Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 50-1 (2017) 6178–6183

Using data mining methods for manufacturing process control

P. Vazan*, D. Janikova, P. Tanuska*, M. Kebisek*, Z. Cervenanska***

**Institute of Applied Informatics, Automation and Mechatronics Slovak University of Technology in Bratislava, Faculty of Materials Science and Technology in Trnava, J. Bottu 25, Trnava 917 24, Slovak Republic (e-mail: pavel.vazan@stuba.sk; pavol.tanuska@stuba.sk; michal.kebisek@stuba.sk, zuzana.cervenanska@stuba.sk).*

***Advanced Technologies Research Institute, Slovak University of Technology in Bratislava, Faculty of Materials Science and Technology in Trnava, J. Bottu 25, Trnava 917 24, Slovak Republic (e-mail: dominika.janikova@stuba.sk).*

Abstract: The Industry 4.0 concept assumes that modern manufacturing systems generate huge amounts of data that must be collected, stored, managed and analysed. The case study is focused on predicting the manufacturing process behaviour according to production data. The paper presents the way of gaining knowledge about the future behaviour of manufacturing system by data mining predictive tasks. The proposed simulation model of the real manufacturing process was designed to obtain the data necessary for the control process. The predictions of the manufacturing process behaviour were implemented varying the input parameters using selected methods and techniques of data mining. The predicted process behaviour was verified using the simulation model.

The authors analysed different methods. The neural network method was selected for deploying new data by PMML files in the final phases. The objectives of the research are to design and verify the data mining tools in order to support the manufacturing system control by aiming at improving the decision-making process. Based on the prediction of the goal production outcomes, the actual control strategies can be precisely modified. Then they can be used in real manufacturing system without risks.

Keywords: decision making, manufacturing process, simulation, data mining, prediction.

1. INTRODUCTION

The life nowadays is characterised by an explosion in data volumes collected and stored in data stores. Enterprises accumulate their data in databases. However, information and knowledge are necessary to be extracted from these databases. Analysts must obtain information to be able to predict trends. Their analyses allow competent managers to make dynamic decisions, which is the process known as knowledge management performed by knowledge workers Repa (2007). The relationship between knowledge management and data mining is described in Tsai (2013). The process of extracting knowledge from databases, often called data mining, is an important step in the knowledge management. Knowledge management and data mining techniques allow to augment the decision-making of domain experts with additional knowledge and provide them with a competitive advantage Tsai (2013). Executives and managers are the final users of these systems at all levels of control.

Data mining is an interdisciplinary field with the general goal of predicting outcomes and uncovering relationships in data, enabling use of automated tools and techniques, employing sophisticated algorithms, to discover hidden patterns, associations, anomalies and/or structures from large amounts of data stored in a data warehouse or other information repositories. Prediction and description are, in the context of manufacturing, two primary goals of data mining.

While descriptive data mining focuses on discovering interesting patterns to describe the data, predictive data mining is aimed at predicting the behaviour of a model and determining future values of key variables based on existing information from available databases. The boundaries between descriptive and predictive data mining are not sharp Choudhary et al. (2006).

2. STATE OF THE ART

According to the surveys conducted by Rexer Analytics (2010) and KDnuggets (2014), data mining is not very often utilised in manufacturing. Less than 10 % of users solve the issues in manufacturing applying data mining.

The theoretical background of data mining applications in engineering design, manufacturing and logistics were laid in Feng et al. (2006). Recently, several reviews concerning data mining in manufacturing industry have appeared. Many possible applications of data mining in manufacturing, such as quality control, scheduling, fault diagnosis, defect analysis, supply chain, decision support system, are included in Bubenik et al. (2014), Choudhary et al. (2009), Trnka (2012). There are some certain specific examples of data mining applications, e.g. defect analysis in ceramics manufacture described in Dengiz et al. (2006), possibility of using the production data in determining the sequence of assemblies and minimizing the risk of producing faulty products Da Cunha et al. (2006). Some other examples

10.1016/j.ifacol.2017.08.986

of data mining applications in industrial, medical and pharmaceutical domains are presented in Kusiak (2006). Braha et al. (2007) brought a new decision-theoretic classification framework and applied it to a real-world semiconductor wafer manufacturing line. Using similar approach, Huyet (2006) considered the optimization and analysis of simulated production systems.

