analyst“ will carry out all the activities related to data mining in the system. This actor will administer the production devices, he will be able to add them, modify their particular parameters, or if not used, delete them from the production process. The other task of the DM Analyst is to administer the production process. He will be able to set particular parameters of the production process, such as production process sequence, production batches size, possible production alternatives, etc.

However, data administration for data mining analyses will be the main task of DM Analyst. He will be able to define the data utilized in the production process and define its type, amount and format.

The DM Analyst can also define the production process. The production process definition will serve as a source for data mining analyses and is always specified for only one product type. The production process can include several production devices, while in the course of the production process these production devices can carry out more production operations. For one production operation more alternative devices can be determined.

3.2.2 Development of relational database data model

Regarding the UML system model proposed in the previous stage it is possible to propose a data model. The data model was developed by mapping the classes in the class diagram into database tables. Oracle 10g Release 2 was selected as the database server and a SQL script was also generated for it. The relational database in RDBMS Oracle was developed via the generated SQL script.

Then in the Oracle 10g database server an independent tablespace was developed for the proposed information system and the new users who can operate it. The DM Analyst user determined for data mining operations was assigned access rights to log in and access to specific database tables of the information system. This permitted the DM Analyst user to log into the database also via the STATISTICA Data Miner Program.

To access the production system simulation model a user ProdSystem was developed. The user had access rights to log into the database to only record in the tables in which the data generated by the production system simulation model are recorded.

The rest of the proposed information system users (Administrator, Seller, etc.) are assigned access rights to log into the database and to access particular database tables with respect to their assigned functionality in the information system proposed.

The resulting data model is illustrated in the following figure.

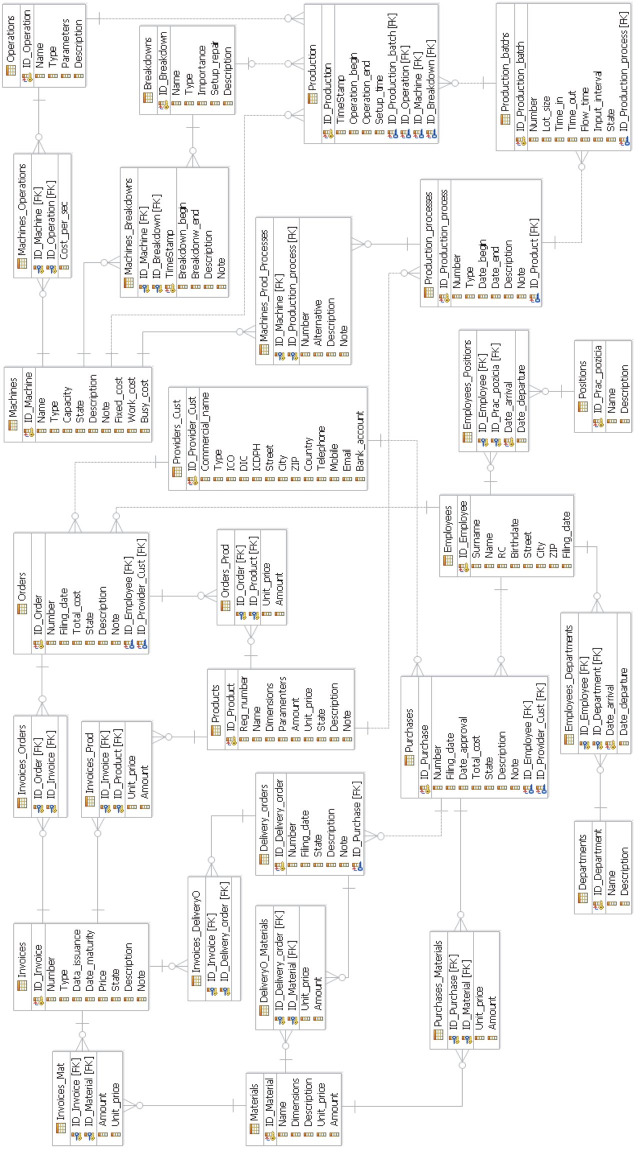


***Fig.** 7 Data model of the proposed relational database*

4. ANALYSIS OF ACQUIRED PRODUCTION SYSTEM DATA

For the production system data analysis STATISTICA Data Miner version 9.1, the tool produced by StatSoft Company was used. The tool allows the application of the process of knowledge discovery in databases from data acquisition, through their transformation and modification, and data mining in evaluation of the achieved results. All process steps can be conducted via a graphical application interface, or it is possible to specify the particular process parts by the related source code assignment, e.g. SQL script for the specific input data selection or by the modification of particular parameters in the source code of employed methods and data mining techniques.

Analysis of the data acquired in the proposed production system was divided into a few steps. The following figure shows the procedure analysis.

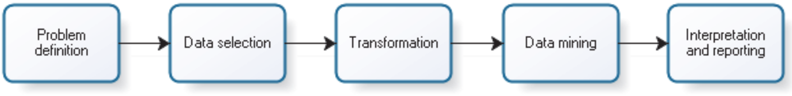


***Fig. 8** Analysis of acquired production system data*

4.1 Definition of data mining objectives

There are several problems (objectives), which should be solved by the application of the process of knowledge discovery in databases:

- analysis of production process influences on production device utilization,
- analysis of production process influences on the flow time of production batches,

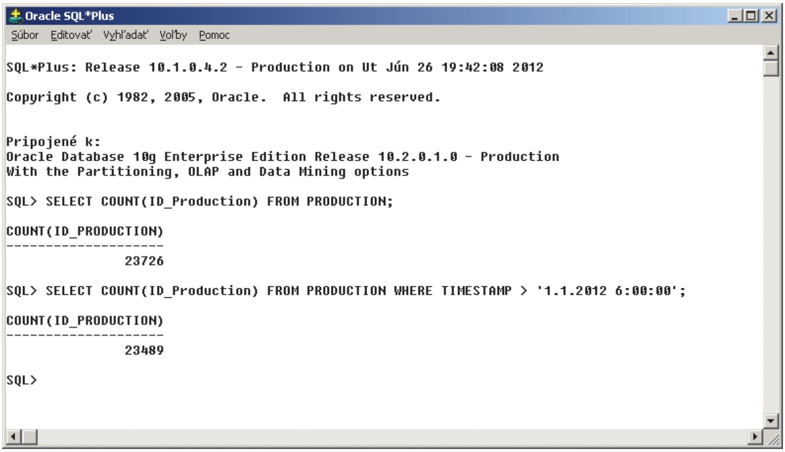
- analysis of production process parameters on the number of manufactured products.

The production process has a precisely determined production route for particular products types. The route represents a precise sequence of individual production batches according to the production plan specified. Therefore, change cannot be considered in the production system arrangement. The main parameters influencing the specified objectives are as follows: sizes of individual production batches, production devices utilization, average production time, number of the products manufactured, and intervals of individual production batches inputs. In the simulation model the production batch size was set as a variable value increasing in the simulation process the values of the interval from one product in the production batch up to six products in a production batch. The interval of the individual production batch input is set in the simulation model as a variable. Other production process parameters change in dependence on the production process operating.

4.2 Data selection from the database

Considering that the proposed simulation model generates the data, not all of the records saved in the information system were selected for the analysis set of data. To prevent undue influence on the results achieved, the data from the first six hours of the production process will be omitted. It represents the time when the production system is gradually „filled up". In general, the first production batch will have a shorter production time, since they are passing through an empty production system and they do not have to wait in front of the particular production devices, until the production operation of the previous production batch is finished. Therefore, the utilization of particular production devices is significantly less as well. The selected time of six hours is enough to fill up the production system sufficiently and avoid influencing the results achieved.

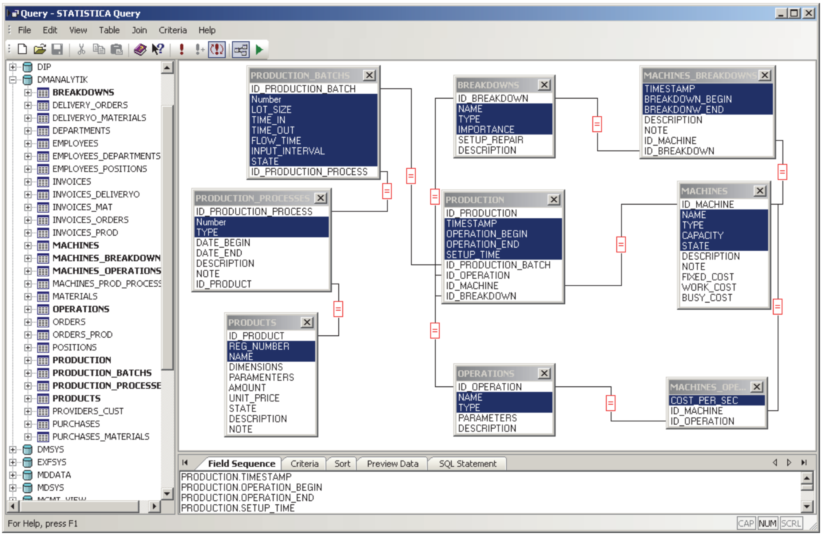
The connection of the STATISTICA Data Miner application to the database server was developed via the STATISTICA Query tool, that allows the definition of the connection to the supported database servers. The connection is utilized via the Oracle 10g Release 2 database server.



```
SQL*Plus: Release 10.1.0.4.2 - Production on Ut Jún 26 19:42:08 2012

Copyright (c) 1982, 2005, Oracle.  All rights reserved.


Pripojené k:
Oracle Database 10g Enterprise Edition Release 10.2.0.1.0 - Production
With the Partitioning, OLAP and Data Mining options

SQL> SELECT COUNT(ID_Production) FROM PRODUCTION;

COUNT(ID_PRODUCTION)
--------------------
               23726

SQL> SELECT COUNT(ID_Production) FROM PRODUCTION WHERE TIMESTAMP > '1.1.2012 6:00:00';

COUNT(ID_PRODUCTION)
--------------------
               23489

SQL>
```

Fig. 9 *Number of records saved in information system*

After connecting to the selected database server, the STATISTICA Query tool provides in its graphical mode the possibility to select the required database tables, for which the required attributes can be determined and hence used in subsequent analyses. It also includes the possibility to define the basic criteria for the connection of individual database tables and other requirements related to the connection.

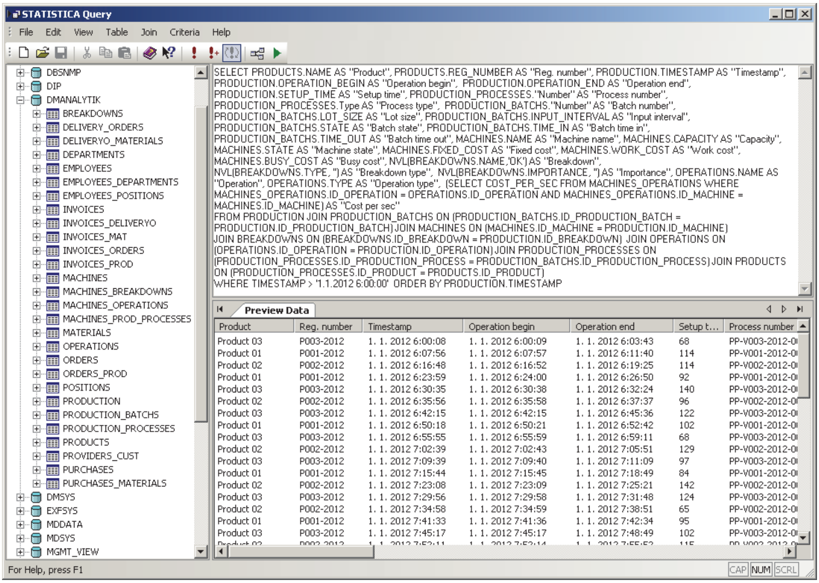


```
PRODUCTION.TIMESTAMP
PRODUCTION.OPERATION_BEGIN
PRODUCTION.OPERATION_END
PRODUCTION.SETUP_TIME
```

Fig. 10 STATISTICA Query – graphical mode

In the graphical mode it is not possible to define more complex selection queries comprising, for example, the definition of the submerged selection query, the replacement of the „null“ attributes, the use of aggregation functions, etc. Therefore, the selection query was developed in the text mode.

This was also the case for the selection queries for other input sets of data regarding the subject matter investigated. For instance, it was necessary to apply the functions for the acquisition of the production batch flow time and the selection of complementary data for the input data analysis, etc.



```
SELECT PRODUCTS.NAME AS "Product", PRODUCTS.REG_NUMBER AS "Reg. number", PRODUCTION.TIMESTAMP AS "Timestamp",
PRODUCTION.OPERATION_BEGIN AS "Operation begin",  PRODUCTION.OPERATION_END AS "Operation end",
PRODUCTION.SETUP_TIME AS "Setup time", PRODUCTION_PROCESSES."Number" AS "Process number",
PRODUCTION_PROCESSES.Type AS "Process type",  PRODUCTION_BATCHS."Number" AS "Batch number",
PRODUCTION_BATCHS.LOT_SIZE AS "Lot size", PRODUCTION_BATCHS.INPUT_INTERVAL AS "Input interval",
PRODUCTION_BATCHS.STATE AS "Batch state", PRODUCTION_BATCHS.TIME_IN AS "Batch time in",
PRODUCTION_BATCHS.TIME_OUT AS "Batch time out", MACHINES.NAME AS "Machine name", MACHINES.CAPACITY AS "Capacity",
MACHINES.STATE AS "Machine state", MACHINES.FIXED_COST AS "Fixed cost", MACHINES.WORK_COST AS "Work cost",
MACHINES.BUSY_COST AS "Busy cost", NVL(BREAKDOWNS.NAME,'OK') AS "Breakdown",
NVL(BREAKDOWNS.TYPE, '') AS "Breakdown type",  NVL(BREAKDOWNS.IMPORTANCE, '') AS "Importance", OPERATIONS.NAME AS
"Operation", OPERATIONS.TYPE AS "Operation type",  (SELECT COST_PER_SEC FROM MACHINES_OPERATIONS WHERE
MACHINES_OPERATIONS.ID_OPERATION = OPERATIONS.ID_OPERATION AND MACHINES_OPERATIONS.ID_MACHINE =
MACHINES.ID_MACHINE) AS "Cost per sec"
FROM PRODUCTION JOIN PRODUCTION_BATCHS ON (PRODUCTION_BATCHS.ID_PRODUCTION_BATCH =
PRODUCTION.ID_PRODUCTION_BATCH) JOIN MACHINES ON (MACHINES.ID_MACHINE = PRODUCTION.ID_MACHINE)
JOIN BREAKDOWNS ON (BREAKDOWNS.ID_BREAKDOWN = PRODUCTION.ID_BREAKDOWN)  JOIN OPERATIONS ON
(OPERATIONS.ID_OPERATION = PRODUCTION.ID_OPERATION) JOIN PRODUCTION_PROCESSES ON
(PRODUCTION_PROCESSES.ID_PRODUCTION_PROCESS = PRODUCTION_BATCHS.ID_PRODUCTION_PROCESS) JOIN PRODUCTS
ON (PRODUCTION_PROCESSES.ID_PRODUCT = PRODUCTS.ID_PRODUCT)
WHERE TIMESTAMP > '1.1.2012 6:00:00'  ORDER BY PRODUCTION.TIMESTAMP
```

