2017 21st International Conference on Process Control (PC)
June 6–9, 2017, Štrbské Pleso, Slovakia

The initial analysis of failures emerging in production process for further data mining analysis

Nemeth M., Michalconok G.
Slovak University of technology in Bratislava, Faculty of Materials Science and Technology in Trnava,
Institute of Applied Informatics, Automation and Mechatronics
Trnava, Slovakia
martin.nemeth@stuba.sk, german.michalconok@stuba.sk

***Abstract*—The aim of this paper is to examine possibilities for the initial data analyses of the failure data from industrial production process. To perform the initial data analysis of the data from production process we have used graphical statistical method and also data mining methods like drill-down analysis and cluster analysis. Before applying mentioned techniques and methods it was necessary to know the principle of the industrial production process itself and also to be aware of the failure data structure. This initial data analysis is vital to be able to review the knowledge potential of given data. Based on this, we are able to point out interesting issues, that can be further solved with KDD (knowledge discovery from databases) techniques.**

***Keywords*—data mining, clustering, industrial production processes**

I. Introduction

The first step to obtain any knowledge from the large data set is the initial data analysis. This analysis is preceded by a proper understanding of the data. It is therefore necessary to know the principle of the researched process, from which the data is obtained. Next step is then determining the structure of the data and analyzing these data with the use of statistical analysis methods. The objective of the data mining is not always clear at the beginning of the analysis. However, based on the initial analysis, it is possible to determine what kind of relevant knowledge can be acquired from given data and then to establish the main objective of discovering knowledge from the data.

Data from the manufacturing process can be obtained easily, in most cases, because the data are automatically generated by the sensors which are installed on devices and monitor operation of the production process. This data is stored in pre-structured database from where they can then be exported using an appropriate interface in various formats, according to user needs. Understanding the data from which the previously unknown knowledge is to be discovered a key step to get a correct representation of this new knowledge. It is therefore necessary to understand the structure of the collected data and to know the data types of the data contained in the data set. This step is important in the process of selecting appropriate methods of data mining and also in the process of transforming data to form of correct input into the data mining algorithms.

Selection of correct data mining is closely related to the initial data analysis. To meet the stated objective, it is possible to use several different methods. This depends on the structure of the data and on the requirements of a particular method of data mining on the input data.

II. Background of the used methods

A. Data mining

One of the definitions of the concept of data mining: "Data mining is the process of analyzing data from different perspectives and their conversion into useful information. From the mathematical and statistical point of view it comes to finding correlations, thus interrelationships or patterns in the data "[1]. Friedman has described data mining process as a set of methods used to discover relationships in data in large data bases. The process of data mining overlaps at some degree with artificial intelligence, machine learning, pattern recognition and data visualization [2]. There are also other definitions, that describe data mining, however they which largely depend on the purpose of usage of the data mining methods. Data mining is a process divided into multiple steps, which are shown in the Figure 1.

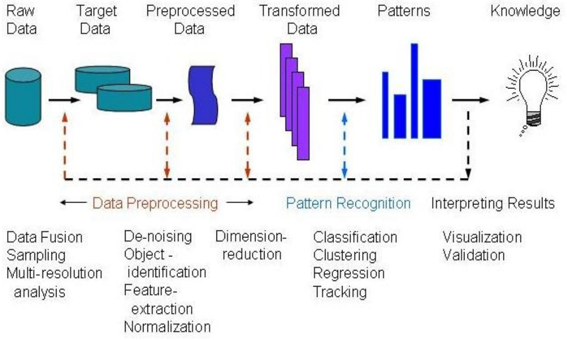


Fig. 1. Data mining scheme that describes the stages of data mining process from raw data to obtained knowledge [3].

Data mining process can be used to solve various types of tasks. All of these tasks are either predictive data mining tasks or descriptive data mining tasks. A data mining system can execute one or more of the below specified tasks as part of data mining:

978-1-5386-4011-1/17/$31.00 ©2017 IEEE

- Exploratory analysis - this analysis is an objective review of the data without prior knowledge.
- Descriptive tasks - describe the dataset. These tasks are using method of clustering similar phenomena into separate clusters.
- Predictive role – The aim of this task is to predict future behavior on the base of the data set.
- Search patterns and rules – This data mining task is looking for patterns and relationships between data in a set.
- Search by the model – The aim of this task is to search the data from the data set that are similar to the searching pattern by which the search is performed [3].