This paper is focused on the use of data mining methods for operational control level. The aim of the case study is to predict the behaviour of the manufacturing system on the basis of changes in the input parameters. Input parameters represent the control variables defining the specific control strategy. According to Li et al. (2009), the relations between these input parameters influence the performance of a manufacturing system. The behaviour of the production system will be evaluated applying quantitative characteristics of the selected control strategies of the manufacturing system. The performance indicators for manufacturing processes are defined in Grabot (1998).

3. METHODOLOGY

The procedure for system implementation is presented in Fig. 1. Each step is described below.

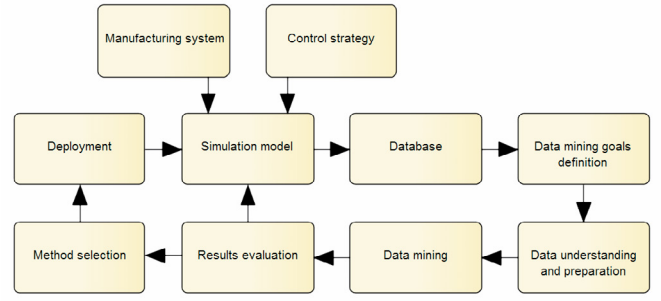

Fig. 1. Procedure for system implementation

3.1 Manufacturing system description

The case study deals with a real manufacturing system of an automotive industry component supplier located in the West of Slovakia. The company required a proposal of the way of prediction for selected system outputs because it was necessary to know the reaction to the rapid and frequent changes of input parameters during the production process.

The given manufacturing system produces two different types of products at the same time. The orders represent inputs of the system, which are released in different lot sizes with different input time intervals. 2D representation of simulation model is shown in Fig. 2.

This system represents a job shop system with batch production, which was implemented to the simulation model. The manufacturing system consists of 5 workplaces for technology operations and two output checkpoints for two final products. Workstations 1-4 include two CNC machines. Input and output buffers are assigned for each workplace. These buffers serve as interoperable storages. A transport of the production batches is carried out using vehicles with exactly defined tracks. These tracks join the workplaces. The material flow of both products respects the technological process and given operation scheduling. Material flow orientation is presented in Fig. 2.

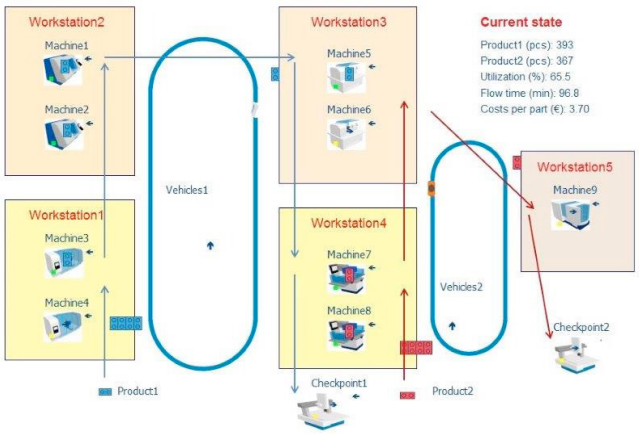

Fig. 2. Material flow of the manufacturing system

3.2 Simulation model

In the project, the simulation was used for generating the production data because the suitable data were missing. Simulation model was also applied for verifying the gained results due to increased risk for real verification in manufacturing system. The simulation model was built based on discrete-event simulation principles. The model of manufacturing system was created in Witness Simulator Lanner Group (2016). Information on the current state of manufacturing process was given by output variables, namely: number of operations, states of machines, state of buffers and state of transport devices. KPIs and KRI represent another group of output variables. These individual output parameters were calculated in entities of the discrete-event simulation model. The discrete-event simulation model allows calculation of these KPIs and KRI if the defined event occurs; for example, when the technological operation is finished. In this way, the KPIs and KRI calculation is very accurate, which was confirmed in the validation process.