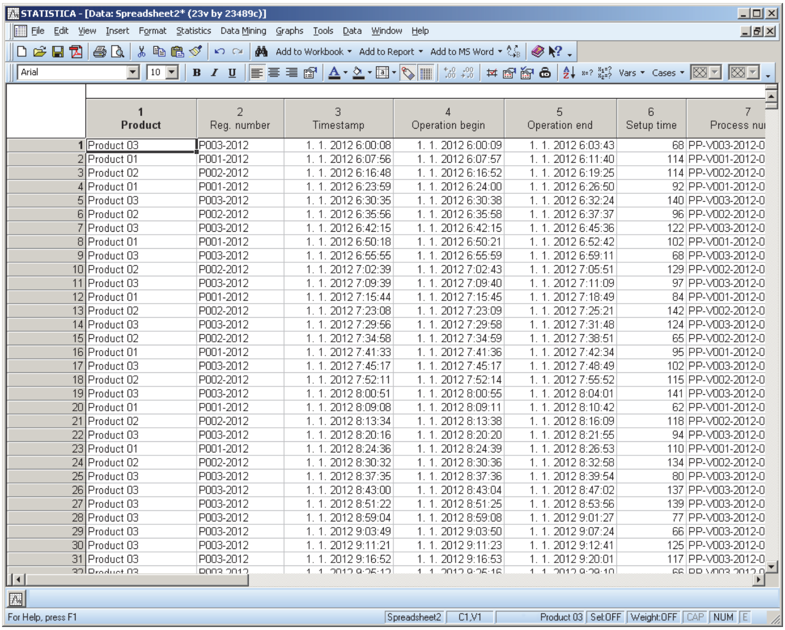
Product	Reg. number	Timestamp	Operation begin	Operation end	Setup t...	Process number
Product 03	P003-2012	1. 1. 2012 6:00:08	1. 1. 2012 6:00:09	1. 1. 2012 6:03:43	68	PP-V003-2012-0
Product 01	P001-2012	1. 1. 2012 6:07:56	1. 1. 2012 6:07:57	1. 1. 2012 6:11:40	114	PP-V001-2012-0
Product 02	P002-2012	1. 1. 2012 6:16:48	1. 1. 2012 6:16:52	1. 1. 2012 6:19:25	114	PP-V002-2012-0
Product 01	P001-2012	1. 1. 2012 6:23:59	1. 1. 2012 6:24:00	1. 1. 2012 6:26:50	92	PP-V001-2012-0
Product 03	P003-2012	1. 1. 2012 6:30:35	1. 1. 2012 6:30:38	1. 1. 2012 6:32:24	140	PP-V003-2012-0
Product 02	P002-2012	1. 1. 2012 6:35:56	1. 1. 2012 6:35:58	1. 1. 2012 6:37:37	96	PP-V002-2012-0
Product 03	P003-2012	1. 1. 2012 6:42:15	1. 1. 2012 6:42:15	1. 1. 2012 6:45:36	122	PP-V003-2012-0
Product 01	P001-2012	1. 1. 2012 6:50:18	1. 1. 2012 6:50:21	1. 1. 2012 6:52:42	102	PP-V001-2012-0
Product 03	P003-2012	1. 1. 2012 6:55:55	1. 1. 2012 6:55:59	1. 1. 2012 6:59:11	68	PP-V003-2012-0
Product 02	P002-2012	1. 1. 2012 7:02:39	1. 1. 2012 7:02:43	1. 1. 2012 7:05:51	129	PP-V002-2012-0
Product 03	P003-2012	1. 1. 2012 7:09:39	1. 1. 2012 7:09:40	1. 1. 2012 7:11:09	97	PP-V003-2012-0
Product 01	P001-2012	1. 1. 2012 7:15:44	1. 1. 2012 7:15:45	1. 1. 2012 7:18:49	84	PP-V001-2012-0
Product 02	P002-2012	1. 1. 2012 7:23:08	1. 1. 2012 7:23:09	1. 1. 2012 7:25:21	142	PP-V002-2012-0
Product 03	P003-2012	1. 1. 2012 7:29:56	1. 1. 2012 7:29:58	1. 1. 2012 7:31:48	124	PP-V003-2012-0
Product 02	P002-2012	1. 1. 2012 7:34:58	1. 1. 2012 7:34:59	1. 1. 2012 7:38:51	65	PP-V002-2012-0
Product 01	P001-2012	1. 1. 2012 7:41:33	1. 1. 2012 7:41:36	1. 1. 2012 7:42:34	95	PP-V001-2012-0
Product 03	P003-2012	1. 1. 2012 7:45:17	1. 1. 2012 7:45:17	1. 1. 2012 7:48:49	102	PP-V003-2012-0
Product 02	P002-2012	1. 1. 2012 7:52:11	1. 1. 2012 7:52:14	1. 1. 2012 7:55:52	115	PP-V002-2012-0

Fig. 11 STATISTICA Query – text mode

After defining the selection query (either in graphical or text modes) and switching to the STATISTICA Data Miner tool, database data was recorded into the internal table (in the application called Spreadsheet). The internal table lines correspond with individual lines obtained from the selected query and the columns (in the application called variables) represent the specific query attributes. This recorded data can be the modified according to the format required; the user can modify the names representing the particular variables, order them, etc.

If a portion of the data will not be used in subsequent analyses, mean independent lines or whole table columns (variables), it can be deleted from the internal tables without redefining the selection query. If it is necessary to add new database columns or read the modified data, the selection query has to be modified and read again via the STATISTICA Query tool.

The STATISTICA Data Miner permits adding a new column – a variable, which comprises not the data read from the database, but the calculated value of the specified expression. This possibility was utilized, for example, for the calculation of the flow time of one production batch in terms of the following variables „Time_in” and „Time_out“, read from the database.



STATISTICA - [Data: Spreadsheet2* (23v by 23489c)]

	1 Product	2 Reg. number	3 Timestamp	4 Operation begin	5 Operation end	6 Setup time	7 Process nu
1	Product 03	P003-2012	1. 1. 2012 6:00:08	1. 1. 2012 6:00:09	1. 1. 2012 6:03:43	68	PP-V003-2012-0
2	Product 01	P001-2012	1. 1. 2012 6:07:56	1. 1. 2012 6:07:57	1. 1. 2012 6:11:40	114	PP-V001-2012-0
3	Product 02	P002-2012	1. 1. 2012 6:16:48	1. 1. 2012 6:16:52	1. 1. 2012 6:19:25	114	PP-V002-2012-0
4	Product 01	P001-2012	1. 1. 2012 6:23:59	1. 1. 2012 6:24:00	1. 1. 2012 6:26:50	92	PP-V001-2012-0
5	Product 03	P003-2012	1. 1. 2012 6:30:35	1. 1. 2012 6:30:38	1. 1. 2012 6:32:24	140	PP-V003-2012-0
6	Product 02	P002-2012	1. 1. 2012 6:35:56	1. 1. 2012 6:35:58	1. 1. 2012 6:37:37	96	PP-V002-2012-0
7	Product 03	P003-2012	1. 1. 2012 6:42:15	1. 1. 2012 6:42:15	1. 1. 2012 6:45:36	122	PP-V003-2012-0
8	Product 01	P001-2012	1. 1. 2012 6:50:18	1. 1. 2012 6:50:21	1. 1. 2012 6:52:42	102	PP-V001-2012-0
9	Product 03	P003-2012	1. 1. 2012 6:55:55	1. 1. 2012 6:55:59	1. 1. 2012 6:59:11	68	PP-V003-2012-0
10	Product 02	P002-2012	1. 1. 2012 7:02:39	1. 1. 2012 7:02:43	1. 1. 2012 7:05:51	129	PP-V002-2012-0
11	Product 03	P003-2012	1. 1. 2012 7:09:39	1. 1. 2012 7:09:40	1. 1. 2012 7:11:09	97	PP-V003-2012-0
12	Product 01	P001-2012	1. 1. 2012 7:15:44	1. 1. 2012 7:15:45	1. 1. 2012 7:18:49	84	PP-V001-2012-0
13	Product 02	P002-2012	1. 1. 2012 7:23:08	1. 1. 2012 7:23:09	1. 1. 2012 7:25:21	142	PP-V002-2012-0
14	Product 03	P003-2012	1. 1. 2012 7:29:56	1. 1. 2012 7:29:58	1. 1. 2012 7:31:48	124	PP-V003-2012-0
15	Product 02	P002-2012	1. 1. 2012 7:34:58	1. 1. 2012 7:34:59	1. 1. 2012 7:38:51	65	PP-V002-2012-0
16	Product 01	P001-2012	1. 1. 2012 7:41:33	1. 1. 2012 7:41:36	1. 1. 2012 7:42:34	95	PP-V001-2012-0
17	Product 03	P003-2012	1. 1. 2012 7:45:17	1. 1. 2012 7:45:17	1. 1. 2012 7:48:49	102	PP-V003-2012-0
18	Product 02	P002-2012	1. 1. 2012 7:52:11	1. 1. 2012 7:52:14	1. 1. 2012 7:55:52	115	PP-V002-2012-0
19	Product 03	P003-2012	1. 1. 2012 8:00:51	1. 1. 2012 8:00:55	1. 1. 2012 8:04:01	141	PP-V003-2012-0
20	Product 01	P001-2012	1. 1. 2012 8:09:08	1. 1. 2012 8:09:11	1. 1. 2012 8:10:42	62	PP-V001-2012-0
21	Product 02	P002-2012	1. 1. 2012 8:13:34	1. 1. 2012 8:13:38	1. 1. 2012 8:16:09	118	PP-V002-2012-0
22	Product 03	P003-2012	1. 1. 2012 8:20:16	1. 1. 2012 8:20:20	1. 1. 2012 8:21:55	94	PP-V003-2012-0
23	Product 01	P001-2012	1. 1. 2012 8:24:36	1. 1. 2012 8:24:39	1. 1. 2012 8:26:53	110	PP-V001-2012-0
24	Product 02	P002-2012	1. 1. 2012 8:30:32	1. 1. 2012 8:30:36	1. 1. 2012 8:32:58	134	PP-V002-2012-0
25	Product 03	P003-2012	1. 1. 2012 8:37:35	1. 1. 2012 8:37:36	1. 1. 2012 8:39:54	80	PP-V003-2012-0
26	Product 03	P003-2012	1. 1. 2012 8:43:00	1. 1. 2012 8:43:04	1. 1. 2012 8:47:02	137	PP-V003-2012-0
27	Product 03	P003-2012	1. 1. 2012 8:51:22	1. 1. 2012 8:51:25	1. 1. 2012 8:53:56	139	PP-V003-2012-0
28	Product 03	P003-2012	1. 1. 2012 8:59:04	1. 1. 2012 8:59:08	1. 1. 2012 9:01:27	77	PP-V003-2012-0
29	Product 03	P003-2012	1. 1. 2012 9:03:49	1. 1. 2012 9:03:50	1. 1. 2012 9:07:24	66	PP-V003-2012-0
30	Product 03	P003-2012	1. 1. 2012 9:11:21	1. 1. 2012 9:11:23	1. 1. 2012 9:12:41	125	PP-V003-2012-0
31	Product 03	P003-2012	1. 1. 2012 9:16:52	1. 1. 2012 9:16:53	1. 1. 2012 9:20:01	117	PP-V003-2012-0
32	Product 03	P003-2012	1. 1. 2012 9:25:12	1. 1. 2012 9:25:16	1. 1. 2012 9:28:10	66	PP-V003-2012-0

For Help, press F1 | Spreadsheet2 | C1,V1 | Product 03 | Sel:OFF | Weight:OFF | CAP | NUM | E

***Fig. 12** Read data in STATISTICA Data Miner Program*

Calculated and read data thus produces the input set for progress in the process of knowledge discovery in databases.

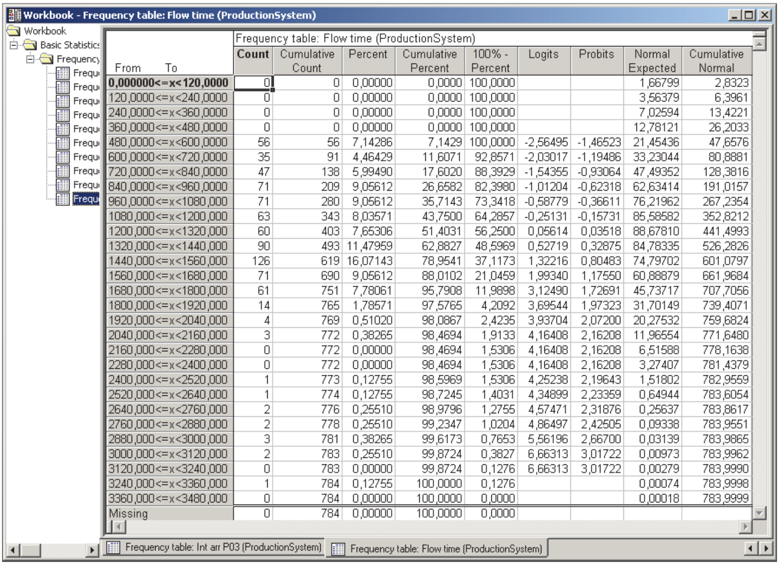
4.3 Data modification and transformation

The modification and transformation of data generated by the simulation model in the course of simulation was the next step in applying the process of knowledge discovery in databases for the proposed production system simulation model. The basic parameters necessary for the analysis were as follows: individual production batch sizes, production system input intervals, production device utilization, and flow time when the production batch remained in the production system. All of these parameters, except flow time, are recorded directly from the production system. Since the production process data is saved after one specific operation execution, it is possible to calculate the flow time of this data. The continuous time is the difference between the end of the last production operation time and the beginning of the first production operation time on the specific production batch. To calculate the value, we utilized the possibility of complementing one variable directly in the STATISTICA Data Miner application table, where the function for continuous time of the production calculation was defined. Since the database server RDBMS Oracle 10g was used for the proposed database, a database trigger was developed for automatically calculating the flow time of individual production batches and saving it directly into the database table called „Production_batchs“. The trigger eases data preparation for subsequent analyses using production batch „Flow_time“ parameter.