B. *Clustering*

Clustering can be understood as the unsupervised classification of patterns like observations, data items, or feature vectors, into groups. The clustering problem can be found many contexts and is used by researchers in many disciplines. Clustering is therefore broadly appeal and useful as one of the steps in exploratory data analysis. However, clustering is a difficult problem. Cluster analysis is a group of various algorithms. This group is considered as one of the methods of data mining. Their aim is to automatically partition given data into a set of partitions. These partitions are also called clusters. These clusters are characterized in that the data in them are similar to each other. The similarity between the data in each cluster is most often defined by the Euclidean distances. In our case, the Euclidean distances were computed between failure types in the data set:

Types of cluster analysis can be divided into different categories based on different criteria [4,5].

- Hard clustering: A given data point in n-dimensional space only belongs to one cluster. This is also known as exclusive clustering. The K-Means clustering mechanism is an example of hard clustering.
- Soft clustering: A given data point can belong to more than one cluster in soft clustering. This is also known as overlapping clustering. The Fuzzy K-Means algorithm is a good example of soft clustering.
- Hierarchical clustering: In hierarchical clustering, a hierarchy of clusters is built using the top-down (divisive) or bottom-up (agglomerative) approach.
- Flat clustering: Is a simple technique where no hierarchy is present.
- Model-based clustering: In model-based clustering, data is modeled using a standard statistical model to work with different distributions. The idea is to find a model that best fits the data.

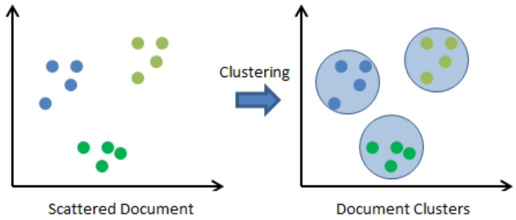
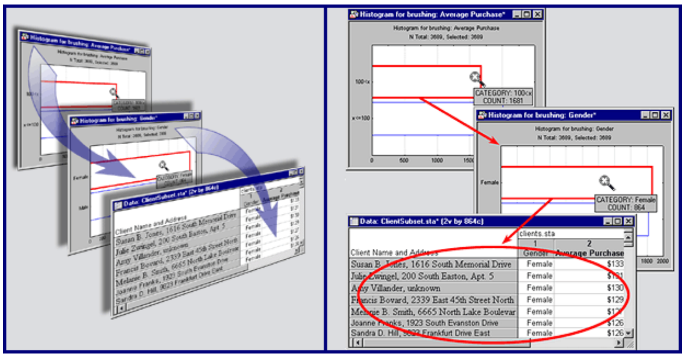


Fig. 2. Cluster analysis scheme that describes the process of grouping raw data into clusters with similar members.

C. *Drill-down analysis*

A first step in solving data mining tasks is often the process of exploration the data interactively to gain a first impression of the types of variables in the analyses, and their possible relationships. The concept of drill-down analysis applies to the field of data mining as the method of interactive exploration of the data, in particular of large databases. The process of interactive drill-down analysis begins by considering some simple break downs of the data by a few variables of interest. It is possible to compute various statistics, tables, histograms, and other graphical summaries for each group. Then, it is possible to "drill-down" to expose and further analyze the data "underneath" one of the categorizations. The advantage of the drill-down analysis is the various auxiliary results that can automatically be updated during the interactive drill-down/up exploration. Outputs of this analysis can be divided into following groups:

- Descriptive statistics and frequency tables
- Box-and-whiskers plots summarizing the distributions of continuous variables
- Scatterplot matrices summarizing the relationships between continuous variables;