Simulation model was validated according to real system. The validation process was an iterative process of comparing the model to actual system behaviour and results of real system. Finally, an accurate representation of the real system was achieved applying the model calibration. The simulation model was considered a black box and its structure was not changed for the next stage of the study. Data was generated by Witness Experimenter module. The possible ranges of input parameters for scenarios were: lot size for both products from 1 to 10 pcs and input intervals for both products from 5 to 35 min. 1981 scenarios were selected from combinations of inputs. Responses were defined as functions in simulation model and they represented the selected KPIs. Length of simulation run was defined as one month without replications for each scenario.

3.3 Control strategy

For the simulation model, the control strategy is proposed to achieve the specific goals of the production. KPIs represent a set of measures focusing on the aspects of organizational performance the most critical for the current and future success of the organization Parmenter (2009). KPIs tend to vary by organizations.

The presented case study is focused on the following KPIs: number of finished products, work in progress products, flow time, capacity utilization and KRI: costs per part unit.

These indicators are measured in short period. They have significant impact on the observed process and they have positive impact on performance of the observed process Parmenter (2009). The fulfilment of all these key indicators is monitored at the same time.

An appropriate adjustment of the input production parameters is necessary to reach the efficient production and the desired objectives of production.

The model able to describe the complex relationships between different settings of input parameters and objectives of control strategy was constructed. The quality and the objectives of the manufacturing process were assessed, based on the quantitative characteristics of individual production objectives. Five objective characteristics and their quantitative evaluation were selected. These were valid for a specific manufacturing system and a particular time period.

When the process fulfils the required conditions of strategy, the production goals can be considered fulfilled. These conditions were defined as follows:

"Costs per part unit < 5€ and Number of finished products > 600 pcs and Flow time < 110 min and Capacity Utilization > 60% and Work in progress production <= 4 pcs".

Based on the above conditions, a categorical variable 'Total goals' was created. This categorical variable takes the value of TRUE or FALSE, as shown in the last column (Table 1).

Table 1. Prediction data

	1 Costs per part	2 VD1	3 VD2	4 V1_ time	5 V2_ time	6 Num. of products	7 Flow time	8 Utiliza-tion	9 W.I.P.	10 Total goals
1	4.985	3	3	15	8	760	133.79	68.929	5.319	FALSE
2	4.141	3	3	15	9	760	96.731	65.765	3.149	TRUE
3	3.502	3	3	15	10	759	61.517	62.865	1.511	TRUE
4	3.236	3	3	15	11	738	44.702	60.083	0.801	TRUE
5	4.157	3	3	16	9	740	96.682	64.028	3.130	TRUE
6	3.498	3	3	16	10	739	61.497	61.087	1.493	TRUE
7	4.173	3	3	17	9	722	96.609	62.501	3.110	TRUE
8	4.166	3	3	18	9	709	96.573	61.077	3.092	TRUE
9	4.726	3	4	15	11	761	125.83	68.412	4.759	FALSE
10	4.130	3	4	15	12	759	99.134	65.83	3.253	TRUE

If the variable 'Total goals' is TRUE – then the process fulfils all the goals. On the other hand, if the value is FALSE – then the process does not meet all the goals simultaneously.

3.4 Database

The data gained from the individual simulation model runs represented values of the selected goal parameters and input variables of control parameters during the monitored period of manufacturing system. The time period was set up to one month and it was the same for each model run. The Oracle 12c relational database management system was chosen for recording generated process data.