The format of the data saved in the database was in the required format, and therefore no further modification was needed.

The next step was to determine whether the set of data did not comprise data whose values significantly differ from the other data (outliers). It was necessary to find not only whether the set of data comprised such data but also the cause of their existence. They might be random data, whose extreme value was, for example, caused by a mistake in assigning. However, it could also concern significant values. Therefore, it was necessary to correctly identify the origin of the values and decide whether the values would be included or eliminated in the set of data used.

To identify the significantly different data a Frequency table was used. The table was used to identify all input parameters. All parameters, except the „Flow_time“ parameter, comprised the assumed extent of data and showed no symptoms of data of different values.


Workbook - Frequency table: Flow time (ProductionSystem)
Workbook
Basic Statistics
Frequency
Frequ
Frequency table: Int arr P03 (ProductionSystem)
Frequency table: Flow time (ProductionSystem)

Frequency table: Flow time (ProductionSystem)

From To	Count	Cumulative Count	Percent	Cumulative Percent	100% - Percent	Logits	Probits	Normal Expected	Cumulative Normal
0,000000<=x<120,0000	0	0	0,00000	0,0000	100,0000			1,66799	2,8323
120,0000<=x<240,0000	0	0	0,00000	0,0000	100,0000			3,56379	6,3961
240,0000<=x<360,0000	0	0	0,00000	0,0000	100,0000			7,02594	13,4221
360,0000<=x<480,0000	0	0	0,00000	0,0000	100,0000			12,78121	26,2033
480,0000<=x<600,0000	56	56	7,14286	7,1429	100,0000	-2,56495	-1,46523	21,45436	47,6576
600,0000<=x<720,0000	35	91	4,46429	11,6071	92,8571	-2,03017	-1,19486	33,23044	80,8881
720,0000<=x<840,0000	47	138	5,99490	17,6020	88,3929	-1,54355	-0,93064	47,49352	128,3816
840,0000<=x<960,0000	71	209	9,05612	26,6582	82,3980	-1,01204	-0,62318	62,63414	191,0157
960,0000<=x<1080,000	71	280	9,05612	35,7143	73,3418	-0,58779	-0,36611	76,21962	267,2354
1080,000<=x<1200,000	63	343	8,03571	43,7500	64,2857	-0,25131	-0,15731	85,58582	352,8212
1200,000<=x<1320,000	60	403	7,65306	51,4031	56,2500	0,05614	0,03518	88,67810	441,4993
1320,000<=x<1440,000	90	493	11,47959	62,8827	48,5969	0,52719	0,32875	84,78335	526,2826
1440,000<=x<1560,000	126	619	16,07143	78,9541	37,1173	1,32216	0,80483	74,79702	601,0797
1560,000<=x<1680,000	71	690	9,05612	88,0102	21,0459	1,99340	1,17550	60,88879	661,9684
1680,000<=x<1800,000	61	751	7,78061	95,7908	11,9898	3,12490	1,72691	45,73717	707,7056
1800,000<=x<1920,000	14	765	1,78571	97,5765	4,2092	3,69544	1,97323	31,70149	739,4071
1920,000<=x<2040,000	4	769	0,51020	98,0867	2,4235	3,93704	2,07200	20,27532	759,6824
2040,000<=x<2160,000	3	772	0,38265	98,4694	1,9133	4,16408	2,16208	11,96554	771,6480
2160,000<=x<2280,000	0	772	0,00000	98,4694	1,5306	4,16408	2,16208	6,51588	778,1638
2280,000<=x<2400,000	0	772	0,00000	98,4694	1,5306	4,16408	2,16208	3,27407	781,4379
2400,000<=x<2520,000	1	773	0,12755	98,5969	1,5306	4,25238	2,19643	1,51802	782,9559
2520,000<=x<2640,000	1	774	0,12755	98,7245	1,4031	4,34899	2,23359	0,64944	783,6054
2640,000<=x<2760,000	2	776	0,25510	98,9796	1,2755	4,57471	2,31876	0,25637	783,8617
2760,000<=x<2880,000	2	778	0,25510	99,2347	1,0204	4,86497	2,42505	0,09338	783,9551
2880,000<=x<3000,000	3	781	0,38265	99,6173	0,7653	5,56196	2,66700	0,03139	783,9865
3000,000<=x<3120,000	2	783	0,25510	99,8724	0,3827	6,66313	3,01722	0,00973	783,9962
3120,000<=x<3240,000	0	783	0,00000	99,8724	0,1276	6,66313	3,01722	0,00279	783,9990
3240,000<=x<3360,000	1	784	0,12755	100,0000	0,1276			0,00074	783,9998
3360,000<=x<3480,000	0	784	0,00000	100,0000	0,0000			0,00018	783,9999
Missing	0	784	0,00000	100,0000	0,0000				

Fig. 13 Identification of significantly different data – Frequency table

Data in the parameter called „Flow_time“ comprised a small group of data, approximately 1.5% of data from the entire set, with significantly higher times. They were subject to further investigation showing that they originated when in the simulation model a malfunction of a production device occurred. Considering the non-standard time of origin of quite a small amount of data, this data was eliminated from the set of data determined for the analysis.

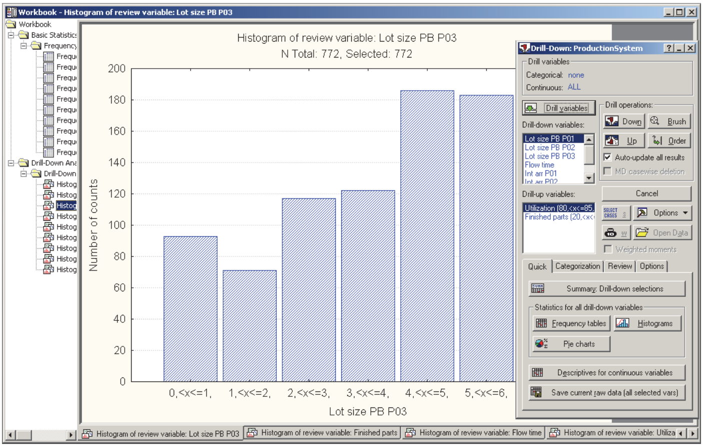
The set of data was then subjected to investigation as to whether it comprised the missing or noised data in some of its parameters. There was no data of the kind in the set of data examined.

This modified set of data produces the input set of data for further progress in the process of knowledge discovery in databases.

4.4 Data mining

Preparations for this specific analysis were finished. The next step in the process of knowledge discovery in databases is data mining.

After data modification and transformation an overview of the data analyzed was obtained. The STATISTICA Data Miner provides an „Interactive Drill Down Tool“, which produced this overview. It allows the development of simple 2D histograms of particular attributes and even 3D histograms for the simultaneous projection of an analyzed attributes couple, and then the basic information on the data analyzed, e.g. maximum and minimum values of particular attributes, average values, standard deviation, etc. The Interactive Drill Down Tool provides access to lower submerged levels, and hence the user can obtain a better overview of the particular data analyzed.



Categorical: none
Continuous: ALL

☑ Auto-update all results

__Fig. 14__ Interactive Drill Down Tool

This possibility was utilized in order to develop the first view of the data analyzed. At this level, the values arrangement for the particular attributes analyzed were found, e.g. their minimum and maximum values, dispersion of values, etc. This facilitated the advancement of the investigation.

To identify the data influencing the particular parameters monitored, „Feature selection and Variable screening“ model was utilized. This can identify and project the dependence of the particular parameters analyzed.

Considering the results achieved for the particular parameters analyzed, the parameters having the most significant influence on the parameters in question were identified. The size of individual production batches have always had the most influence on the three parameters analyzed. The following analysis is dealing with these parameters.

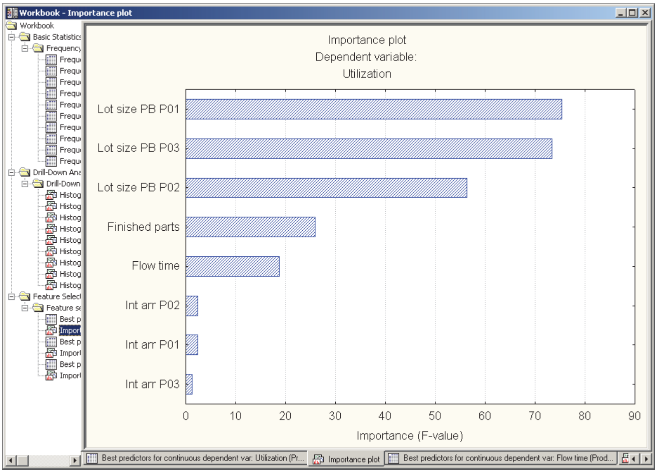


Fig. 15 Identification of dependences – Feature selection and Variable screening

The STATISTICA Data Miner application in version 9.1 provides two basic possibilities to develop a data-mining model, i.e. the use of Data Miner Recipes or Data Miner Workspace.

Data Miner Recipes provides the development of a data-mining model via a graphical guide. It also provides means to prepare the analyzed data, which can be further analyzed via the five basic mining models used by Classification and Regression Trees, Random Forest, Boosted Trees, Neural Networks and SVM (Support Vector Machines). It is possible to modify the particular model's parameters, and hence adjust them to the specific subject matter researched. Data Miner Recipes provide the projection of results obtained from the particular models in graphical, text or numerical modes.

The Data Miner Workspace affords the analyst even more possibilities in the mining model development. The analyst can select from more data resources, and then precisely specify the modification, cleaning and transformation of input data. The user can also select from a range of data mining methods and techniques, and other possibilities in the visualization and analysis of the achieved results. It is useful to utilize Data Miner Workspace for deeper and more complex data analyses. The variety of the possibilities in data mining model development offered provides the analyst with more data analysis possibilities as well as with more obtained results analysis.

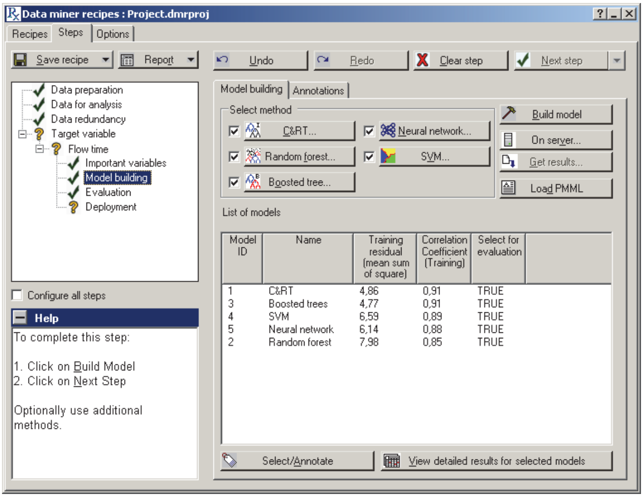
In both cases the models developed can be saved and used again for the analysis of another data set. The selection depends on the type of the investigated problem and the proposed solution methods.

In this case both possibilities for the analysis were utilized.

Analysis of production process parameters influence on production devices utilization

For the analysis of the production process parameters influence the Data Miner Recipes tool was first utilized. The use of the tool is suitable also for the less experienced users if the STATISTICA Data Miner application, since it provides a rich graphical interface that contains the basic components of the data mining process. It also helps the user to keep the sequence of the process steps (selection of input data set, modification, cleaning and transformation of the data set, data mining and verification of the model developed as well as the overview of the achieved results). At the same time, it provides the user with all necessary information on the specific activities needed in each process phase. It is appropriate to use the tool as a first step before the development of the complex data-mining model, so as to verify the suitability of using the provided models for the input data set.

A modified data set recorded from the information system database was used as the input data set, and it was complemented by the calculated flow times of individual production batches. It was not necessary to modify such a modified data set any further, and the definition of data mining process could be directly advanced to.



- ✓ Data preparation
- ✓ Data for analysis
- ✓ Data redundancy
- ? Target variable
  - ? Flow time
    - ✓ Important variables
    - ✓ Model building
    - ✓ Evaluation
    - ? Deployment

□ Configure all steps

To complete this step:

1. Click on Build Model
2. Click on Next Step

Optionally use additional methods.

Select method

☑ C&RT... ☑ Neural network...
☑ Random forest... ☑ SVM...
☑ Boosted tree...

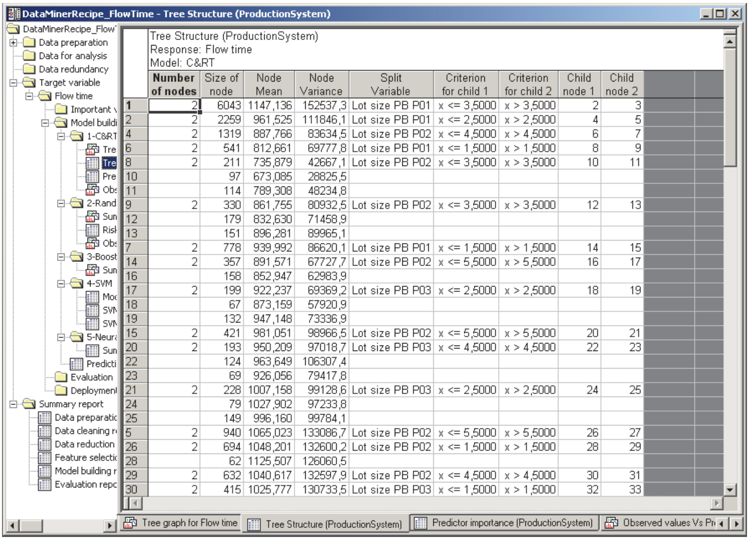
List of models

Model ID	Name	Training residual (mean sum of square)	Correlation Coefficient (Training)	Select for evaluation
1	C&RT	4,86	0,91	TRUE
3	Boosted trees	4,77	0,91	TRUE
4	SVM	6,59	0,89	TRUE
5	Neural network	6,14	0,88	TRUE
2	Random forest	7,98	0,85	TRUE

Fig. 16 Data Miner Recipes – Production devices utilization

All standard data mining models, provided by the Data Miner Recipes tool, were used for the input data analysis. Each model was modified according to the specific requirements of the analyzed problem and all other essential parameters of the data mining process were also determined.

After the data mining process ended, the Data Miner Recipes tool provided the results independently for all of the selected mining models. The results achieved were saved in a new „Workbook – Results" file for future detailed analyses. However, the Data Miner Recipes tool provides only a limited amount of output results of the mining process. The overview of the results achieved can be seen in the following figure.