Histogram for brushing: Average Purchase*
Histogram for brushing: Average Purchase
N Total: 3609, Selected: 3609
100<x
x<=100
CATEGORY: 100<x
COUNT: 1681
0
500
1000
1500
Histogram for brushing: Gender*
Histogram for brushing: Gender
N Total: 3609, Selected: 3609
Female
CATEGORY: Female
COUNT: 864
1800
2000
Data: ClientSubset.sta* (2v by 864c)
clients.sta

Client Name and Address	1 Gender	2 Average Purchase
Susan B. Jones, 1616 South Memorial Drive	Female	$133
Julie Zwingel, 200 South Easton, Apt. 5	Female	[illegible]
Amy Villander, unknown	Female	$130
Francis Bovard, 2339 East 45th Street North	Female	$129
Melanie B. Smith, 6665 North Lake Boulevar	Female	[illegible]
Joanne Franks, 1923 South Evanston Drive	Female	$126
Sandra D. Hill, [illegible] Frankfurt Drive East	Female	$126

Fig. 3. The demonstration of process of drill-down analisis in STATISTICA 13 software [6].

III. THE OBJECT OF INVESTIGATION AND DATA STRUCTURE

A. The object of investigation and its failures

We have chosen the process of assembly of the rear wheel as the object of investigation for our research. The assembly line can by characterized as semi-automatic. This means that some of the workplaces require a human operator on the position. In this paper, we are dealing with failure data from this semi-automatic assembly line. There are four pre-defined failure types defined as follows:

- Emergency stops: These failures require immediate emergency stop, which applies to the whole line.
- Delayed stops: These failures cause emergency stop of the line at the end of the cycle. These failures must be repaired by the operator before restarting.
- General alarms: These failures do not cause stopping of the line. These are informational messages for the operator. Operator do not have to remove this failure.
- Missing initial conditions: These failures are notifications for the operator, which say that the operation is not in expected position.

B. Data structure

The failure dataset produced by the assembly line consists of 620635 records. These records were acquired in a time frame of six months. Each of these records consists of several parameters. These parameters are: start date and time (in dd/mm/yyyy hh/mm/ss format), end date and time (same format as start date and time), duration, localization and the name of the failure. Localization points to the workplace where the failure occurred, start date and time carries the exact information about when the failure occurred in the system, end date and time says when exactly the failure state was solved, duration is the difference between start and end date and time and it is in hh/mm/ss format and name describes the failure itself. These raw data were exported into a .xls file format.

C. Derived parameters

The data structure of the raw data has several parameters, which are describing the emerged failure in some degree, however it has not very big initial knowledge potential. To be able to mine previously unknown knowledge from this data, it was necessary to derive new parameters based on the knowledge about the object to supplement the failure dataset. Three new derived parameters were computed from the existing mentioned initial parameters. These parameters were day, working team and shift. It is clear, that the raw data was not sufficient to make predictions of emerging failures based on the technical information from the sensors in the process, because this information is missing in given dataset. However, it is possible to examine the data from different point of view and to study other possible parameters that may have impact on emerging of the failures.