3.5 CRISP-DM

Further, the CRISP-DM methodology was used (Fig. 1). The Rapid Deployment module allows to be applied for the pre-used models (PMML files – Predictive Model Markup Language) on the new data set. This module enables rewriting the variables to the predicted values presenting the input data set, as well as to external databases or a data warehouse. The PMML document is the output of these methods and provides a standard way to represent data mining models. These models can be shared between different statistical applications Statsoft Inc. (2013).

New set of input parameters was implemented for the simulation model proposed in Witness simulator to verify the predicted values. This data was used as input data in the process of knowledge discovery from databases. The aim was to compare the simulated and predicted values, and at the same time, to determine whether the selected models and specific algorithms in PMML files can be used for decision making in the process control. The comparison is mentioned in discussion of results.

3.6 Used data mining methods

Cognition of techniques and methods is needed for correct applying into the solved prediction problem. Based on the surveys from KDnuggets (2014) and Rexer Analytics (2010), and of course, regarding the data mining tasks and the actual available data, the authors decided to implement prediction using the following most commonly used predictive methods and techniques:

- Standard Classification and Regression Trees (CART)
- Boosting Tree (Boost)
- Random Forest (RF)
- Multivariate Adaptive Regression Splines (MARS)
- Support Vector Machine (SVM)
- K-Nearest Neighbour (KNN)
- Multiple Regression (MR)
- Neural Networks (used abbrev. NN).

3.7 Used metrics for data mining model evaluation

Accuracy of the categorical prediction models can be determined in various ways, from simple to complex. Several

metrics for classification and other metrics aimed at regression, such as the percentage of correct classification, overall accuracy, lift chart, error rate, and goodness of fit were used.

Overall accuracy – means the proportion of the total number of correct predictions.

Positive Predictive Value or Precision – is the proportion of correctly identified positive cases.

Negative Predictive Value – is the proportion of correctly identified negative cases.

Lift chart – is a measure of the effectiveness of a classification model calculated as the ratio of the results obtained with and without the model. Lift charts are visual aid for evaluating performance of classification models.

Error rate – for regression-type problems, when observed values for the outcome variable exist, the STATISTICA programme is used for computing the average squared error (residual) for each prediction model. For classification type problems, overall error rates are calculated.

Goodness of fit – Chi-square statistic, G-square statistic, and Percentage disagreement are normally used for classification models. For the numerical prediction, the following metrics are the most frequently used: mean square error, mean absolute error, relative mean squared error and correlation coefficient Statsoft Inc. (2013).

4. RESULTS

4.1 Classification – Predicting the behaviour of manufacturing system

Design of the data mining model using the selected methods for classification is shown in Fig. 3. The designed model includes categorical dependent variable – Total goals. The predictors are controllable continuous variables: lot size VD1, VD2 and their input time intervals V1_time, V2_time.

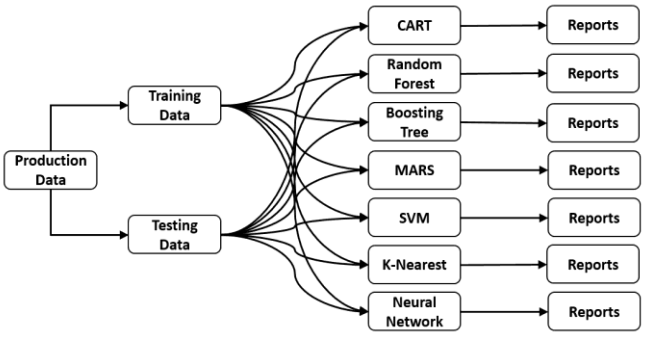

Fig. 3. Design of classification data mining model

The basic data file was divided by Split Input data node into training and testing data sets using random sampling and setting the approximate 30 percent of cases for testing. Therefore, the training set with size of 1395 observations and test set with 586 observations were created. The created data mining model was applied on the transformed data set and consequently, the accuracy of classification was evaluated according to the results of each model using test data.

