Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 72 (2018) 1051–1056

www.elsevier.com/locate/procedia

51st CIRP Conference on Manufacturing Systems

Lead time prediction using machine learning algorithms: A case study by a semiconductor manufacturer

Lukas Lingitz[a], Viola Gallina[a,*], Fazel Ansari[a,b], Dávid Gyulai[c,d], András Pfeiffer[c], Wilfried Sihn[a,b], László Monostori[c,d]

[a]*Fraunhofer Austria Research GmbH, Theresianumgasse 27, A-1040 Wien, Austria*

[b]*Vienna University of Technology, Institute of Management Science, Theresianumgasse 27, A-1040 Wien, Austria*

[c]*Centre of Excellence in Production Informatics and Control (EPIC), Institute for Computer Science and Control (SZTAKI), Hungarian Academy of Sciences (MTA), Budapest, Hungary, Kende str. 13-17, H-1111 Budapest, Hungary*

[d]*Department of Manufacturing Science and Engineering, Budapest University of Technology and Economics, Müegyetem rkp. 3, H-1111 Budapest, Hungary*

[*] Corresponding author. Tel.: +43-676-888-616-46; *E-mail address:* viola.gallina@fraunhofer.at

Abstract

The accurate prediction of manufacturing lead times (LT) significantly influences the quality and efficiency of production planning and scheduling (PPS). Traditional planning and control methods mostly calculate average lead times, derived from historical data. This often results in the deficiency of PPS, as production planners cannot consider the variability of LT, affected by multiple criteria in today's complex manufacturing environment. In case of semiconductor manufacturing, sophisticated LT prediction methods are needed, due to complex operations, mass production, multiple routings and demands to high process resource efficiency. To overcome these challenges, supervised machine learning (ML) approaches can be employed for LT prediction, relying on historical production data obtained from manufacturing execution systems (MES). The paper examines the use of state-of-the-art regression algorithms and their effect on increasing accuracy of LT prediction. Through a real industrial case study, a multi-criteria comparison of the methods is provided, and conclusions are drawn about the selection of features and applicability of the methods in the semiconductor industry.

Peer-review under responsibility of the scientific committee of the 51st CIRP Conference on Manufacturing Systems.

Keywords: Lead time; prediction; machine learning; regression methods; comparison; features

1. Intoduction

Predictive data analytics refers to building and using models that make predictions based on the patterns extracted from historical data [1]. Considering the six key phases of a predictive data analytics project lifecycle —as defined by the Cross Industry Standard Process for Data Mining (CRISP-DM) [2], namely (i) business understanding, (ii) data understanding, (iii) data preparation, (iv) modeling, (v) evaluation and (vi) deployment—, the fourth phase (modeling) is where machine learning (ML) algorithms are employed to build prediction models [1,3]. The best model which fits for the purpose of prediction, for instance lead-time prediction, will be evaluated and proved for deployment e.g. in manufacturing execution systems (MES). In particular, ML is defined as an automated process that extracts patterns from (historical) data [1]. To this end, one can distinguish between two main approaches: (1) Supervised ML, which assumes that training examples are classified (labeled) (i.e. learning relationship between a set of descriptive features and a target feature), and (2) Unsupervised ML, which concerns the analysis of unclassified examples [1,3]. Other types of ML include semi-supervised and reinforcement learning. In this paper, we focus exclusively on supervised ML in particular on regression algorithms.

The task of predicting a continuous target or value is referred to as a regression task/problem [1]. The most common algorithms for a regression task are linear and logistic regression, i.e. modeling relationship between the continuous variable (e.g. lead-time) and one or more predictors (e.g. operation time, number of orders in progress, work in progress (WIP), etc.) using a linear or logit link function, respectively [1].Addressing this line of research in the semi-conductor manufacturing, the key research question is "which logistical measures are needed in order to achieve an accurate lead time prediction for single, connected and repetitive production steps using ML algorithms"? The challenge is to identify ML algorithms that are suitable for discovering interrelations between the aforementioned measures. To tackle this problem, we employ supervised ML methods applied to solve regression tasks.

The paper is structured as follows: Section 2 presents the state-of-the-art methods in lead-time prediction using regression algorithms. Section 3 provides a methodology for handling

2212-8271
Peer-review under responsibility of the scientific committee of the 51st CIRP Conference on Manufacturing Systems.
10.1016/j.procir.2018.03.148

regression tasks which is exemplified in the context of semiconductor manufacturing. Finally, we discuss the key findings and identify future research potentials.