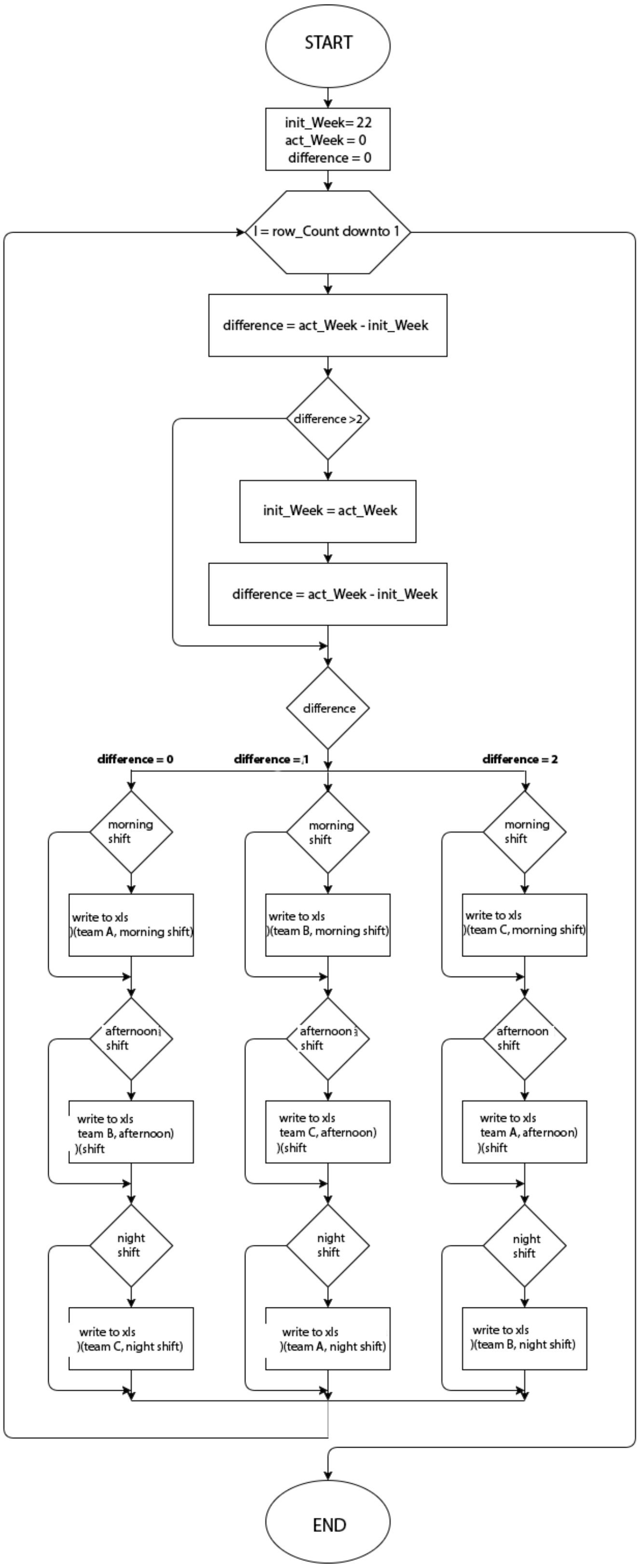

START
init_Week= 22
act_Week = 0
difference = 0
I = row_Count downto 1
difference = act_Week - init_Week
difference >2
init_Week = act_Week
difference = act_Week - init_Week
difference
difference = 0
difference = 1
difference = 2
morning shift
write to xls)(team A, morning shift)
morning shift
write to xls)(team B, morning shift)
morning shift
write to xls)(team C, morning shift)
afternoon shift
write to xls team B, afternoon))(shift
afternoon shift
write to xls team C, afternoon))(shift
afternoon shift
write to xls team A, afternoon))(shift
night shift
write to xls)(team C, night shift)
night shift
write to xls)(team A, night shift)
night shift
write to xls)(team B, night shift)
END

Fig. 4. The flowchart that represents the algorithm of computing derived parameters working team and working shift

To be able to derive some of these new parameters additional information about the production process were needed. We have discovered, that the production is running non-stop for six days a week (from Monday to Saturday). The assembly takes place in three working shifts (morning, afternoon and night) with three working teams (A, B and C). These working teams cycle through these shifts in a backwards manner. This means, that if working team A has worked one week in a morning shift, next week they will be working in a night shift. The first derived categorical parameter day was derived from the parameter start date and time parameter in MS Excel with the use of TEXT function. The remaining two derived categorical parameters working team and working shift were computed with the use of the programming language C#. In C# it is possible to use COM objects to communicate with and manipulate the .xls files. The figure 4 is showing the algorithm of computing these derived parameters. After computing these parameters, it was possible to perform the initial data analysis with chosen data mining methods.

IV. DATA MINING

After getting the information about the object of investigation, collecting the necessary data and after deriving possible parameters, data mining methods and knowledge discovery from the data can be applied to the data. The process of data mining does not only apply the selected algorithms to the dataset, but it also consists some necessary preparatory stages. These phases are primarily aimed at preparing the data for the actual application of selected data mining algorithms. This phase is primary aimed on adaptation and transformation of the data to a proper form, which can serve as a correct input for the data mining algorithms. After the phase of data adjustments it is possible to proceed to applying chosen methods and algorithms of data mining. Since the process of data mining is an iterative process, it is possible to return to previous phases (for example data transformation phase), in order to successfully apply other data mining algorithms. Sometimes, it is possible to return in the process of data mining to even earlier phases, for example like deriving new parameters.