Based on the results of all the methods, a summary Table 2 was created depicting the observed order of the model success in the given metrics.

Table 2. Summary order of classification methods

	NN	Rand	CART	KNN	SVM	Boost	MARS
Overall accuracy	1	2	3	4	5	6	7
Correctly classified TRUE	1	3	2	4	5	6	7
Correctly classified FALSE	1	3	2	4	5	6	7
Error rate	1	2	3	4	5	6	7
Lift chart	1	3	2	4	5	6	7
Chi-square	1	3	2	4	5	6	7
G-square	1	3	2	4	5	6	7
Percentage disagreem.	1	3	2	4	5	6	7
Summary	8	22	18	32	40	48	56

The mathematical-statistical scoring method was used to evaluate the obtained results for multi-objective evaluation. Order and points from 1 to 7 were assigned and evaluated from the most successful methods to the least successful method in the metrics. Seven models and eight metrics reached the best total score of eight and the worst total score of 56. The method with the lowest assigned number of points was the best and on the contrary, the method with the most points was the worst.

Based on the comparison and evaluation of all the methods, the NN method was chosen to classify the new data set. The number of neurons in input, hidden and output layers of the NN was automatically set to the specific task, defined by number of input parameters, output values and the input data set. Neural network has achieved the best results in all metrics.

4.2 Numerical prediction – Regression

Behaviour of the system is monitored by numerical prediction of production goal indicator. Created models can predict numerical values based on the relation between different settings of individual input control parameters and goals of the production. The numerical prediction is carried out particularly for each production goal. Each data mining model includes dependent continuous variable representing the indicator of production goal and predictors of lot size VD1 and VD2 with their input time intervals. Five data mining models were created for each individual production goals – costs per part unit, number of finished products, flow time, capacity utilization, and work in progress products (semi-finished products). Each created data mining model was applied on the transformed data set. Then the prediction accuracy was evaluated according to several metrics from the results of testing data set. Rapid Deployment Module was used for evaluating the models. First, the error rate (mean squared error residual) was found. The individual methods in ascending order are presented in Table 3.

Table 3. Error rate for costs per part

	Error rate
NN	0.001600
KNN	0.018314
CART	0.021583
RF	0.026241
SVM	0.052129
Boost	0.144016
MARS	0.166651
MR	0.166850

This table contains error prediction of the selected models in the testing data set. Smaller value of error indicates better and more powerful model. In this case, the NN model reached the smallest residual error. Furthermore, according to metrics of fit for the numerical prediction, the accuracy of each model can be compared. In Table 4, the methods are listed in ascending order according to the values obtained in individual metrics. The smallest error in the prediction was reached when NN method was used; this has been the most accurate solution.

Table 4. Goodness of fit for variable costs per part

	Mean square error	Mean absolute error	Mean relative square error	Correlation coefficient
NN	0.0030243	0.0412196	0.0001731	0.9955259
CART	0.0196653	0.0816930	0.0012026	0.9706029
KNN	0.0364194	0.1252114	0.0021226	0.9477075
SVN	0.0506216	0.1813791	0.0032855	0.9253509
RF	0.0685247	0.2017735	0.0042767	0.9231864
Boost	0.1632220	0.3192827	0.0112108	0.7250672
MARS	0.1639706	0.3263381	0.0115438	0.7196053
MR	0.1740447	0.3352871	0.0122151	0.6990720

The data mining models for each production goal were created similarly. In comparison with other methods, NN method reaches the bests results for each production goal. The mean squared error is used for evaluation of predictive power Kumari (2012). The sufficient predicting power of the model is given by reached results presented in Table 4.

5. RESULTS EVALUATION

All individual phases were performed according to procedure of implementing. In the final phases, the NN method was selected for deploying new data by PMML files.