2. Lead time prediction with ML algortihms

In this section, we firstly discuss knowledge discovery approaches in production planning and control (PPC) (cf. Section 2.1). Secondly, we explore state-of-the-art regression-based reproaches in lead time prediction (cf. Section 2.2.).

2.1. Knowledge discovery in production planning and control

Statistical learning, data mining or knowledge discovery was first defined in 1989 [4] as a new intelligent tool for extracting useful information and knowledge (actionable information or hidden patterns) from different databases. In the first time period, knowledge discovery was extensively applied in many different fields —such as in medicine- and biotechnology, finance, marketing— but compared to these fields, there has been less research interest in the manufacturing domain [5]. However, in the past few years there have been a significant growth in the number of papers discussing the usage of different data analytics methods and techniques in production management [4,6,7]. According to Rainer, after applying different data mining techniques for converting *big data* to *smart data*, companies have experienced payback of at least ten times of their investment [5].

The output information of data mining can be split into two main categories based on the functions and the goals of the applied technique. *Descriptive statistical analytics* (like association analysis or clustering) focuses on the discovery of rules or patterns to describe the data, while *predictive data mining* (such as classification or regression) is used to analyze the relevant and actual data in order to predict future values for one or more key variables [4]. Regression is one part of predictive data mining, where the objective is to predict the value of a continuous variable. Cheng et al. revised the relevant papers since 2010 and discussed the typical knowledge mining techniques in production management [4]. According to this survey, the four most reflected typical application areas are advanced planning and scheduling, quality improvement, fault diagnosis and defect analysis. A fifth category was defined, in which flow time/cycle time prediction could be found –among life and yield prediction. PPC was identified as a research gap already in 2009 [8], and the review [4] in 2017 has revealed just a few applications in this passed 9 years. Consequently, more attention from the research community would be needed to data mining in PPC.

2.2. Lead time prediction with regression: state-of-the-art methods

In the present paper, lead time –as one of the most important control parameter and target figures of PPC– is analyzed and predicted with the help of different ML algorithms and based on MES data. The literature survey revealed that most research of time related data mining analysis (flow time, (lot) cycle time, lead time): i) have focused on the whole process flow, ii) have used a dataset generated by simulation and iii) have applied and compared just a few ML algorithms. Pfeiffer et al. [9] compared the results of three ML models predicting the lead time with eight features and data gained by discrete-event simulation. In particular, they employed random forest model outperformed linear regression and regression tree models. Ozturk et al. [10] applied regression trees to a simulated data source of four shop types in order to determine the most relevant attributes having a relatively high predictive power. Meidan et al. [11] focused on the waiting time rather than the whole lead time. After discretization of all continuous variables selective naive Bayesian classifier, decision tree, artificial neural network (ANN) and multinomial logistic regression were evaluated. It is revealed that, the 182 features of the original dataset generated by simulation could be reduced to 20. Alenzi et al. [12] used a thoroughly tuned support vector regression model (SVM) for real-time flow time prediction and compared it with an ANN model and traditional time series models. On the applied dataset obtained from a simulation model the SVM performed best. The comparison of a Bayesian network to an ANN and to a SVM model was executed by Mori et al. [13] for production times in the steel industry. It was concluded that all methods could accurately predict the output variables in case of completely observed variables, however, with partially known input data the Bayesian network had the best performance for the simulated example with binarized variables. De Cos Juez et al. [14] analyzed the results of a SVM model with 8 features (reduced from 12) to predict whether a batch is going to be completed in the forecasted time or not. A stepwise regression model was implemented by Li et al. [15] in order to estimate the relationship between the characteristic of the flow time distribution and the predictor variables. In the work of Raaymakers et al. [16] the ANN performed slightly better than regression models for estimating the makespan of job sets in batch process industries.

3. Research methodology

The methodological approach applied in the use case orientates on the CRISP-DM model, with the focus on the first five phases. According to Figure 1, the first three phases, namely business understanding, data understanding and data preparation are consolidated in the section *Description of process and data*. The section *ML algorithm - Toolbox* of Figure 1 contains the modeling phase, while the evaluation is done in the *Results* section, where relevant features have been selected and the accuracy of prediction model itself has been evaluated.