Before we use any of data mining method, the drill-down analysis is possible and useful to use.

A. Drill-down analysis

Drill-down analysis was performed in the software Statistica 13. This analysis goes through selected parameters and examines the impact that have selected parameters to each other. At the beginning of the drill-down analysis it is needed to choose one of the parameters, which will serve as the first break-down. As the first output of the analysis, the histogram of the selected first parameter was displayed. Subsequently, based on the choice of the category of the first parameter a histogram of other parameters is displayed, but this already has a relationship with the previous selection. Before applying the so-called drill-down analysis, we have selected relevant data from the dataset, which we then subjected to the analysis. From initial data file, we have excluded all records that fell into categories of states or alarms with using filtering in MS Excel. Thus, the analysis was performed only for records that fell into the category of faults. Following plots captures drill down analysis with the selection of the initial parameter with the name Start date.

The drill-down analysis module has split the date range of occurrence of the events throughout the data file into several intervals. As shown in figures 1 to figure 5, this module represents the intervals, which were originally in the format of date as a numeric value. The following table shows the date intervals converted from numeric format to a date format using MS Excel.

TABLE I.

THE CONVERTATION OF NUMERIC REPRESENTATION OF THE DATA INTERVALS

Numeric representation of dating intervals	Representation by the date format
42700 - 42750	26.11.2016 - 15.1.2017
42650 - 42700	7.10.2016 - 26.11.2016
42600 - 42650	18.8.2016 - 7.10.2016
42550 - 42600	29.6.2016 - 18.8.2016
42500 - 42550	10.5.2016 - 29.6.2016
42450 - 42500	21.3.2016 - 10.5.2016
42700 - 42750	26.11.2016 - 15.1.2017

The first output from the drill-down module was histogram of frequencies of errors within the date intervals to which drill-down module divides the entire date range from the data file. As can be seen from the graph, the greatest number of emerged failures are located between 05/10/2016 to 06/29/2016. The next step in the drill-down analysis is to manually select the part of the graph, which is interesting for us. In our case, we chose disorders just mentioned in dating range.

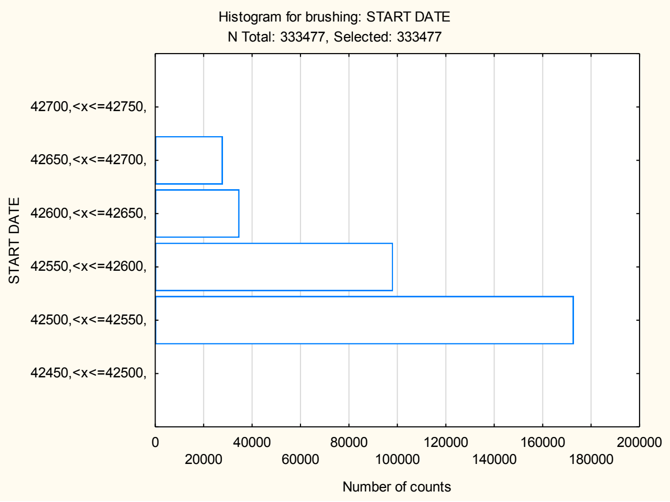


Fig. 5. The histogram shows the distribution of the emerged failures with respect to the parameter START DATE.

The following chart depicts the distribution of the following parameters, in this case the parameter weekday, in the selection of the first step of the drill-down analysis. The chart thus shows the distribution of the number of failures emerging after various days of the week between 05/10/2016 and 29/06/2016. To continue to the next step, it was necessary to make a further selection from the new plot. It is obvious that the days with the biggest number of emerged failures for the selected period are Tuesday and Wednesday. Although the frequency of failure for these days were very similar, in this step, we have selected Wednesday as the most abundant presence of emerged failures.