In the first step of deployment, the classification was carried out and regression (numerical prediction) was performed in the next step. The system behaviour was determined using the classification method. Table 5 presents the classification result of input parameters with selected settings. The first three settings were classified as TRUE using the NN method, which means, simultaneous fulfilment of all the characteristics was predicted. The process with the first three settings can be considered successful.

Table 5. NN – Categorical prediction to new data

Set of parameters	VD1	VD2	V1_time	V2_time	Total goals
I	3	7	10	21	TRUE
II	5	4	15	19	TRUE
III	6	3	16	18	TRUE
IV	7	5	20	11	FALSE
V	8	6	18	20	FALSE

Setting the parameters IV and V did not fulfil the characteristics for all production targets at the same time. Verifying the classification using simulation is shown in Table 7. Based on the values of individual characteristics obtained by simulation, the required conditions of the management strategy can be considered fulfilled.

The values of the specific indicators (Table 6) were detected. The results were rounded to two decimal places. These predicted values confirmed the previous classification. The setting of parameters in cases IV and V did not fulfil the goals simultaneously. These settings did not meet the requirement of work in progress products; especially the number of products in buffers which should be maximum 4 pieces and the flow time required to be less than 110 minutes.

Table 6. Predicted values with new setting of parameters

Set of parameters	I	II	III	IV	V
VD1 [pcs]	3	5	6	7	9
VD2 [pcs]	7	4	3	5	6
VD1 time [min]	10	15	16	20	18
VD2 time [min]	21	19	18	11	20
Costs per part [€]	3.93	3.27	3.39	5.67	4.30
Num. of products [pcs]	912.54	822.01	844.50	977.08	1048.97
Flow time [min]	107.78	54.24	60.78	212.12	145.24
Utilisation [%]	80.54	72.25	73.17	92.83	93.16
W.I.P. [pcs]	4.00	1.10	1.68	9.52	6.15

The simulation confirmed the classification of each setting as shown in the last column 'Total goals' in Table 7. The values in this row correspond to the same column in Table 5. Thus, the classification using the NN method can be considered the most successful. Afterwards, the numerical values of simulation and prediction were compared in Tables 6 and 7.

Table 7. Values gained using simulation

Set of parameters	I	II	III	IV	V
Costs per part [€]	4.01	3.27	3.35	5.77	4.19
Num. of products [pcs]	907	841	835	973	1043
Flow time [min]	110.38	55.68	57.07	212.10	134.25
Utilisation [%]	80.05	71.71	72.47	90.96	92.63
W.I.P. [pcs]	3.99	1.40	1.52	11.15	6.03
Total goals [0/1]	TRUE	TRUE	TRUE	FALSE	FALSE

The highest differences occurred in predicting the number of finished products. Number of products predicted using the NN model differed from the simulation values by between 4 and 19 pieces (average deviation 0.96%). The simulation confirmed that the chosen methods were well-defined for all the production goals. Predicted costs per part unit were very accurate and the average deviation was 1.6%. Prediction of flow time was also very accurate; a maximum deviation of ca. 11 minutes was reached for setting V (average deviation 3.29%). The value for the capacity utilization was predicted very accurately as well, where the maximum deviation was 1.87% for parameter IV. In the numerical prediction of work in progress, there was the highest deviation of 1.63% in the setting IV, a difference corresponds to 2 pieces. All the above deviations can be considered acceptable for manufacturing system in our case study.

6. CONCLUSION

The achieved results demonstrate that data mining methods are suitable powerful tools for supporting the decision-making of managers. Based on the past data from the controlled process, the future states and goals of the control system can be clearly predicted using specific selected input parameters. Managers should be able to deeply understand the system behaviour to properly control the system. They need to comprehend the interactions between the decision parameters of the system, as well as their impact on its performance.

Based on the results, the drawn conclusion was that the selected PMML files of the neural network method (NN) for classification and numerical prediction are suitable for implementing them into business intelligent solution. By forecasting the behaviour of manufacturing systems and processes according to the defined KPIs and KRI, the hypothesis was confirmed that the selected input parameters might lead either to achieving or failing the declared objectives of the process. The specific value of goals can be predicted with tolerable accuracy. For each input, the required outputs can be predicted. All predicted values were verified on the simulation model of the manufacturing system.