Tree Structure (ProductionSystem)
Response: Flow time
Model: C&RT

	Number of nodes	Size of node	Node Mean	Node Variance	Split Variable	Criterion for child 1	Criterion for child 2	Child node 1	Child node 2
1	2	6043	1147,136	152537,3	Lot size PB P01	x <= 3,5000	x > 3,5000	2	3
2	2	2259	961,525	111846,1	Lot size PB P01	x <= 2,5000	x > 2,5000	4	5
4	2	1319	887,766	83634,5	Lot size PB P02	x <= 4,5000	x > 4,5000	6	7
6	2	541	812,661	69777,8	Lot size PB P01	x <= 1,5000	x > 1,5000	8	9
8	2	211	735,879	42667,1	Lot size PB P02	x <= 3,5000	x > 3,5000	10	11
10		97	673,085	28825,5					
11		114	789,308	48234,8					
9	2	330	861,755	80932,5	Lot size PB P02	x <= 3,5000	x > 3,5000	12	13
12		179	832,630	71458,9					
13		151	896,281	89965,1					
7	2	778	939,992	86620,1	Lot size PB P01	x <= 1,5000	x > 1,5000	14	15
14	2	357	891,571	67727,7	Lot size PB P02	x <= 5,5000	x > 5,5000	16	17
16		158	852,947	62983,9					
17	2	199	922,237	69369,2	Lot size PB P03	x <= 2,5000	x > 2,5000	18	19
18		67	873,159	57920,9					
19		132	947,148	73336,9					
15	2	421	981,051	98966,5	Lot size PB P02	x <= 5,5000	x > 5,5000	20	21
20	2	193	950,209	97018,7	Lot size PB P03	x <= 4,5000	x > 4,5000	22	23
22		124	963,649	106307,4					
23		69	926,056	79417,8					
21	2	228	1007,158	99128,6	Lot size PB P03	x <= 2,5000	x > 2,5000	24	25
24		79	1027,902	97233,8					
25		149	996,160	99784,1					
5	2	940	1065,023	133086,7	Lot size PB P02	x <= 5,5000	x > 5,5000	26	27
26	2	694	1048,201	132600,2	Lot size PB P02	x <= 1,5000	x > 1,5000	28	29
28		62	1125,507	126060,5					
29	2	632	1040,617	132597,9	Lot size PB P02	x <= 4,5000	x > 4,5000	30	31
30	2	415	1025,777	130733,5	Lot size PB P03	x <= 1,5000	x > 1,5000	32	33

Fig. 17 *Results from Data Miner Recipes tool*

The use of Data Miner Recipes does not provide a larger amount of more detailed results of the mining process, nor does it provide the definition of particular mining models. To conduct a more detailed and precise analysis Data Miner Workspace tool was utilized.

The proposed data-mining model in Data Miner Workspace tool was executed in the following three steps:

- data set selection serving as an input data set for the data mining process,
- modification and transformation of the selected data set,
- selection, definition and modification of the necessary parameters of particular data mining methods and techniques.

For the selection of the data set the modified and transformed set modified previously in the process was considered. The proposed model also utilized the possibilities provided by the model and the data was tested for missing values, values arrangement. A search for extreme values in the data was conducted and then the modification, arrangement and final transformation were completed. This modified and transformed data set served as the input data set for the data mining process.

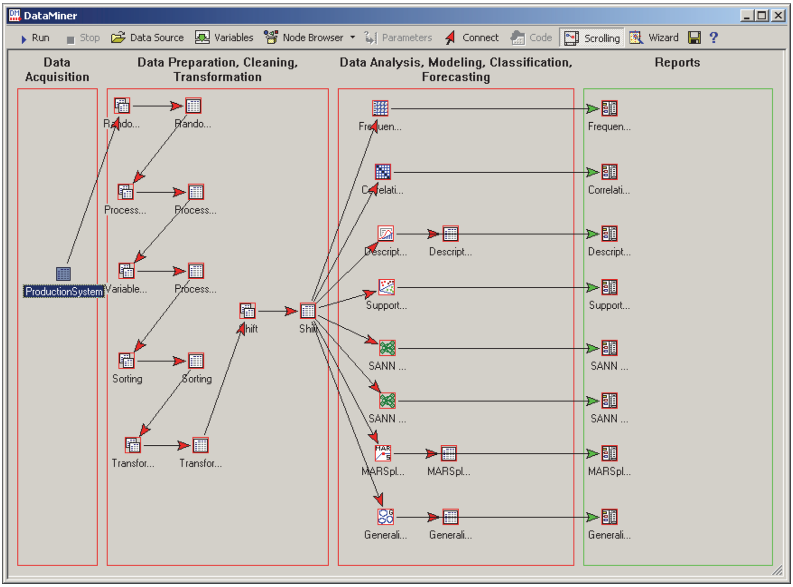


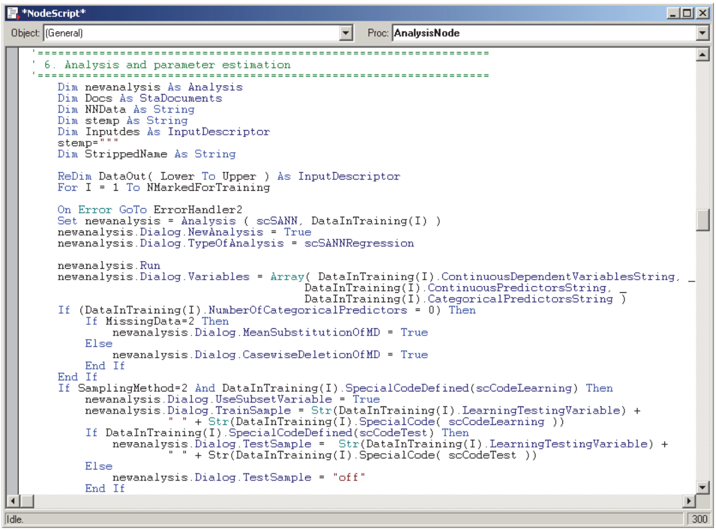
***Fig. 18** Proposed model in Data Miner tool*

The selection of data mining methods and techniques was the next step in the development of the data-mining model. The following data mining methods and techniques were included in the proposed model:

- STATISTICA Automated Neural Network Regression – Custom Neural Network,
- STATISTICA Automated Neural Network Regression – Automated Network Search,
- Generalized K-Means Cluster Analysis,
- MARSplines,
- SVM (Support Vector Machines).

The capacity to modify the individually used methods is an advantage of using the Data Miner tool. To define the model's particular parts STATISTICA Data Miner application utilizes STATISTICA Visual Basic programming language, which is a modification of Microsoft Visual Basic programming language. This capacity was used by several methods where it was not possible to set some of the specific properties via the graphical interface, and it was necessary to directly intervene in the resource code of the method used. This modified method was saved in user methods and later on used for other analyze issues.

The particular parts of the data-mining model were connected to functional units. To make a functional model, it was necessary to connect the input data resource to individual modules determined for the processing and modification of the set of data. Each module developed a modified data set. Data connected with the following module was determined for the data set modification. The resulting modified data set was connected to individual data-mining methods. This prepared and modified data-mining model was set for further activities.



```
'==================================================================
' 6. Analysis and parameter estimation
'==================================================================
    Dim newanalysis As Analysis
    Dim Docs As StaDocuments
    Dim NNData As String
    Dim stemp As String
    Dim Inputdes As InputDescriptor
    stemp=""" 
    Dim StrippedName As String

    ReDim DataOut( Lower To Upper ) As InputDescriptor
    For I = 1 To NMarkedForTraining

    On Error GoTo ErrorHandler2
    Set newanalysis = Analysis ( scSANN, DataInTraining(I) )
    newanalysis.Dialog.NewAnalysis = True
    newanalysis.Dialog.TypeOfAnalysis = scSANNRegression

    newanalysis.Run
    newanalysis.Dialog.Variables = Array( DataInTraining(I).ContinuousDependentVariablesString, _
                                    DataInTraining(I).ContinuousPredictorsString, _
                                    DataInTraining(I).CategoricalPredictorsString )
    If (DataInTraining(I).NumberOfCategoricalPredictors = 0) Then
        If MissingData=2 Then
            newanalysis.Dialog.MeanSubstitutionOfMD = True
        Else
            newanalysis.Dialog.CasewiseDeletionOfMD = True
        End If
    End If
    If SamplingMethod=2 And DataInTraining(I).SpecialCodeDefined(scCodeLearning) Then
        newanalysis.Dialog.UseSubsetVariable = True
        newanalysis.Dialog.TrainSample = Str(DataInTraining(I).LearningTestingVariable) +
                " " + Str(DataInTraining(I).SpecialCode( scCodeLearning ))
        If DataInTraining(I).SpecialCodeDefined(scCodeTest) Then
            newanalysis.Dialog.TestSample =  Str(DataInTraining(I).LearningTestingVariable) +
                " " + Str(DataInTraining(I).SpecialCode( scCodeTest ))
        Else
            newanalysis.Dialog.TestSample = "off"
        End If
```

Fig. 19 *Fragment of used neural network resource code*

After the data-mining model development the process of data mining began. After the initiation of the process the outcome reports for every employed data-mining method or technique were created. The reports were saved in the Workbook for subsequent detailed analyses. Individual reports consisted of text or numerical outcomes in the form of summary information on the executed process in the particular method, or of partial information on specific parts of the model or graphical outcome in graphs, e.g. histograms, mean graphs, Box and Whisker Graphs, etc. The format and type of individual outcome reports depended on the specific data mining method or technique.

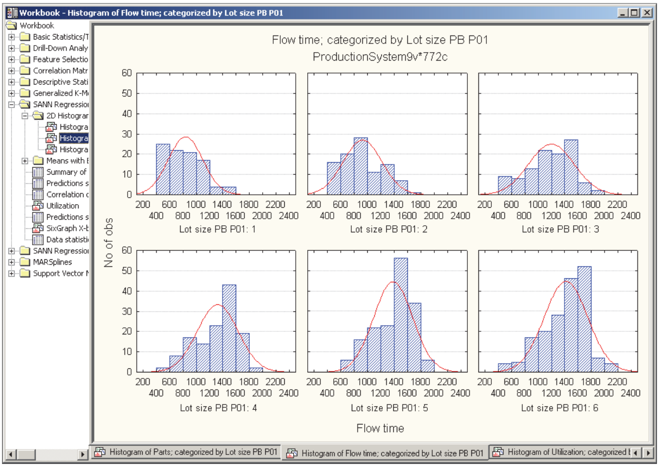


*Fig. **20** Histogram of production devices utilization in dependence on the production batch size*

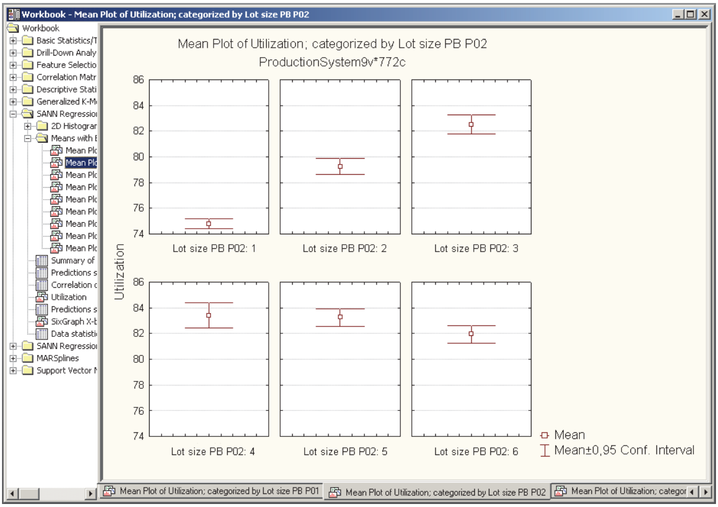
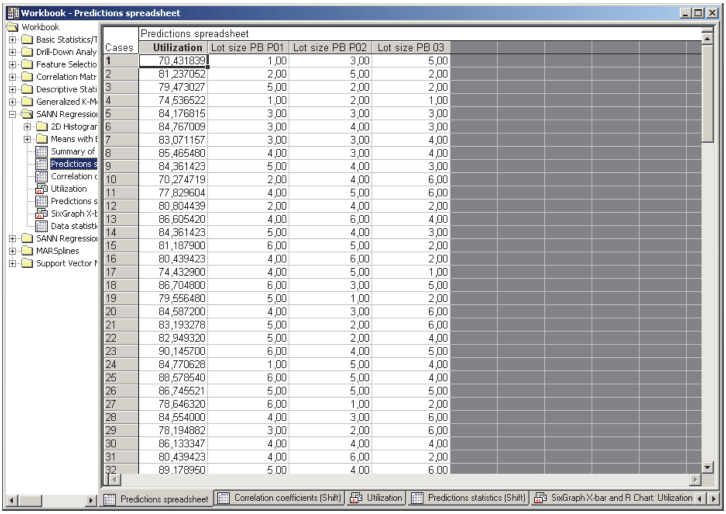


*Fig. **21** Relation of production batch size and production devices utilization*



Predictions spreadsheet

Cases	Utilization	Lot size PB P01	Lot size PB P02	Lot size PB 03
1	70,431839	1,00	3,00	5,00
2	81,237052	2,00	5,00	2,00
3	79,473027	5,00	2,00	2,00
4	74,536522	1,00	2,00	1,00
5	84,176815	3,00	3,00	3,00
6	84,767009	3,00	4,00	3,00
7	83,071157	3,00	3,00	4,00
8	85,465480	4,00	3,00	4,00
9	84,361423	5,00	4,00	3,00
10	70,274719	2,00	4,00	6,00
11	77,829604	4,00	5,00	6,00
12	80,804439	2,00	4,00	2,00
13	86,605420	4,00	6,00	4,00
14	84,361423	5,00	4,00	3,00
15	81,187900	6,00	5,00	2,00
16	80,439423	4,00	6,00	2,00
17	74,432900	4,00	5,00	1,00
18	86,704800	6,00	3,00	5,00
19	79,556480	5,00	1,00	2,00
20	84,587200	4,00	3,00	6,00
21	83,193278	5,00	2,00	6,00
22	82,949320	5,00	2,00	4,00
23	90,145700	6,00	4,00	5,00
24	84,770628	1,00	5,00	4,00
25	88,578540	6,00	5,00	4,00
26	86,745521	5,00	5,00	5,00
27	78,646320	6,00	1,00	2,00
28	84,554000	4,00	3,00	6,00
29	78,194882	3,00	2,00	6,00
30	86,133347	4,00	4,00	4,00
31	80,439423	4,00	6,00	2,00
32	89,178950	5,00	4,00	6,00

Fig. 22 *Results of proposed neural network*

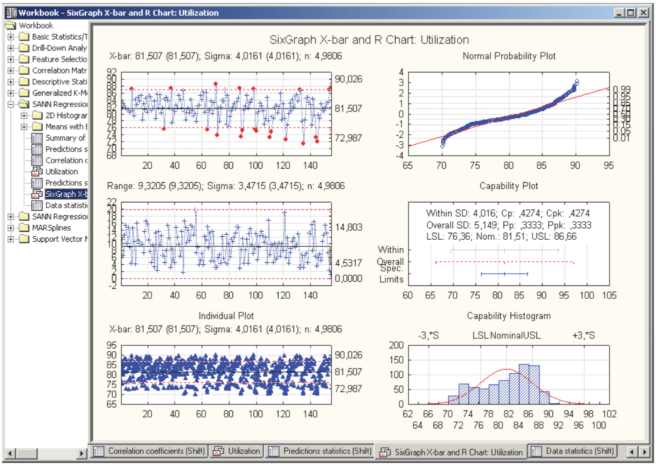


Fig. 23 *Results of proposed data mining model for production devices utilization*

Similarly, new data-mining models for other defined objectives were developed – the analysis of production process parameter influence on production batch flow times, as well as the analysis of the production process parameters influence on the number of manufactured products. All of the reports for each data-mining model were again independently saved in the Workbook. Once this work was completed, the analysis of the achieved results began.