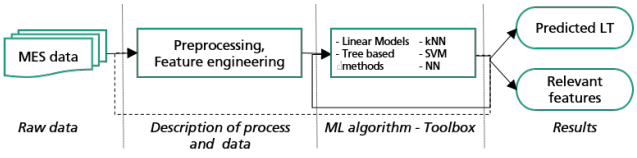

Fig. 1. Data analysis process steps applied in PPC

3.1. Description of manufacturing process and data

As stated above, the case study has been conducted in the semiconductor industry. A special characteristic of this industry is that most of the products are built in multiple layers.

Moreover, most production plants do not have rigidly linked machines, but the machines are organized in operation specific workcenters. This leads to a job shop organized production system, where products have to run on the same machines several times in order to build integrated circuits on the layers. Although the classification for the production type is mainly mass production, normally many different products and also smaller lot sizes for qualification of new products have to be handled on the same equipment. Due to the complexity described above, the semiconductor industry traditionally invested a lot in advanced IT-systems in the production area, resulting in a large amount, detailed data available about products, processes and equipment.

The company of under study is characterized by these attributes. Within their use case, we concentrated on a sequence of three process steps, namely *Sorter, Bakefuse* and *Sputter* illustrated on Figure 2. The *Sorter* initializes the process sequence and is carried out to sort all wafers of one lot and bring them into the correct sequence for the next process steps. This is needed, as throughout the production, it might occur that the sequence of the wafers within one lot can be mixed up, thus the sorter reorganizes the lot. After this process, there is the main buffer before the lot goes to the *Bakefuse* stage. This process step is necessary for the preparation of the wafers for the *Sputter* process where a thin layer of metal gets sputterd on the wafer's surface. There is a process related maximum lay time between the *Bakefuse* and the *Sputter* process that limits the buffer size and leads to the blocking of lots before. Therefore, the measurement and later on also the prediction of the lead time is done not only per lot but also per layer. This means, that a lot can arrive several times —depending on the product structure, up to three times, what means that the product has 3 layers— in the observed process sequence, within a time span of several days.

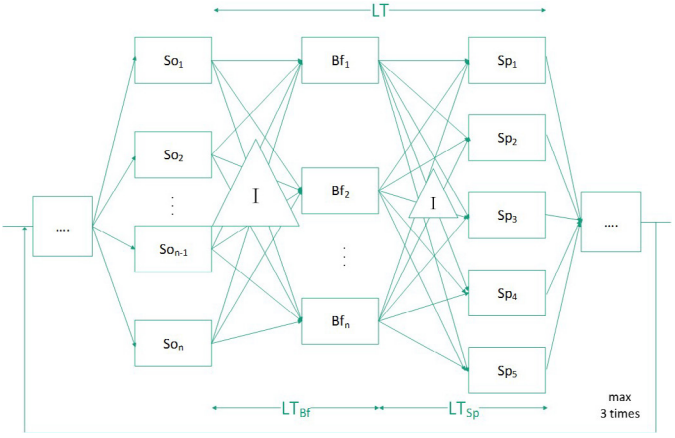

Fig. 2. Illustration of the analyzed manufacturing process steps

Process confirmation data from the MES system, historical information about the equipment/machine status and customer relating information were provided by the company for a period of two years. The characteristics of input data are summarized in Table 1. In the first two columns the name and the type of the data are described. The third column contains information about the range of the particular variable — how many unique values the variable has or how detailed information about the variable was provided. The company supplies 33 customers with 106 different products. The production of wavers is facilitated in these three analyzed process steps by 43 different equipments and 38 various routings. During these three analyzed process steps 14 different operations can be distinguished. Time stamp information about move-in and move-out of the various operations are available with the precision of seconds. Graphical illustration of the time stamp data is presented on Figure 3. In case of the analysis of machine utilization we keep our focus on the 5 machines in the process *Sputter* illustrated on Figure 2.