7. FUTURE RESEARCH

Future research will be focused on the study of more objective parameters and application of the gained results into real practice. It will be oriented on proposing the methodology of data mining method selection for identified problems of manufacturing systems, and trying to determine the suitability of the analysed methods for specific sets of problems. Conceptual proposal for knowledge discovery in hierarchical control systems will be formulated as a holistic approach for solving the problems related to process large amount of data for the purpose of complex system control.

ACKNOWLEDGEMENT

This contribution was written with a financial support VEGA agency in the frame of the project 1/0673/15 “Knowledge discovery for hierarchical control of technological and production processes“.

REFERENCES

Bubenik, P. and Horak, F. (2014). Knowledge-based systems to support production planning. *Tehnicki Vjesnik-Technical Gazette*, 21(3), 505-509.

Braha, D., Elovici, Y., and M. Last. (2007). Theory of actionable data mining with application to semiconductor manufacturing control. *International 19 Journal of Production Research*, 45(13), 3059-3084.

Choudhary, A. K., Harding, J. A., and Tiwari, M. K. (2009). Data mining in manufacturing: a review based on the kind of knowledge. *Journal of Intelligent Manufacturing*, 20(5), 501-521.

Da Cunha, C., Agard, B., and Kusiak, A. (2006). Data mining for improvement of product quality. *International Journal of Production Research*, 44(18-19), 4027-4041.

Dengiz, O., Smith, A. E., and Nettleship, I. (2006). Two-stage data mining for awidentication in ceramics manufacture. *International Journal of Production Research*, 44(14), 2839-2851.

Feng, J. C. X. and Kusiak, A. (2006). Data mining applications in engineering design, manufacturing and logistics. *International Journal of Production Research*, 44(14), 2689-2694.

Grabot, B. (1998). Objective satisfaction assessment using neural nets for balancing objectives. *International Journal of Production Research*, 36(9), 2377-2395.

Huyet, A. L. (2006). Optimization and analysis aid via data-mining for simulated production systems. *European Journal of Operational Research*, 173(3), 827-838.

KDnuggets (2014). *Where analytics, data mining, data science is applied.* URL http://www.kdnuggets.com/2014/12/where-analytics-data-mining-data-science-applied.html.

Kumari, K. A., Boiroju, N. K., Ganesh, T., and Reddy, P. R. (2012). Forecasting surface air temperature using neural networks. *International Journal of Mathematics and Computer Applications Research*, 3, 65–78.

Kusiak, A. (2006). Data mining: manufacturing and service applications. *International Journal of Production Research*, 44(18-19), 4175-4191.

Lanner Group (2016). *Lanner group - provider of simulation tool witness*. URL http://www.lanner.com/technology/witness-simulation-software.html.

Li, J. and Meerkov, S. M. (2009). Production systems engineering introduction. *Production Systems Engineering*, 3-12.

Parmenter, D. (2009). *Key Performance Indicators (KPI): Developing, Implementing, and Using Winning KPIs*. John Wiley and Sons, Inc., United States of America.

Rexer Analytics (2010). *4th annual data miner survey 2010 report*. URL http://www.rexeranalytics.com/Data-Miner-Survey-2010-Intro2.html.

Repa, V. (2007). *Business processes: Process Control and Modeling*. Grada, Praha.

StatSoft, Inc. (2013). *Electronic statistics textbook*. URL http://www.statsoft.com/textbook/.

Trnka, A. (2012). Results of application data mining algorithms to (Lean) Six Sigma methodology. *International Journal of Engineering*. 10(1), 141-144.

Tsai, H. H. (2013). Knowledge management vs. data mining: Research trend, forecast and citation approach. *Expert Systems with Applications*, 40(8), 3160-3173.