4.5 Evaluation of discovered knowledge

The results were evaluated with respect to the data-mining outcome reports on particular data-mining methods and techniques. The evaluation was carried out independently of defined objectives. The results are discussed below.

Analysis of production process parameters influence on production devices utilization

The following knowledge is based on the results of the data mining model developed for the analysis. To ensure the increase of production device utilization it was necessary to select the following sizes of individual production batches:

- production batch size of Product 1 is six products in one production batch,
- production batch size of Product 2 is four products in a production batch,
- production batch size of Product 3 is five products in a production batch.

Analysis of production process parameters influence on flow times of production batches

The following knowledge is based on the results of the other data mining model developed for the analysis. To ensure the reduction of flow time the following production batch sizes are suitable:

- production batch size of Product 1 is one product in one production batch,
- production batch size of Product 2 is two products in a production batch,
- production batch size of Product 3 is two products in a production batch.

Analysis of production process parameters influence on number of manufactured products

The third data-mining model focused on the production process parameters and produced the following knowledge. To ensure the increase of the quantity of manufactured products, it is necessary to set up the sizes of the production batches as follows:

- production batch size of Product 1 is four products in a production batch,
- production batch size of Product 2 is five products in a production batch,
- production batch size of Product 3 is three products in a production batch.

To verify whether the discovered knowledge ensured the defined objectives, it is necessary to verify their correctness on the proposed simulation model of the production system.

5. DISCOVERED KNOWLEDGE APPLICATION IN THE PRODUCTION SYSTEM

To verify the discovered knowledge in applying the process of knowledge discovery in databases to the achieved results in the production system simulation model, the discovered knowledge in the proposed simulation model was applied. The objective was the acquisition of results from the modified production process and to compare them with the knowledge acquired in previous stages of the process.

In all cases the production process parameters were modified according to the knowledge obtained. The length of production process simulation was 30 days as was also the case in the previous simulation.

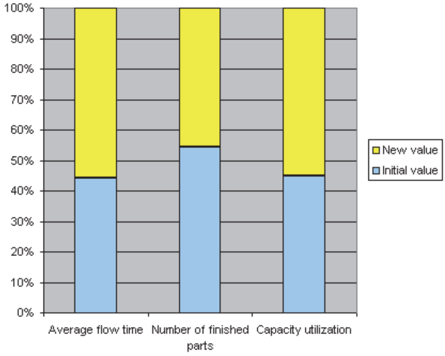
Analysis of production process parameters influence on production devices utilization

The sizes of production batches were set as follows:

- production batch size of Product 1 – 6 products in a production batch,
- production batch size of Product 2 – 4 products in a production batch,
- production batch size of Product 3 – 5 products in a production batch.

By modifying the production batch sizes it resulted in the change of the parameter required – production device utilization. The overall utilization of production devices rose from the original 73.31% to a new value of 90.37%. The discovered knowledge in the data-mining model was verified and by changing production batches sizes the utilization of production devices was increased.

The results acquired in the simulation were processed and are illustrated in the following figure.



	Initial value	New value	Difference
Average flow time	1329 s	1674 s	-25,96%
Number of finished parts	8746 pcs	7358 pcs	-15,87%
Capacity utilization	73,31%	90,37%	23,27%

***Fig. 24** Results of simulation model in 1st analysis*

At the same time, two other production process parameters changed. The number of manufactured products decreased and the average flow time increased. In contrast to the aforementioned improvement in the production devices utilization, the result of the change of these two parameters is a negative since both parameters worsened.

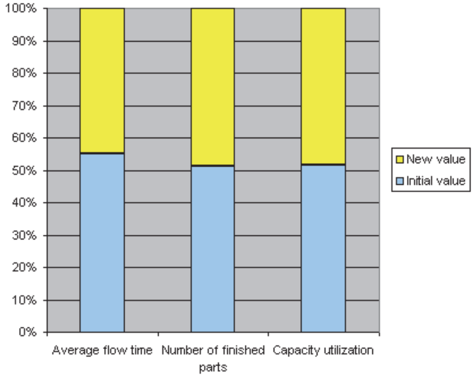
Analysis of production process parameters influence on flow times of production batches

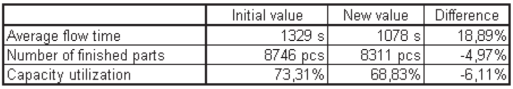
The sizes of production batches were set as follows:

- production batch size of Product 1 – one product in a production batch,
- production batch size of Product 2 – two products in a production batch,

- production batch size of Product 3 – two products in a production batch.

The results acquired in the simulation were processed and are illustrated in the following figure.





	Initial value	New value	Difference
Average flow time	1329 s	1078 s	18,89%
Number of finished parts	8746 pcs	8311 pcs	-4,97%
Capacity utilization	73,31%	68,83%	-6,11%

***Fig. 25** Results achieved in 2nd analysis simulation model*

By applying new production batches sizes the improvement of the monitored parameter – average flow time was achieved. The average flow time shortened from the original 1329 seconds to 1078 seconds. This case proved that the discovered knowledge in the data-mining model contributed to the improvement of the analyzed parameter.

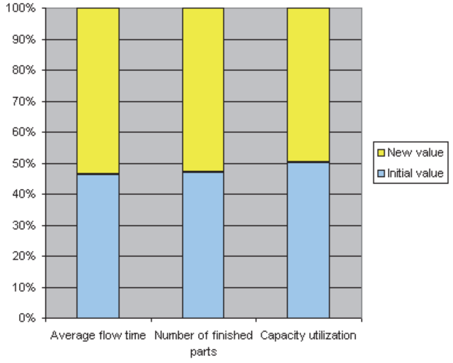
As in the first case, in this case the change of other production process parameters occurred. The overall number of manufactured products decreased by 5% and the production devices utilization fell by more than 6%.

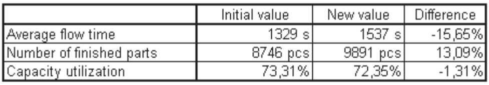
Analysis of production process parameters influence on number of manufactured products

The sizes of production batches were set as follows:

- production batch size of Product 1 – four products in a production batch,
- production batch size of Product 2 – five products in a production batch,
- production batch size of Product 3 – three products in a production batch,

The results acquired in the simulation were processed and are illustrated in the following figure.





	Initial value	New value	Difference
Average flow time	1329 s	1537 s	-15,65%
Number of finished parts	8746 pcs	9891 pcs	13,09%
Capacity utilization	73,31%	72,35%	-1,31%

***Fig. 26** Results achieved in 3rd analysis simulation model*

By changing production batch sizes the improvement of the monitored parameter – the number of manufactured products was observed. The number of manufactured products rose from original 8746 pieces to 9891 pieces. As in previous cases, this case proved that the discovered knowledge in the data mining model contributed to the improvement of the production process parameter analyzed.

Similarly, it also led to the change of two other monitored parameters. The average flow time rose from original flow time and the utilization of production devices decreased as well. Both parameters worsened.

Discovered knowledge evaluation

The verification of knowledge discovered in previous analyses resulted in proving their accuracy. In all three cases to the improvement of the production process parameter occurred.

At the same time, in all cases the worsening of the other monitored production process parameters was observed. Regarding the achieved results in particular analyses, another analysis was carried out. This analysis focused on the influence of particular production batch sizes on all three simultaneously monitored parameters – total flow time, number of manufactured products and production devices utilization.

Therefore a new data-mining model was proposed. The input set of data was the same as in the previous analyses. However, given the objective, the data mining model structure was changed. The selection of data mining methods and techniques corresponded to the methods and techniques verified in data mining models of the previous analyses. The proposed model included the following data mining methods and techniques:

- STATISTICA Automated Neural Network Regression – Custom Neural Network,
- STATISTICA Automated Neural Network Regression – Automated Network Search,

- STATISTICA Automated Neural Network Classification – Custom Neural Network,
- Generalized K-Means Cluster Analysis,
- Generalized EM Cluster Analysis,
- Generalized Additive Models,
- Standard Regression Trees (Classification and Regression Tree),
- Standard Regression CHAID,
- MARSplines,
- SVM (Support Vector Machines).

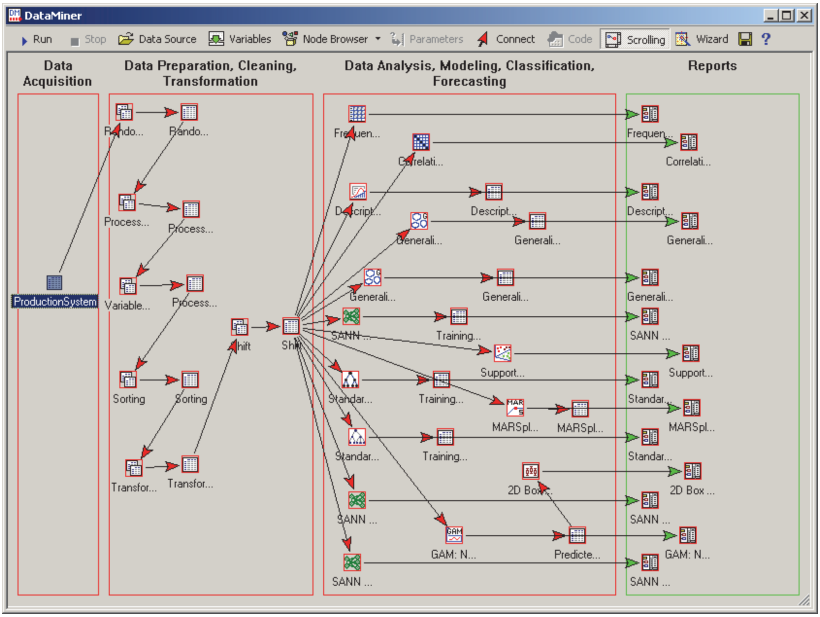
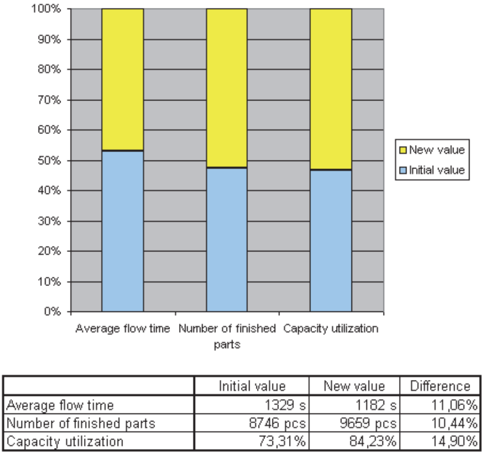


Fig. 27 Proposed model in Data Miner tool

The achieved results produced the following knowledge. To be able to modify the production process so that it does not result in the improvement of only one parameter and in the worsening of other production process parameters, it is necessary to select the following sizes of production batches for the particular products:

- production batch size of Product 1 is four products,
- production batch size of Product 2 is three products,
- production batch size of Product 3 is five products.

With respect to the knowledge discovered, the sizes of individual production batches in the simulation model were modified and another simulation was conducted so as to verify the findings. The results acquired in the simulation were processed and are illustrated in the following figure.



	Initial value	New value	Difference
Average flow time	1329 s	1182 s	11,06%
Number of finished parts	8746 pcs	9659 pcs	10,44%
Capacity utilization	73,31%	84,23%	14,90%

***Fig. 28** Results achieved in 4th analysis simulation model*

With respect to the results achieved in the simulation model the improvement of all three production process parameters was observed. In comparison to the results obtained in the analysis of individual production parameters, the improvement is not very significant. Nevertheless, the other monitored production process parameters did not worsen. Similarly, and as in previous cases, the analysis proved that the discovered knowledge contributed to the improvement of the analyzed production process parameters.

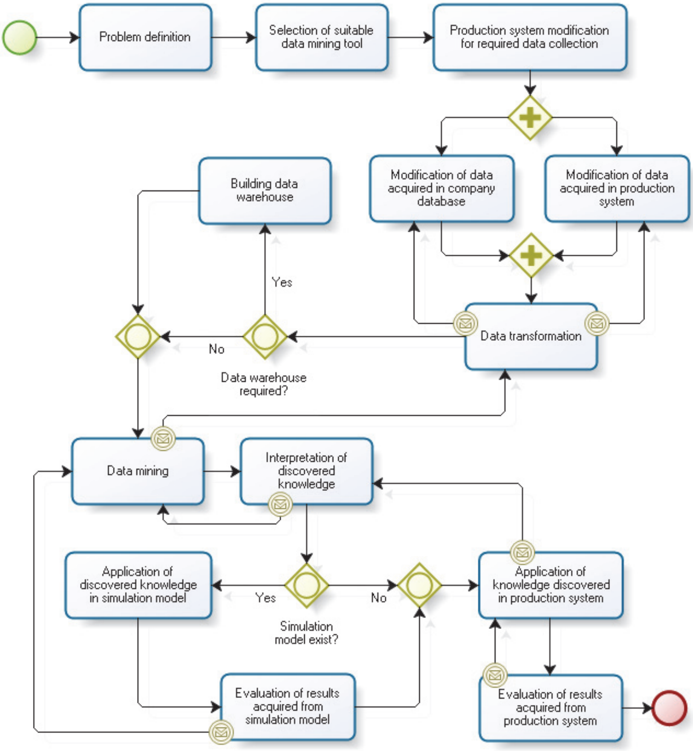
Discovered knowledge led to production process improvement and via their help the defined objectives were achieved. In some cases, the worsening of other production process parameters occurred. Therefore, it is necessary to determine the priorities of the particular objectives and with these in mind decide how the process of knowledge discovery in databases will be applied to the production process and which of the production process parameters will be monitored and processed.