Table 1. Description of raw data

Name	Type	Granularity
Product number	Alphanumeric string	106
Customer	Alphanumeric string	33
Production lot	Alphanumeric string	23819
Operations	Alphanumeric string	14
Routings	Alphanumeric string	38
Time stamp	Date and time	Seconds
Production quantity	Integer	0-25
Equipment	Alphanumeric string	43
Priority	Integer	3
Status of operations	Alphanumeric string	22

These process confirmation data and information about the machine status and customers were used to generate a ML database, which included a number of 18,532 observations about lot-layer combinations. For the evaluation, the ML dataset was split up into training and testing datasets randomly, applying a 70/30 sampling ratio. Out of the raw data, we derived 41 features (35 numerical and 6 categorical features) that were analyzed according their impact on the lead time. The features can be divided according to their characteristics into static features that characterize the lot —e.g, product, customer, planned time— and dynamic ones. The latter features reflect the status of the production system, especially the three observed processes, at the entry time of the lot at a specific layer.

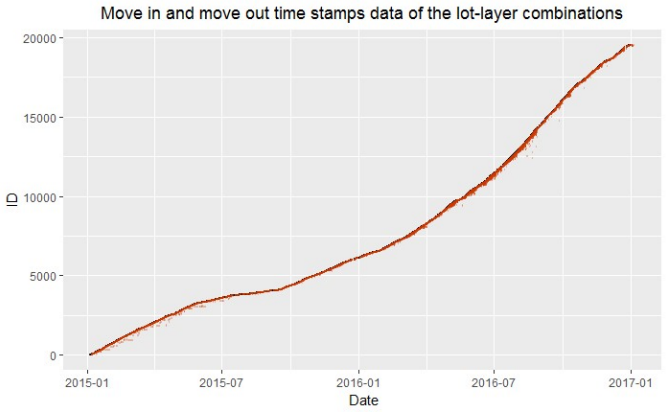

Fig. 3. Time stamp data of the lot-layer combinations

The lead time is calculated for the processes steps *Bakefuse* and *Sputter* separately. For those processes the lead time is defined as time span from the end confirmation of the previous process and the observed process, i.e. the lead time in *Bakefuse* is calculated from the end confirmation date and time in *Bakefuse* minus the end confirmation date and time of the previous process, the *Sorter* process. For the *Sputter* the pre-

vious process is *Bakefuse*. The overall leadtime for all three process steps is calculated from the end confirmation at *Sputter* and the end confirmation at *Sorter*. According to domain experts from the company the WIP before *Sorter* can be neglected due to technical reasons and was therefore not considered. The calculated overall lead time is illustrated on Figure 4 .

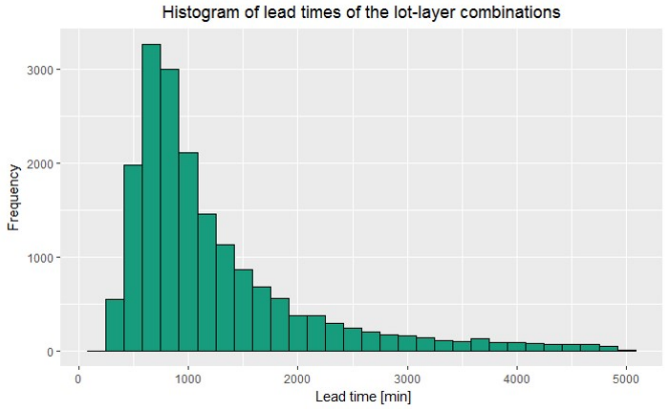

Fig. 4. Histogram of the lead time in the analyzed process steps

3.2. Exploring ML algorithms

Basically, statistical learning refers to a set of methods for understanding data, recovering relationships between parameters and for providing estimations about desired variables. In case of regression, the value of a continuous variable is to be predicted. The most fundamental regression methods are the easily interpretable linear regression models (LM), which assume an approximately linear relationship among the variables. These models seek to find estimates of the parameters so that the sum of mean squared error is minimized. While ordinary linear regression focuses on the model bias, ridge and lasso regression concentrate on the model variance [17]. There are various regression models that are inherently nonlinear in nature, such as ANN, multivariate adaptive regression (MARS), SVM, k-nearest neighbor (kNN) and tree based models. ANN is a powerful regression technique inspired by the working of the human brain. The output variable is modeled by an intermediary set of unobserved variables, by the so called hidden layer(s) [13,16,17]. MARS is a non-parametric regression method and can be used for modeling complex nonlinearities and relationships between variables [17]. The main idea of SVMs –which can be used for both classification and regression– is the individualization of hyperplanes parallel to error minimization [12,14,17]. kNN approach predicts outcomes using the k closest samples from the training set [17]. Regression trees (RT) partition data into more homogeneous with respect to the response variable groups. Although regression trees are easy to interpret and implement, they have less-than-optimal predictive performance and they are instable. Ensemble techniques could be used in order to reduce the variance of the prediction and to increase the accuracy and stability of the model. But the improved performance has its trade-offs, such as computational costs, memory requirements and loss in interpretation. Bagging and boosting are among the most widespread ensemble methods. Bagging (bootstrap aggregation) is a general approach that uses bootstrapping. The prediction of a bagged regression tree model is the average of the predictions of each regression tree in the bagged ensemble. Random forest (RF) is a special case of a bagged regression tree, where just randomly selected predictors are used in the tree construction process, in order to reduce the correlation among predictors. While in RF all trees are independent, they contribute to the final model equally and each tree is created to have maximum depth, in boosting however the trees are dependent on past trees, have different contributions to the final model and they have a minimum depth [17].