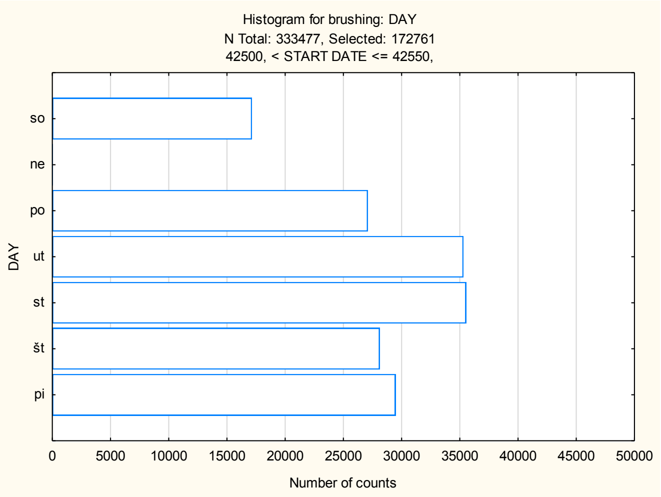


Fig. 6. The second histogram above shows the distribution of the emerged failures with respect to the parameter DAY within the selection of date range from the figure 1.

The next level of drill-down analysis shows the distribution of the parameter Working team. This distribution is shown just as in the previous graph with respect to the selection of the previous level. The graph's clear that the three working teams were, as regards the fault frequency, approximately equal. Working Group C showed as well as on initial statistical analysis the largest number emerged failures. Therefore, in the selection in this level of drill-down analysis, we have chosen the group C as a basis for the next level of the analysis.

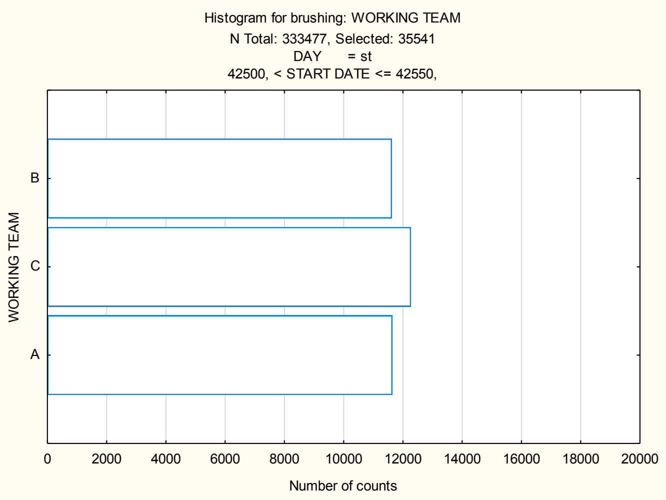


Fig. 7. The histogram shows the distribution of the emerged failures with respect to the parameter WORKING TEAM within the selection from figure 2.

The following graph shows the distribution of the number of faults that occured within the individual working shifts, ranging from 10/05/2016 to 29/06/2016. These failures emerged on Wednesdays, when the working team C was working on the assembly line. It can be also seen that in the context of individual work shifts, there is no noticeable difference in the frequency of emerging failures. Yet we have chosen night shift for the last drill-down iteration, because it had the greatest number of faults that occurred during this shift.

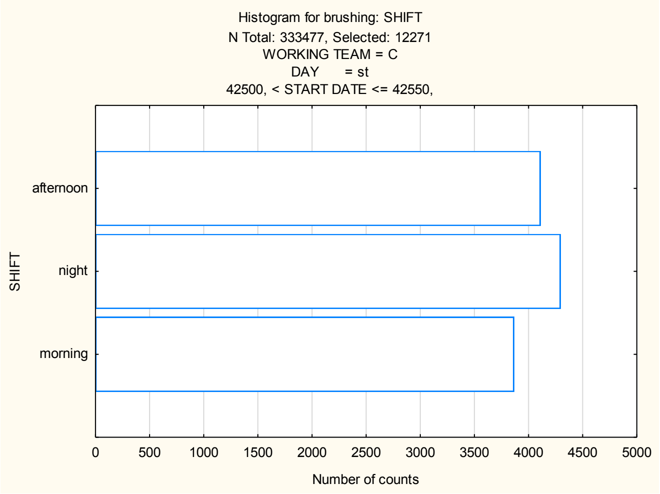


Fig. 8. The histogram shows the distribution of the emerged failures with respect to the parameter SHIFT within the selection from the figure 3.