Considering the fact that it is not possible to determine generally valid priorities for the particular objectives of the production process analysis, it is necessary to individually approach each specific case. If, for instance, it is necessary to produce a larger amount of products in a short time interval, it is necessary to assign this parameter a higher priority in the analysis. As a result this the discovery of knowledge is ensured and based on this the production process is modified so that the number of manufactured products is increased, even if this means that some of the other production process parameters worsen at the same time.

The determination of priorities in production process analyses is a very important component. It influences specific steps of the process of knowledge discovery in databases and hence also the knowledge discovered.

6. METHODOLOGY PROPOSAL OF A DATA MINING PROCESS IN INDUSTRY TO IMPROVE CONTROL

The methodology is generally determined and facilitates the processes in the specific environment, while at the same time it also determines the rules, conditions and factors for decisions. In this Chapter we evaluate the knowledge acquired so far, experience from previous research and propose a methodology that describes the data mining application in industry in detail.



Problem definition
Selection of suitable data mining tool
Production system modification for required data collection
Building data warehouse
Modification of data acquired in company database
Modification of data acquired in production system
Yes
No
Data warehouse required?
Data transformation
Data mining
Interpretation of discovered knowledge
Application of discovered knowledge in simulation model
Yes
No
Simulation model exist?
Application of knowledge discovered in production system
Evaluation of results acquired from simulation model
Evaluation of results acquired from production system

***Fig. 29** Application methodology of data mining in industry*

Problem definition

The defining of the problem is the important first step in the implementation of data mining. In the beginning it is necessary to determine the basic objective as well as partial objectives to be solved by the process of knowledge discovery in databases. These can be problems such as making the production process more efficient, the discovery of new knowledge in the production process, the identification of possible problems of the production process, making the production devices utilization more efficient, the reduction of production costs, etc. Some of the objectives can be contradictory, e.g. maximization of production device utilization and minimization in process production. It is therefore essential to define the priorities of the particular objectives. With respect to the objectives and priorities it is possible to elaborate further procedure.

To enhance the discovered knowledge in the production process data can be included, which is not directly acquired from the production system in the process. If the company has its own company information system, comprising, for example, data on employees, suppliers, customers and their orders, monitoring of customers' and suppliers' liabilities, production process planning, etc., it can use this data. From the data saved in the company information system it is possible to discover new behavioral knowledge and hence better prediction of future behaviours, e.g. trends development in products ordering, etc.

The defined objectives determine the subsequent steps – selecting a suitable tool for data mining and modifying the production system appropriately so that it can collect the necessary data and so on.

Selection of suitable data mining tool

Tool selection will reflect the defined data mining objectives. Their determination is based in the basic requirements of the „mining“ tool. Particular tools differ from one another in parameters, offered data mining methods and techniques, ways of discovered knowledge display and last but not least, the means of proliferation of the tool.

Some tools are quite narrowly specialized so as to solve a smaller set of problems and comprise a limited set of data mining methods and techniques. They can be used when the objectives determined and their investigation is more tools oriented and can cover the whole subject matter under analysis. It is also possible to utilize the combination of more tools to solve the partial problems. However, it is more demanding in terms of tool operation and input data format and modification because the input data formats for a particular tool can differ (e.g. various data unification forms, incapacity to cooperate with a data warehouse, inability to access the data in specific RDBMS types, etc.).

Price availability of the tool is another important criterion in tool selection. It is possible to obtain without monetary cost downloadable tools such as Weka (65) or TANAGRA (61) that contain many data mining methods and techniques, etc. Nevertheless, the majority of non-commercial tools are oriented to the solution of a smaller number of specific problems, and therefore they include a limited number of basic data mining methods and techniques. The other possibility is to utilize the commercially available tools providing a larger number of data mining methods and techniques than the non-commercial tools. They can be integrated with the widely used

RDBMS, or even possibly with data warehouses. They also provide the tools for data pre-preparation and transformation, results visualization, etc.

At this stage of the tool selection it is necessary to consider the subject matter under investigation, its extent, as well as the further possible utilization of the mining tool for other company problems.

After the results of this stage are considered, it is possible to change the production process and data modification so that they can be collected and analyzed.

Production system modification for required data collection

In order to work with the production system data, the production system must be able to produce, save and in a suitable format and make the data on its activities available. These properties are not suitable to all commonly used production devices. If the production device does not allow production data collection, a module for collecting the production data and ensuring that it is saved can be added to the production device. More straightforward devices are capable of acquiring and processes at a minimum basic data, such as bar code readers, industrial cameras or optical sensor gates. To ensure that the scanning and saving of a larger amount of data can be used e.g. intelligent conveyor belts or palettes, RFID (Radio Frequency Identification) codes readers, PLC (Programmable Logic Controller) devices, etc. can be employed.

For some production devices it is not possible to apply the complementary device for production data collection and saving. It is thus necessary to ensure the production data is recorded, for example, by an employee operating the production device.

Newer production devices, such as NC (Numerical Control) and CNC (Computer Numerical Controlled) devices, service robots and manipulators, devices with OPC (OLE for Process Control) architecture and so on, already have their own modules for production data collection and saving. In this case it should only what data and in which format will be collected and how they will be made available in the production database should be modified.

Modification of data acquired in company database

Data acquired in company database should be appropriately modified. If all of the necessary data is found in one database, then the modifications are simple and only result in the modification of a format, the changing of some definitions, the unification of some parameter values, or the aggregation of some values can be carried out.

If the data is taken from multiple data sources, such as relational or file databases, text files, etc., then besides basic modifications other modifications have to be executed as well. Common elements in different data sources have to unambiguously identified, their names unified and the exact rules for their valid identification defined.

Noised and inconsistent data then have to be removed and data significantly different from others in their value identified, etc. Identification of significantly different data (so called outliers) must be properly investigated. It can be random data, whose extreme value was caused, for example, by a mistake in setting. However, it can be important values caused, for example, by an over-standard order. Therefore, the origin of these values has to be identified and whether or not they will be included in the used data set or eliminated must be decided upon.

The result of this step directly influences the quality of further steps.

Modification of data acquired in production system

As in the case of the modification of data acquired from company information system, the data acquired from a production system has to be properly modified. The modification is similar as in the previous case.

There can be issues with the missing data. This could be due to the connection failure between the production device and database, or missing data in the production devices where the data cannot be collected, etc. It is necessary to deal with this issue – the missing data from temporary connection failure between the production device and database due to a collection system failure, or network connection failure caused, etc. can be complemented via the generation of similar data. Some data mining tools include a module that considers the data produced in the course of operation can generate missing data as well. They also respect the same production device and production process setups. The addition of medium values, or the most probable value, constant, or possibly a smaller amount of manually missing data represents another possibility.

If it is not possible to sufficiently complement the missing data with suitable data, then the issue in data mining and especially in the evaluation of the results achieved has to be considered.

Data transformation

In this process stage the data acquired in the production process and the data acquired in the company information system was merged. Their formats, names, etc. need to be unified in previous stages. If unification did not occur in the previous stage then it needs to be carried out now. It is essential from the perspective of accurate identification not only in the course of the data mining process itself, but also for the unambiguous identification in the evaluation of discovered knowledge. Then redundant data in the production system and company database has to be eliminated.

If the data mining method or proposed technique does not have to operate within the entire set of data, it is now possible to select a suitable subset of data to be used in the process. The used set of data with respect to the time interval can also be limited e.g. to the data of the last week, month, year, etc.

This transformation should be conducted with respect to the defined objectives and assumed methods and techniques that will be utilized in the process as well as with respect to the data-mining tool employed.

Building data warehouse

In practice one case often occurs – the specific data is saved in multiple databases. They can be independent databases administered by the same RDBMS, however, in many cases these databases are controlled by different RDBMS, or possibly the necessary data is in other forms, e.g. text files, MS Excel Program files, etc. In such cases, the process of data modification is much more complex. As a result there has to be a proposed mutual connection of independent data resources, which identify proper data, modify the format, unify data types of particular attributes, etc.

For the most part, the process leads to the construction of a new data depository in which the modified data will be saved. Typically it is a data warehouse representing a suitable place for data placing and saving. Data warehouses allow the accumulation and consolidation of analyzed data from relational databases of various RDBMS as well as from other resources, such as file databases, text files, etc. A retained history of saved data as well as their transparency and simple availability when operating with saved data is an advantage of using the data warehouses (20), (24), (34), (41).

Data can be suitably modified before being saved in the data warehouse. It is possible to modify the formats, unify the names, modify the same data aggregated from various resources, etc. Such modified data can be utilized in further applications of the process of knowledge discovery in databases.

The organization can utilize the data warehouse not only in applying the process of knowledge discovery in databases but also for other operational needs, or analyses of decisions as well as of a part of managerial systems, etc.

Data mining

Data mining represents the key step of the process. Regarding the objectives defined it is possible to build a model and apply relevant data mining methods and techniques to the model to acquire patterns of analyzed data.

Some methods and techniques require training and the verification of data set serving for learning, e.g. training and verification of a neural network and then apply this learned and verified neural network to selected data.

Data mining methods and techniques used are selected according to their application. Subsequently, the patterns achieved will be analyzed in detail.

Interpretation of discovered knowledge

Patterns achieved in the previous stages have to be analyzed and evaluated. Considering the objectives defined, not all the results achieved need to represent interesting and useful knowledge. Their correct evaluation is essential so that the analyst evaluating them has experience not only in the application and evaluation of the process, but also in the analyzed production system and activities running in related organization.

If the knowledge discovered is not interesting or useful enough, or when the discovered knowledge is unexpectedly useful, it is possible to return to previous steps and conduct process iteration. By returning to data mining it is possible to modify the data mining parameters or utilize another data mining method or technique. Also, it is possible return to the data transformation and modify this aspect of the process. Sometimes it is even necessary to return to the data selection and to modify the input data. These steps can lead to the discovery of new, interesting and useful knowledge or possibly a better comprehension of knowledge acquired in previous process iterations. Iterations can be carried out until the analyst is satisfied with the discovered knowledge in respect to the defined objectives.

The task of the analyst is to transform the discovered knowledge into specific solutions applicable to the production system.

Application of discovered knowledge in simulation model

The application of discovered knowledge in the real process is frequently accompanied b a certain risk. The risk means that the knowledge discovered is always connected with a certain level of probability. Therefore, if possible, it is useful to verify the executed modifications and changes on a simulated model of a production process. In contrast to using it in a real process, the use of a production process simulation model also brings another advantage – significant time factor elimination. To verify some new solutions the specific time in which the proposed new solution appears is needed and the use of the simulation model can shorten the time significantly.

Nowadays, there are many simulation tools that allow the construction of a simulation model and the evaluation of the results achieved, e.g. Witness by Lanner Group Company, Inc., Arena by Optimization Technologies Company, Inc., Simul8 by Visual Thinking International Company, Ltd., AutoMod by AutoSimulations Company, Inc., ProModel by ProModel Company Corp., etc. (71).

Firstly, the proposed simulation model has to be verified via the original production system setup. These results can be compared to the results produced by a real production system and thus discover to what extent the simulation model proposed corresponds with a real production system. Only then it is possible to apply the changes based on the discovered knowledge and proposed solutions.

Evaluation of results acquired from simulation model

Results acquired from simulation model can be compared to the results assumed with respect to the discovered knowledge. Based on these analyzes, it is possible to see how the changes are reflected and that they met defined objectives. If the results achieved do not correspond with the anticipated results, it is possible to execute certain changes in previous steps so as to improve the whole process. If the results achieved correspond with the anticipated results and meet the determined objectives, they can be applied in a real production system.

In the analysis of the achieved results it is necessary to take into account that the results are achieved from a simulation model extent to which the model corresponds with a real production system.

Application of knowledge discovered in production system

The application of acquired knowledge or proposed solutions in the production system is the last step in the application of data mining. The implementation of this stage can be carried out when the analyst is sufficiently satisfied with discovered knowledge, especially with respect to the defined objectives, and he is confident that the discovered knowledge and proposed solutions could be suitably applied to a real production system.

Evaluation of results acquired in production system

In contrast to the application of knowledge and solutions to the simulation model, in the application a longer time interval is required in the production system, so that the real contribution of the solution acquired can be evaluated.

The results can be compared to assumed results. With respect to these analyses it is possible to discover how the executed changes have appeared and whether the determined objectives have been met. If the results achieved do not correspond with the anticipated results, it is possible to implement certain changes in the previous steps so as to improve the process or to discover new knowledge in the production system.

7. CONCLUSION

This scientific monograph aims at proposing a suitable application process of knowledge discovery in databases in industry. The entire process was divided into several distinct stages.

In first stage the problems to be solved via the knowledge discovery in databases were identified (Chapter 2). The problems specifically depend on the field the process is applied in. There the most frequent problems encountered in the application of the process of knowledge discovery in databases in industry are also identified.

The next stage concerns the analysis of production process data (Chapter 4). In this Chapter it was argued that it should be based on defined objectives. Regarding the defined objectives the selection of a suitable KDD tool, methods and techniques helpful in the investigation are also mentioned. Data acquired in the proposed simulation model of production system (Chapter 3.1) and placed in the proposed relational database (Chapter 3.2) were analyzed via STATISTICA Data Miner, KDD tool. It was utilized since it sufficiently covers the entire process of knowledge discovery in databases from the access to data sources, through their modification, the data mining itself and right up to the visualization of the results achieved. Several data mining models were developed in which the data mining methods and techniques were applied in conjunction with analyzed input data and the subject matter investigated. The discovered knowledge was applied in the next stage.

To be able to discern to what extent the knowledge is interesting and useful, it was applied to the production system whose data served as input data to the process of knowledge discovery in databases. In this stage the discovered knowledge to the proposed simulation model of production system are applied (Chapter 5). The proposed simulation model was modified on the basis of discovered knowledge. After the application the achieved results were compared to the achieved results by the production system before the application of knowledge concerned. The achieved results proved that the discovered knowledge is useful and the modified simulation model had the assumed behaviour.

The proposal of the methodology of the process of knowledge discovery in databases in industry was the final stage of the monograph (Chapter 6). The proposed methodology provides particular steps of this processes' implementation. The methodology can help identify individual requirements and potential problems in the process, which can occur in the course of its application in industry.

7.1 Monograph contributions

The main contributions of the monograph can be summarized as follows:

- The identification of basic problems resolvable via the process of knowledge discovery in databases in industry,
- The application of the process of knowledge discovery in databases to the production process and new knowledge discovery on the process analyzes, evaluation of discovered knowledge and their subsequent application to the production process to verify the knowledge discovered,

- A methodology proposal of the process of knowledge discovery in databases in industry to improve control.