3.3. Application and evaluation of the selected ML techniques

Eleven various statistical learning methods were evaluated in the analysis, by using *R* and *R Studio* [18,19]. In case of linear models (LM, Ridge and Lasso) and ANN regression, only numerical features were applied. During the model building and feature selection, 10-fold cross validation was performed to estimate the prediction accuracy of the models on the independent test set. The accuracy of the models were measured with 5 various error measures, namely with mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared error (MSE),root mean squared error (RMSE) and normalized root mean squared error (NRMSE). The *NRMSE* gives the average prediction error in the percentage of the real lead time values. The calculation of NRMSE is provided by Equation 1, where P_i and R_i are the predicted and real lead time values, respectively. The results of the regression models are summarized in Table 2.

$$\mathrm{NRMSE} = 100\frac{\sqrt[2]{\frac{1}{n}\sum_{i=1}^{N}(P_i - R_i)}}{R_{max} - R_{min}} \tag{1}$$

It can be concluded that nearly the same accuracy was achieved by all the three linear models (see the different error measures in Table 2). The results of the RF model were calculated with the default parameters: 500 trees, 13 random variables from the 41 —as the suggested value is the third of the number of all predictors in case of regression. The boosted RT model had 20000 trees. In the SVM model radial kernel was applied. The optimal value of k of the kNN model proved to be 9, after resampling k from 5 to 23. The ANN model had one hidden layer with 26 neurons —as the recommended value is two third of the number of input variables (a minor increase in

Table 2. Accuracy of the tested ML algorithms based on different error measures

	LM	Ridge	Lasso	RT	bagged RT	RF	boosted RT	SVM	MARS	kNN	ANN
MAE	487	510	508	563	394	390	397	423	488	504	535
MAPE	42.7	45.0	44.7	53.5	33.9	33.8	33.9	30.9	43.15	44.0	53.4
MSE	529408	573520	572939	639617	369993	360780	369414	500693	513638	554897	658852
RMSE	727	757	756	799	608	600	607	707	716	745	771
NRMSE	15.2	15.8	15.8	16.7	12.7	12.5	12.7	14.77	14.9	15.5	17.3

Table 3. Description and sensitivity analysis results of the ten most important variables of the models

Feature	Description	RF	boosted RT
MovDeparture	Moving average of the inter departure times of the last 20 lot-layers	2,7	3.3
ArrivalHour	Hour of the arrival time	-7,6	-13.9
WD	Weekday of the arrival time	3.6	2.7
SumMedOTs	Median of the product's lead time (in the analyzed process steps)	5.2	1.5
WIP	Work in progress: number of lot-layers in the analyzed process steps	5.3	2.6
WIPtimeBfMed	Work in progress: expected work content in minutes by process step bakefuse	2.9	1.7
SpEffPrevDay	Capacity utilization of machines in process step sputter on the previous day	0.1	1.9
MovArrival	Moving average of the inter arrival times of the last 10 lot-layers	-1.3	-1.3
medOTProdRout	Mean of median operations times of a product on a given route	3.6	2.1
SBPrevDay	Time in standby status of the machines of sputter on the previous day	1.2	0.8