The latest graphical plot of selected drill-down analysis shows the distribution of the number of occurring failures based on a selection from previous levels of the drill-down analysis. From the figure 5, it is clear that the largest representation, with the number of emerged failures of around 3500 has a failure called *Error on reverse rotation sensor SQP + PI35 on OP35*. Where SPQ + PI35 is the name of the sensor and OP35 denotes the number of the operation in the process.

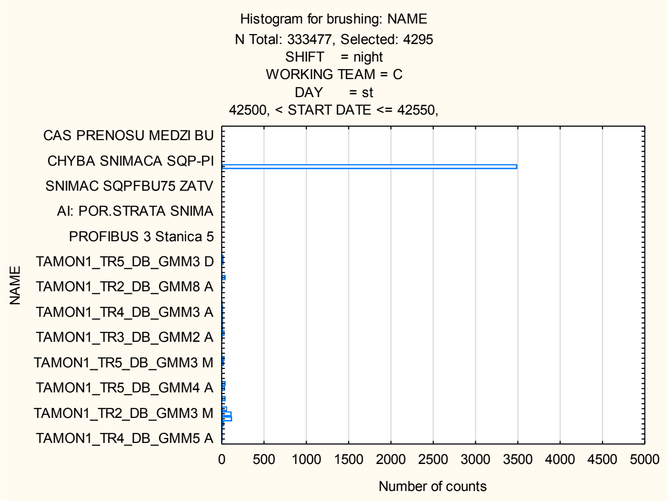


Fig. 9. The last histogram shows the distribution of the emerged failures with respect to the parameter NAME within the selection from the figure 4.

CONCLUSION

This paper dealt with the initial phase of data analysis with basic data mining method called drill-down analysis. This analysis was performed on real failure data from production process from automotive industry. The object of investigation was a small assembly line where the assembly of the rear wheel is performed.

In practice, a huge amount of data is collected every day. This data is however often only stored in databases and no analysis is performed with it. In many cases, there can be found valuable new knowledge in this data, but the knowledge potential is not always clear from the first sight. To discover this potential, it is necessary to examine the data with the use of basic methods like statistical analysis, or drill-down analysis, to discover the initial relationships between various parameters in the dataset.

When performing data mining or initial data analysis, it is also important to be aware of the principle of the object of investigation. Based on this knowledge it is then possible to derive new parameters from the dataset, which can expand the knowledge potential of the data for further applying data mining methods.

In future work, we would like to examine and optimize algorithms to find repetitive patterns in the data about merging failures in the system. These patterns can subsequently identify group and order of failures which are emerging one after another. The aim of our research is to be able to find this kind of related failures in the system even without extensive dataset with detailed technical parameters.

ACKNOWLEDGEMENTS

This publication is the result of implementation of the project: "UNIVERSITY SCIENTIFIC PARK: CAMPUS MTF STU - CAMBO" (ITMS: 26220220179) supported by the Research & Development Operational Program funded by the EFRR.

This publication is the result of implementation of the project VEGA 1/0673/15: “Knowledge discovery for hierarchical control of technological and production processes” supported by the VEGA.

REFERENCES

[1] CHEN, Ming-Syan; HAN, Jiawei; YU, Philip S.. . Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and data Engineering, 1996, 8.6: 866-883.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] FRIEDMAN, Jerome H. Data mining and statistics: What's the connection?. Computing Science and Statistics, 1998, 29.1: 3-9. K. Elissa, “Title of paper if known,” unpublished.

[3] GROSSMAN, Robert L., et al. (ed.). Data mining for scientific and engineering applications. Springer Science & Business Media, 2013.

[4] Y ANDERBERG, Michael R. Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks. Academic press, 2014.

[5] KAUFMAN, Leonard; ROUSSEEUW, Peter J. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, 2009.

[6] 59. STATSOFT: STATISTICA Data miner. Dostupné na internete: www.statsoft.com/products/statistica-data-miner