The application of the process of knowledge discovery in databases can help in the identification of the influence of production parameters on the production process and subsequent production process optimization. It was utilized in the expectation of malfunctions, emergencies or states, which can negatively influence the production process, and hence to discover knowledge helpful in the control of the process concerned. In the field of prediction, it was used also for the prediction of preventive checkups of production devices, or for production process costs, organization customers' behaviour, etc.

In the implementation of the process of knowledge discovery in databases into the production systems control can be improved through the following: inserting knowledge in the production system, better understanding the controlled system, and obtaining new and interesting knowledge predicting future behaviour of the production system.

Newly discovered knowledge can help managers in their decisions.

7.2 Future development prospects

The monograph mainly dealt with the data typically produced by the common production systems. Currently it is increasingly possible to deal with the integration of production databases with databases, which contain mainly business data organization. The data acquired in the production and business databases can provide new sources of potential knowledge. They can help understand the analyzed subject matter better as well as provide new knowledge. The proposed model can be enhanced, for example, by the prediction of and organization of customer behaviour. With respect to the data saved in the company information system, it is possible to obtain new knowledge about their behaviour and permit further prediction of the development of behaviours such as trends development in products ordering, etc.

If the sources of data are built by various data sources such as relational databases, file databases, or possibly text files, it is possible to utilize data warehouses in the process of knowledge discovery in databases. The data warehouse allows the integration of the data from various sources into one place. Before saving in the data warehouse, the data can be suitably modified, e.g. by the unification of attribute values, transformation and modification of data formats, etc. This provides further advantages in applying the process of knowledge discovery in databases for such a modified set of input data.

REFERENCES

1. AGRAWAL, R., PSAILA, G. Active Data Mining. In: *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95*), Menlo Park, Calif.: American Association for Artificial Intelligence, 1995.
2. ARLOW, J., NEUSTADT, I. UML a unifikovaný proces vývoje aplikací *(UML and Unified Process of Applications Development)*. Computer Press, 2003. ISBN 80-72269-47-X
3. ARLOW, J., NEUSTADT, I. UML 2 a unifikovaný proces vývoje aplikací *(UML 2 and Unified Process of Applications Development)*. Computer Press, 2007. ISBN 80-25115-03-8
4. BERKA, P. *Dobývaní znalostí z databází (Knowledge Discovery in Databases)*. Praha: Academia, 2003. ISBN 80-200-1062-9
5. BERKA, P., RAUCH, J., ŠIMUMEK, M. LISp-Miner: systém pro získávání znalostí z dat *(LISp-Miner System for Knowledge Discovery in Data)*. In: *Znalostný manažment (Knowledge Management)* 2007. VŠM Trenčín, 2007. ISBN 978-80-89306-02-2
6. BEZÁK, P., IRINGOVÁ, M. Optimization Using Genetic Algorithms. In: *Materials Science and Technology*, 2008, **8**(8). (online). [cit. 2012-04-15]. ISSN 1335-9053
7. CERRITO, P. *Introduction to Data Mining Using SAS Enterprise Miner*. SAS Press, 2006. ISBN 978-1-59047-829-5
8. DOVRTĚL, M. Data mining a jeho použití v komerční praxi (Data Mining and Its Use in Commercial Practice). In: *Moderní databáze* '99 *(Modern Database '99)*. Beroun, 1999.
9. ĎUĎÁK, J., GAŠPAR, G. Efficiency of Database Systems on Hardware Platforms fot Industrial Use. In *Dependability in Complex System Modelling*. Wroclaw: 2012, pp. 25-36. ISBN 978-83-7493-584-5
10. ĎUĎÁK, J., PAVLÍKOVÁ, S., GAŠPAR, G. Methods of Temperature Estimation on Given Area in System of Data Collection. In: *Proceedings of 14 Inernational Conference on Mechatronics*. Trenčianske Teplice, 2011, pp. 39-42. ISBN 978-80-8075-476-1
11. FAYYAD, U. M. Data Mining and Knowledge Discovery in Databases: Implication for Scientific Databases. Proc. In: *Ninth Int. Conf. on Scientific and Statistical Database Management*. Olympia Washington, Computer Society, 1997, pp. 2-11.

12. FAYYAD, U. M. Data Mining and Knowledge Discovery: Making Sense Out of Data. In: *IEEE Expert/Intelligent Systems & Their Applications*, 1996, (**11**) 5, pp. 20-26.
13. FAYYAD, U. M., PIATETSKI-SHAPIRO, G., SMYTH, P. From Data Mining to Knowledge Discovery: An Overview. In: *Advances in Knowledge Discovery and Data Mining*. California, Menlo Park, 1996AAAI Press/The MIT Press, pp. 1-34. ISBN 02-62560-97-6
14. FAYYAD, U. M., WEIR, N., DJORGOVSKI, S. Automated cataloging and analysis of sky survey image databases: the SKICAT system. In: *Conference on Information and Knowledge Management.* Proceedings of the 2nd International Conference on Information and Knowledge Management Washington, D.C. New York: Association for Computing Machinery, 1993, pp. 527-536. ISBN 0-89791-626-3
15. GIUDICI, P., FIGINI, S. *Applied Data Mining for Business and Industry*. 2nd Edition. Wiley Computer Publishing, 2009. ISBN 978-0-470-05887-9
16. GREISER, G., LANGE, S., MEMMEL, M. DaMiT: Ein adaptives Tutorsystem für Data-Mining *(Adaptive Tutor System for Data Mining)*. In: *Tagungsband Leipziger Informatik-Tage München: Akademische Verlagsgesellschaft*, 2003, pp. 192-203.
17. HALENÁR, I., ELIÁŠ, M. Using Neural Networks to Securing Communications Management Systems. In: *Proceedings of the International Workshop „Innovation Information Technologies: Theory and Practice“ Dresden*. Forschungszentrum Dresden – Rossendorf, 2010, pp. 36-40. ISBN 978-3-941405-10-3
18. HALENÁR, I., HALENÁR, R. The Self-organizing Maps and Their Exploitation by Designing and Controlling of an IDS. In: *International Conference Applied Natural Sciences 2009 (online)*. Editors Pipíška, M. Horník, M., Gajdošová, J. Univerzita sv. Cyrila a Metoda v Trnave *(University of Ss. Cyril and Methodius in Trnava)*, 2009, pp. 53-59. ISBN 978-80-8105-129-6
19. HALENÁR, R. Contribution of Near Real Time ETL. In: *2011 International Conference on Database and Data Mining* (ICDDM 2011). Editors: Steve Thatcher and Liu Guiping. Sanya. IEEE, 2011. pp. 243-247. ISBN 978-1-4244-9610-5
20. HALENÁR, R. Improved ETL process in KDD. In: *Aktuaľnyje problemy i innovacii v ekonomike, upravlenii, obrazovanii, informacionnych technologijach (Current Issues and Innovations in*

Economics, Administration, Education, and Information Technologies). Vol. 6, No. 2011. pp. 82-85. ISSN 2074-1685

21. HRNČIAR, M. CIMVIEW a CIMWORK – nástroje pre podnikovú integráciu *(CIMVIEW and CIMWORK – Tools for Company Integration)*. *AT&P Journal*, 2000, No. 5.
22. HUMUSOFT: WITNESS – simulace podnikových procesů *(Simulation of Company Processes)*. (online). [cit. 2012-04-15]. Available at <http://www.humusoft.cz/produkty/witness>
23. IBM: InfoSphere Warehouse. (online). [cit. 2012-04-15]. Available at<http://www-01.ibm.com/software/data/infosphere/warehouse/mining.html>
24. INMON, W. H. *Building the Data Warehouse.* Wiley Computer Publishing, 2002. ISBN 04-71081-30-2
25. JEŽEK, K., SOUKENÍK, K., TONCAR, V. Data Mining a získávání znalostí z databází. (Data Mining and Knowledge Discovery in Databases). *Softwarové noviny (Software Newspaper)*, 1998, No. 2, pp. 106-110.
26. JOHN, G. H. Enhancements To The Data Mining Process. Dissertation submitted to the Department of Computer Science and the Committee on graduate studies of Stanford University. Stanford, 1997.
27. KANISOVÁ, H., MÜLLER, M. UML srozumitelně *(Comprehensive UML)*. 2. aktualizované vydání *(2nd updated edition)*. Computer Press, 2006. ISBN 80-25110-83-4
28. KEBÍSEK, M. Využitie dolovania v riadení výrobných procesov *(Utilization of Data Mining in Production Processes Control)*. Písomná práca k dizertačnej skúške *(Written Thesis Submitted to Dissertation Examination)*. Trnava: MTF STU, 2003.
29. KEBÍSEK, M., ELIÁŠ, M. The Possibility of Utilization of Knowledge Discovery in Databases in the Industry. In: *Annals of MTeM for 2009 & Proceedings of the 9th International Conference Modern Technologies in Manufacturing.* Cluj-Napoca, Romania. Technical University of Cluj-Napoca, 2009. pp. 139-142. ISBN 97-37937-07-04
30. KEBÍSEK, M., SCHREIBER, P. The Possibility of Utilization of Neural Networks at the Data Mining. In: *CO-MAT-TECH 2004: 12th International Scientific Conference*. Trnava, Slovak Republic, ISBN 80-22721-17-4
31. KDNUGGETS: Data Mining Community‘s Top Resource. (online). [cit. 2012-04-15]. Available at <http://www.kdnuggets.com>

32. KREJČÍ, J. Automatizované získávání znalostí z dat *(Automated Knowledge Discovery in Data)*. In: *Moderní databáze '99 (Modern Database '99)*. Beroun, 1999.
33. LACKO, L. Business Intelligence v SQL Serveru 2008 *(Business Intelligence in SQL Server 2008)*. Reportovací, analytické a další datové služby *(Reporting, Analytical and Other Data Services)*. Computer Press, 2009. ISBN 978-80-251-2887-9
34. LACKO, L. Databáze: datové sklady, OLAP a dolování dat *(Database: Data Warehouses, OLAP and Data Mining)*. Computer Press, 2003. ISBN 80-72269-69-0
35. LACKO, L. Oracle – Správa, programování a použití databázového systému *(Oracle – Administration, Programming and Use of Database System)*. Computer Press, 2007. ISBN 978-80-251-1490-2
36. LANNER GROUP LIMITED: WITNESS – Bussines Simulation Software System. (online). [cit. 2012-04-15]. Available at <http://www.lanner.com/en/witness.cfm>
37. LISP–MINER: The official site of the LISp–Miner project. (online). [cit. 2012-04-15]. Available at <http://lispminer.vse.cz/index.html>
38. LUPA: Letní škola data miningu SPSS s lektory ČVUT *Summer School of OSPSS Data Mining with ČVUT Lecturers)*. (online). [cit. 2012-04-15]. Available at <http://www.lupa.cz/tiskove-zpravy/letni-skola-data-miningu-spss-s-lektory-cvut>
39. MACLENNAN, J., TANG, Z. H.,CRIVAT, B. *Data Mining with Microsoft SQL Server 2008*. Wiley Computer Publishing, 2006. ISBN 978-0-470-27774-4
40. MAKYŠ, P., KEBÍSEK, M. Možnosti spracovania dát pre podporu rozhodovania pri riadení výrobných systémov *(Possibilities of Data Processing to Support Deciding in Production Systems Control)*. In: Informačné technológie v manažmente výrobných systémov *(Information Technologies in Management of Production Systems)*. Medzinárodná vedecká konferencia *(International Scientific Conference)*. Nitra: SPU, pp. 144-149. ISBN 80-80693-64-1
41. MATIAŠKO, K., VNUK, L., ŠEVČÍKOVÁ, K. Dátové sklady ako informačný zdroj pre podporu rozhodovania *(Data Warehouses as Source of Information for Deciding Process Support)*. Žilina: FRI ŽU, 1999.
42. MATIGNON, R. *Data Mining Using SAS Enterprise Miner*. Wiley Computer Publishing 2007. ISBN 978-0-470-14901-0

43. MICROSOFT CORPORATION: SQL Server 2008 – Data Mining. (online). [cit. 2012-04-15]. Available at <http://www.microsoft.com/sqlserver/2008/en/us/data-mining.aspx>

44. ORACLE CORPORATION: Oracle Data Mining. (online). [cit. 2012-04-15]. Available at <http://www.oracle.com/technology/products/bi/odm/index.html>

45. PAGE–JONES, M. *Základy objektově orientovaného návrhu v UML (Basics of Object-oriented Proposal in UML)*. Praha: Grada, 2001. ISBN 80-24702-10-X

46. PANDIYAN, G. An Overview Of Knowlegde Discovery In Database (KDD) Process Towards Data Mining. (online). [cit. 2012-04-15]. Available at <http://www.articlesbase.com/computers-articles/an-overview-of-knowledge-discovery-in-database-kdd-process-towards-data-mining-1297459.html>

47. PAVLÍKOVÁ, S. Diagonálna podobnosť matíc. (Diagonal similarity matrix). In: *Academic Dubnica 98.* Bratislava: STU, 1998, Vol. 1. pp. 79-82. ISBN 80-227-111-X

48. PAVLÍKOVÁ, S. Konštrukcia grafu s minimálnym duálnym indexom (Construction of a graph with minimum dual index). In: *25. conference VŠTEP.* Trnava, pp. 283-287.

49. PMC: Witness Simulation Suite. (online). [cit. 2012-04-15]. Available at <http://www.pmcorp.com/Asprova/Witness_Simulation.asp>

50. POKORNÝ, J. OLAM = Skladování dat + OLAP + Dolování dat *(OLAM = Data Warehousing + OLAP + Data Mining)*. In: *Moderní databáze '99 (Modern Database '99)*. Beroun, 1999. pp. 90-99.

51. POPELÍNSKÝ, L. Knowledge Discovery in Spatial Data by Means of ILP. In: *Zytkow J.M*., Quafafaou M.(eds.): Principles of Data Mining and Knowledge Discovery. Porc. of 2nd Eur.Symposium, PKDD'98, Nantes, France. LNCS 1510, Springer Verlag, 1998.

52. RAKOTOMALALA, R.,TANAGRA. A Free Software for Research and Academic Purposes. In: *Proceedings of EGC,* Paris, France, 2005.