the NRMSE was found with with 20 and 30 neurons). According to the results summarized in Table 2, ensemble tree based methods could outperform all models, and the best result was provided by RF. Bagging and RF models were built by using *randomForest* package [20,21] and the boosted RT model was constructed with the *gbm* package [22]. The ensemble nature of these models makes it impossible to gain an understanding of the relationship between the input and output variables. However, it is possible to quantify the impact of predictors in the ensemble [17]. The *importance* function of the *randomForest* package can evaluate the importance of all variables with two different approaches: i) random permutation of the values of each predictor or ii) improvement in node purity based on the performance metric for each predictor [17]. Variable importance for boosting is a function of the reduction in squared error due to each predictor. The intersection of the most important variables, found by *randomForest* and *gbm* packages, are summarized and described in the first two columns of Table 3. With the help of the ten most important variables a new RF and a boosted RT model were built —as suggestion for lead time prediction. During the fine tuning of the model, it was discovered, that the optimal number of trees in the RF model construction is between 100-125, as the NRMSE of 13.1 with 500 trees (running time is about 2 minutes) does not increases till the number of trees are decreased to 125 (running time is less then a minute). The running time is about 20-30 seconds with 100 trees and the increase in the NRMSE is only 0.1. The running time of the boosted RT model was significantly more (6 minutes with 25000 trees and 5 minutes with 20000 trees) and the accuracy results never outperformed the results of the RF model (NRMSE of 13.5 and 13.6 respectively).

After the construction of the final RF and boosted RT models, the variable importance of all 10 variables was studied with the help of sensitivity analysis. The sensitivity analysis results are summarized in the last two columns of Table 3, where the increase in MSE (in %) without the particular variable can be seen (in case of the RF model the results of 5 runnings are averaged). According to the results, the number of lot-layers in the analyzed process steps (WIP) appears to be the most important variable, as disregarding this feature from the final model had the most significant impact on the MSE. Removing the WIP as a feature from the final RF and boosted RT model resulted in a 5.3% and in a 2.6% increase in the MSE. The sensitivity analysis results for two variables (ArrivalHour and MovArrival) are negative, so in these particular cases better results were achieved without these features. As mentioned in Kuhn and Johnson [17], there could be bias in RF variable importance measure. Two examples are named that could have serious effect on the importance values; (i) the correlations between predictors and the (ii) number of random variables during the model construction. Further analysis is needed in order to understand the interrelations between these two and the other features.

Our final model suggested for lead time prediction in this particular case is a RF model with all the eight variables with positive sensitivity analysis results in Table 3. With 125 trees and 2 random variables during the model construction — running time is 20 seconds— the original (with all the 41 features) NRMSE of 12.5 is reached.

4. Future research agenda

Further research will be conducted by the authors. First, the scope of the analysis will be extended and the developed approach will be applied to other process steps and to the production system as a whole to analyse, if the identified approach with its eight variables is also suitable for different processes. Another future research agenda would be the application of the approach to other industries to check the suitability of the variables and learning methods.

From the literature studies and from the conducted research the need for a *feature codebook* came up. Therefore new features need to be defined and tested and existing features need to be tuned and their applicability and suitability for different production process types (e.g. batch processes, continuous processes etc.) and industries need to be studied and documented.

Last but not least, the analysis of interrelations between variables with negative sensitivity will be investigated.

5. Acknowledgements

This research work has been performed in the EU project Power Semiconductor and Electronics Manufacturing 4.0 (SemI40), which is funded by the programme Electronic Component Systems for European Leadership (ECSEL) Joint Undertaking (Grant Agreement No. 692466) and the programme IKT der Zukunft (project number: 853343) of the Austrian Ministry for Transport, Innovation and Technology (bmvit) between May 2016 and April 2019. More information on IKT der Zukunft can be found at https://iktderzukunft.at/en/. Moreover, the project SemI40 is co-funded by grants from Germany, Italy, France, and Portugal. We also would like to thank our partners within the SemI40 project for their support.

The authors would like to acknowledge the financial support of the European Commission for funding the H2020 research project EPIC (https://www.centre-epic.eu/) under grant No. 739592.

References

[1] Kelleher, J., Mac Namee, B., D'Arcy, A.. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press; 2015.

[2] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., T., R., Shearer, C., et al. Step-by-step data mining guide: Step-by-step data mining guide. 2000.

[3] Fürnkranz, J., Gamberger, D., Lavrač, N.. Foundations of Rule Learning. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. ISBN 978-3-540-75196-0. doi:10.1007/978-3-540-75197-7.

[4] Cheng, Y., Chen, K., H., S., Zhang, Y., Tao, F.. Data and knowledge mining with big data towards smart production. Journal of Industrial Information Integration 2017;doi:10.1016/j.jii.2017.08.001.