53. SAS INSTITUTE INC.: Data mining with SAS Enterprise Miner. (online). [cit. 2012-04-15]. Available at <http://www.sas.com/technologies/analytics/datamining/miner>

54. SCHMULLER, J. *Myslíme v jazyku UML (We Think in UML Language)*. Praha: Grada, 2001. ISBN 80-24700-29-8

55. SIPINA RESEARCH: *Classification Trees Software*. (online). [cit. 2012-04-15]. at <http://eric.univ-lyon2.fr/~ricco/sipina>

56. SPSS INC.: PASW Modeler. (online). [cit. 2012-04-15]. Available at <http://www.spss.com/software/modeling/modeler>
57. STATSOFT: STATISTICA Data Miner. (online). [cit. 2012-04-15]. Available at <http://www.statsoft.com/products/statistica-data-miner>
58. STATSOFT ČR: STATISTICA Data Miner. (online). [cit. 2012-04-15]. Available at <http://www.statsoft.cz/produkty/5-dataminingove-nastroje/21-statistica-data-miner>
59. ŠTOFKO, M., OCELÍKOVÁ, E. Získavanie znalostí z databáz a ich využívanie vo vizualizácii informačných a riadiacich systémov (1) (Knowledge Discovery in Databases and Their Utilization in Information and Control Systems Visualisation (1). *AT&P Journal,* HMH s.r.o Bratislava, 2005, No. 6, pp. 53-54.
60. ŠTOFKO, M., OCELÍKOVÁ, E. Získavanie znalostí z databáz a ich využívanie vo vizualizácii informačných a riadiacich systémov (2) (Knowledge Discovery in Databases and Their Utilization in Information and Control Systems Visualisation (2). *AT&P Journal*, 2005, No. 7, pp. 94-96.
61. TANAGRA: A free data mining software for teaching and research. (online). [cit. 2012-04-15].Available at <http://chirouble.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>
62. TANG, Z. H., MACLENNAN, J. *Data Mining with SQL Server 2005.* Wiley Computer Publishing, 2005. ISBN 978-0-471-46261-3
63. TANUŠKA, P., KEBÍSEK, M., SCHREIBER, P.,VAŽAN, P. *Využitie RUP a UML pri tvorbe softvérových projektov (RUP and UML Utilization in Software Projects Development)*. Trnava: Tripsoft, 2007. ISBN 978-80-89291-10-6
64. THE OBJECT MANAGEMENT GROUP. (online). [cit. 2012-04-15]. Available at <http://www.omg.org>
65. THE UNIVERSITY OF WAIKATO: Weka 3. – Data Mining with open source machine learning software in Java. (online). [cit. 2012-04-15]. Available at <http://www.cs.waikato.ac.nz/ml/weka>
66. TRNKA, A.: Classification and Regression Methods with Data Mining Algorithms. *Computer Technology and Application,* 2011. Vol. 2, No. 3, pp. 227-231. ISSN 1934-7332
67. TRNKA, A. Market Basket Analysis with Data Mining Methods: Six Sigma methodology improvement. In: *2010 International Conference on Networking and Information Technology.* Editors: Thatcher, S., Yi, X. Piscataway. IEEE, 2010, pp. 446-450. ISBN 978-1-4244-7577-3

68. TRNKA, A. Results of application data mining algorithms to (Lean) Six Sigma methodology. In: *Annals of Faculty Engineering Hunedoara.* International Journal of Engineering, 2012, **10**(1), pp. 141-144. ISSN 1584-2665
69. TRNKA, A. RFM Analysis as a Part of DMAIC Phases. In: *2011 International Conference on Database and Data Mining* (ICDDM 2011). Editors: Thatcher, S., Guiping, L. Sanya. IEEE, 2011, pp.278-281. ISBN 978-1-4244-9610-5
70. TRNKA, A. Six Sigma Methodology with Fraud Detection. In: *Advances in Data Networks, Communications, Computers. 9th WSEAS International Conference on Data Networks, Communications, Computers. University of Algarve*. Portugal: Faro, WSEAS Press, 2010, pp. 162-165. ISBN 978-960-474-245-5
71. VAŽAN, P. Simulačná optimalizácia – jej možnosti a problémy (Simulation Optimization – Its Possibilities and Issues). In: *Witness 2006, 9. ročník mezinárodní konference, Česká republika (9th International Conference, Czech Republic)*. Brno: Vysoké učení technické v Brně*)*, 2006, pp. 25-31. ISBN 80-21431-98-9
72. VAŽAN, P., SCHREIBER, P. KRIŽANOVÁ, G. Simulation Optimization with the Witness Simulator. In: *Ecumict 2008: Proceedings of the Third European Conference on the Use of Modern Information and Communication Technologies*. Gent, Nevelland v.z.w., 2008, pp. 461-469. ISBN 978-908082-553-6
73. VAŽAN, P., SCHREIBER, P., TANUŠKA, P. *Modelovanie a simulácia systémo. (Modelling and Simulation of Systems)*. Trnava: Tripsoft, 2005. ISBN 80-96939-02-5
74. VAŽAN, P. , SCHREIBER, P., TANUŠKA, P. The opportunities and problems of simulation optimization. In: *MOSIS '06 : Proceedings Spring International Conference.* 40th. Modelling and Simulation of Systems. Přerov, Ostrava: MARQ, 2006, pp. 59-65. ISBN 80-86840-21-2
75. VŠE: Vysoká škola ekonomická v Praze *(VŠE: University of Economics, Prague)*. (online). [cit. 2012-04-15]. Available at <http://www.vse.cz>

LIST OF PUBLICATIONS

BEZÁK, P., IRINGOVÁ, M., KEBÍSEK, M., URCIKÁN, M. Motion planning algorithms for general closed-chain mechanisms. (Algoritmy plánovania pohybu všeobecných mechanizmov s uzavretým reťazcom). In: *CO-MAT-TECH 2006. 14. medzinárodná vedecká konferencia (14th International Conference)* (Trnava, 19-20 Oct 2006). STU Bratislava, 2006. pp. 100-102. ISBN 80-227-2472-6

ĎUĎÁK, J., GAŠPAR, G., KEBÍSEK, M. Application of modified MODBUS protocol for the needs of modern measurement systems. In: Advances in Mechatronics 2011 : Proceedings of the 6th International Conference on Advances in Mechatronics 2011 (AIM'11), Brno, ČR. University of Defence, 2011. pp. 9-14. ISBN 978-80-7231-848-3

ELIÁŠ, M., ĎURČI, M., KEBÍSEK, M., MAKYŠ, P. Team development of database application. In: *CO-MAT-TECH 2001: 9. medzinárodná vedecká konferencia (9th International Conference)*. Trnava, 25-26 Oct 2001. Volume 2. STU Bratislava, 2001. pp. 245-250. ISBN 80-227-1591-3

ELIÁŠ, M., KEBÍSEK, M. An Overview of Methods for 3D Model Reconstruction from 2D Orthographic Views. In: *Proceedings of the International Workshop „Innovation Information Technologies: Theory and Practice"*. Forschungszentrum Dresden – Rossendorf, 2010. pp. 65-69. ISBN 978-3-941405-10-3

ELIÁŠ, M., KEBÍSEK, M. Checking position using laser scanners. In: *Annals of MTeM for 2009 & Proceedings of the 9th International Conference Modern Technologies in Manufacturing:* 8-10 Oct 2009, Cluj-Napoca, Romania. Technical University of Cluj-Napoca, 2009, pp. 91-94. ISBN 973-7937-07-04

ELIÁŠ, M., KEBÍSEK, M. *Transport table position checking using laser scanners. In: Process Control 2008: Proceedings of the 8th International Scientific-Technical Conference.* Kouty nad Desnou, Czech Republic, 9-12 Jun 2008. University of Pardubice, 2008. ISBN 978-80-7395-077-4

HALENÁR, I., KEBÍSEK, M., KUNÍK, S. Bezpečnostné riziká informačných technológií v ASR. (Security risks of information technology in ASR). In: *Process Control 2006 7th International Scientific-Technical Conference. K*outy nad Desnou, Czech Republic. University of Pardubice, 2006, pp. R155-1/R155-5. ISBN 80-7194-860-8

KEBÍSEK, M. Knowledge discovery in databases and data mining tools. In: *International Doctoral Seminar 2010 Smolenice.* Trnava: AlumniPress, 2010, pp. 240-249. ISBN 978-80-8096-118-3

KEBÍSEK, M. *Object-oriented programming.* Qintec s.r.o., Trnava, 2010, 120 p. ISBN 978-80-969846-8-8

KEBÍSEK, M., ELIÁŠ, A., BEZÁK, P. Text mining a jeho využitie v praxi (*Text mining and utilization in practice)*. In: *Process Control 2006: 7th International Scientific-Technical Conference.* Kouty nad Desnou, Czech Republic, 13-16 Jun 2006. University of Pardubice, 2006. R159-1/R159-5. ISBN 80-7194-860-8

KEBÍSEK, M., ELIÁŠ, M. The possibility of utilization of knowledge discovery in databases in the industry. In: *Annals of MTeM for 2009 & Proceedings of the 9th International Conference Modern Technologies in Manufacturing: 8-10 Oct 2009, Cluj-Napoca, Romania.* Technical University of Cluj-Napoca, 2009, pp. 139-142. ISBN 973-7937-07-04

KEBÍSEK, M., MAKYŠ, P. Získavanie znalostí z databáz *(Knowledge Discovery in Databases)*. In: *1st International Conference on Applied Mathematics and Informatics at Universities 2001.* STU Bratislava, 2001, pp. 212-217. ISBN 80-227-1568-9

KEBÍSEK, M., MAKYŠ, P., Získavanie znalostí z databáz a možnosti ich využitia v riadení výrobných procesov. (Knowledge discovery in database and possibilities of they usage in industrial process control). In: *Process Control 2004*: Kouty nad Desnou, Czech Republic. Univerzita Pardubice, 2004. R300/1-6. ISBN 80-7194-6621

KEBÍSEK, M., SCHREIBER, P. The possibility of utilization of neural networks at the data mining. In: *CO-MAT-TECH 2004*: 12th. CO-MAT-TECH. Trnava, Slovak Republic, 14-15 Oct 2004. STU Bratislava, 2004. pp. 589-595. ISBN 80-227-2117-4

KEBÍSEK, M., SCHREIBER, P., HALENÁR, I. Knowledge Discovery in Databases and its Application in Manufacturing. In: *Proceedings of the International Workshop „Innovation Information Technologies: Theory and Practice*“. Forschungszentrum Dresden – Rossendorf, 2010, pp. 204-207. ISBN 978-3-941405-10-3

KEBÍSEK, M., TANUŠKA, P. Možnosti využitia genetických algoritmov pri získavaní údajov z rozsiahlych databáz *(The utilization possibility of genetic Algorithms at the knowledge discovery in large databases)*. In: *Materials Science and Technology*, **4**(3), 2004. (online). [cit. 2004-08-09]. ISSN 1335-9053

KEBÍSEK, M., TANUŠKA, P., ELIÁŠ, M. Návrh a implementácia dátového skladu pre potreby MTF STU Trnava (Design and implementation of data warehouse for FMT SUT Trnava). Vega 1/4078/07). In: *Materials Science and Technology,* 2008, **8**(7). (online) [cit. 2008-12-01]. ISSN 1335-9053

MAKYŠ, P., KEBÍSEK, M. Možnosti spracovania dát pre podporu rozhodovania pri riadení výrobných systémov. (Data processing possibilities decision support in production system control). *In: Informačné technológie v manažmente výrobných systémov (Information Technologies in Management of Production Syst*ems): Medzinárodná vedecká konferencia *(International Scientific Conference)*. Nitra: Slovenská poľnohospodárska univerzita, 2004. pp. 144-149. ISBN 80-8069-364-1

MAKYŠ, P., KEBÍSEK, M. Softvérová podpora vyhodnocovania výsledkov meraní vypínačov (Software support of results evaluation of breakers measurements). In: *CO-MAT-TECH 2002. 10. medzinárodná vedecká konferencia. (10th International Scientific Conference)* (Trnava, 24-25 Oct 2002): Vol. 2. Manažment a kvalita. Aplikované prírodné a inžinierske vedy *(Management and Quality. Applied Natural and Engineering Sciences*. Bratislava: STU, 2002. pp. 324-328. ISBN 80-227-1768-1

TANUŠKA, P., KEBÍSEK, M. The knowledge discovery in huge databases. In: *CO-MAT-TECH 2005: 13th International Scientific Conference,* Trnava, Slovak Republic, 20-21 Oct 2005. STU Bratislava, 2005. pp. 1175-1180. ISBN 80-227-2286-3

TANUŠKA, P., KEBÍSEK, M.,MORAVČÍK, O., VAŽAN, P. The Proposal of Data Warehouse Validation. *Computer Technology and Application*, 2011, **2**(8), pp. 650-657. ISSN 1934-7332

TANUŠKA, P.,KEBÍSEK, M., SCHREIBER, P., VAŽAN, P.: Využitie RUP a UML pri tvorbe softvérových projektov *(RUP and UML Utilization in Software Projects Development)*. 1. vyd *(Edition 1).* Trnava: Tripsoft, 2007. ISBN 978-80-89291-10-6

TANUŠKA, P., MAKYŠ, P., KEBÍSEK, M. Automatizácia vyhodnocovania meraní integrálnej netesnosti kontajnerov C-30 (Automation the evaluation of integral measurements of leakage of containers C-30). *Materials Science and Technology,* 2003, **3**(2). (online). [cit. 2003-10-07]. ISSN 1335-9053

TANUŠKA, P., VAŽAN, P., KEBÍSEK, M., MORAVČÍK, O. SCHREIBER, P. Data Mining Model Building as a Support for Decision

Making in Production Management. In: *Advances in Intelligent and Soft Computing*. ISSN 1867-5662. Vol. 166. Advances in Computer Science, Engineering and Applications. Proceedings of the Second International Conference on Computer Science, Engineering and Applications (ICCSEA 2012), New Delhi, India, Volume 1. Springer-Verlag Berlin Heidelberg, 2012, pp. 695-701. ISBN 978-3-642-30156-8

VAŽAN, P., KEBÍSEK, M., TANUŠKA, P., JUROVATÁ, D. The data warehouse suggestion for production system. In: *Annals of DAAAM and Proceedings of DAAAM Symposium Vienna*, 2011, **22**(1), pp. 0017-0018. ISBN 978-3-901509-834

VAŽAN, P., TANUŠKA, P., KEBÍSEK, M. The data mining usage in production system management. *World Academy of Science, Engineering and Technology*, 2011, Year 7, Issue 77, pp. 1304-1308. ISSN 2010-376X

VOSTERMANS, A., TANUŠKA, P., VERSCHELDE, W., KEBÍSEK, M. A Student-oriented University Ontology. In: *WORLDCOMP'10 : The 2010 World Congress in Computer Science*, Computer Engineering and Applied Computing. Las Vegas Nevada, CSREA Press, 2010. pp. 228-231. ISBN 1-60132-138-4

ZEMAN, J., TANUŠKA, P., KEBÍSEK, M. The Utilization of Metrics Usability To Evaluate The Software Quality. In: *ICCTD 2009: International Conference on Computer Technology and Development.* 13-15 Nov 2009, Kota Kinabalu, Malaysia. Los Alamitos: IEEE Computer Society, 2009, pp. 243-246. ISBN 978-0-7695-3892-1

CONTENTS