[5] Rainer, C.. Data mining as technique to generate planning rules for manufacturing control in a complex production system. In: Windt, K., editor. Robust Manufacturing Control. Lecture Notes in Production Engineering; Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-30748-5; 2013, p. 203–214. doi:10.1007/978-3-642-30749-2_15.

[6] Esmaeilian, B., Behdad, S., Wang, B.. The evolution and future of manufacturing: A review. Journal of Manufacturing Systems 2016;39:79–100. doi:10.1016/j.jmsy.2016.03.001.

[7] Ansari, F., Dienst, S., Uhr, P., Fathi, M.. Using data analysis for discovering improvement potentials in production process. Joint IEEE International Conference on Industrial Electronics (IEEE ICIT) 2011;:279–284.

[8] Choudhary, A.K., Harding, J.A., Tiwari, M.K.. Data mining in manufacturing: A review based on the kind of knowledge. Journal of Intelligent Manufacturing 2009;20(5):501–521. doi:10.1007/s10845-008-0145-x.

[9] Pfeiffer, A., Gyulai, D., Kádár, B., Monostori, L.. Manufacturing lead time estimation with the combination of simulation and statistical learning methods. Procedia CIRP 2016;41:75–80. doi:10.1016/j.procir.2015.12.018.

[10] Öztürk, A., Kayalıgil, S., Özdemirel, N.E.. Manufacturing lead time estimation using data mining. European Journal of Operational Research 2006;173(2):683–700. doi:10.1016/j.ejor.2005.03.015.

[11] Meidan, Y., Lerner, B., Rabinowitz, G., Hassoun, M.. Cycle-time key factor identification and prediction in semiconductor manufacturing using machine learning and data mining. IEEE Transactions on Semiconductor Manufacturing 2011;24(2):237–248. doi:10.1109/TSM.2011.2118775.

[12] Alenezi, A., Moses, S.A., Trafalis, T.B.. Real-time prediction of order flowtimes using support vector regression. Computers & Operations Research 2008;35(11):3489–3503. doi:10.1016/j.cor.2007.01.026.

[13] Mori, J., Mahalec, V.. Planning and scheduling of steel plates production. part i: Estimation of production times via hybrid bayesian networks for large domain of discrete variables. Computers & Chemical Engineering 2015;79:113–134. doi:10.1016/j.compchemeng.2015.02.005.

[14] De Cos Juez, F. J., , García Nieto, P.J., Martínez Torres, J., Taboada Castro, J.. Analysis of lead times of metallic components in the aerospace industry through a supported vector machine model. Mathematical and Computer Modelling 2010;52(7-8):1177–1184. doi:10.1016/j.mcm.2010.03.017.

[15] Li, M., Yang, F., Wan, H., Fowler, J.W.. Simulation-based experimental design and statistical modeling for lead time quotation. Journal of Manufacturing Systems 2015;37:362–374. doi:10.1016/j.jmsy.2014.07.012.

[16] Raaymakers, W., Weijters, A.. Makespan estimation in batch process industries: A comparison between regression analysis and neural networks. European Journal of Operational Research 2003;145:14–30. doi:10.1016/S0377-2217(02)00173-X.

[17] Kuhn, M., Johnson, K.. Applied predictive Modelling; vol. Springer. 2013. ISBN 978-1-4614-6848-6.

[18] RCoreTeam, . R: A language and environment for statistical computing 2017;URL: `https://www.R-project.org`; visited on 01.03.2018.

[19] RStudioTeam, . Rstudio: Integrated development environment for r 2016;URL: `https://www.rstudio.com/`; visited on 01.03.2018.

[20] Liaw, A., Wiener, M.. Classification and regression by randomforest. The Newsletter of the R Project 2002;3(2):18–22.

[21] Breiman, L., Cutler, A., Liaw, A., Wiener, M.. Package 'randomforest': Breiman and cutler's random forests for classification and regression 2015;URL: `https://cran.r-project.org/web/packages/randomForest/randomForest.pdf. R package version 4.6.12`; visited on 01.03.2018.

[22] Ridgeway, G.. Package 'gbm': Generalized boosted regression models 2017;URL: `https://cran.r-project.org/web/packages/gbm/gbm.pdf. R package version 2.1.3`; visited on 01.03.2018.