

PRODUCTION LOG DATA ANALYSIS FOR REJECT RATE PREDICTION AND WORKLOAD ESTIMATION

András Pfeiffer
Dávid Gyulai
Ádám Szaller
László Monostori

MTA SZTAKI – Institute for Computer Science and Control
Hungarian Academy of Sciences
Kende str. 13-17.
Budapest 1111, HUNGARY

ABSTRACT

The main focus of the research presented in this paper is to propose new methods for filtering and cleaning large-scale production log data by applying statistical learning models. Successful application of the methods in consideration of a production optimization and a simulation-based prediction framework for decision support is presented through an industrial case study. Key parameters analysed in the computational experiments are fluctuating reject rates that make capacity estimations on a shift basis difficult to cope with. The most relevant features of simulation-based workload estimation are extracted from the products' final test log, which process has the greatest impact on the variance of workload parameters.

1 INTRODUCTION

At the operational level of manufacturing systems, difficulties arise from unexpected tasks and events, non-linearities, and a multitude of interactions while attempting to control various activities in dynamic shop floors. The selection of the most appropriate production control decision for a given assignment, as well as the prediction of cycle times, waiting times, workloads, or utilizations of the resources are no trivial tasks, although they can be supported by simulation-based evaluations (Pfeiffer 2007; Bagchi et al. 2008; Watt 1998; Honkomp et al. 1999; Rose 2007; Sivakumar and Chong 2001).

The discrete-event simulation (hereafter referred to as simulation) approach has been applied to decisions in scheduling and control, related to production applications (Banks 1998; Law and Kelton 2000; O'Rielly and Lilegdon 1999; Kim et al. 1998). The simulation models that are used for making or evaluating these decisions (e.g., by projecting different key performance indicators, KPIs) generally represent the flow of materials to and from processing machines and the operations of machines themselves (Rabelo et al. 2003). Potential problems can be identified and corrected using a simulation model. By far the most common use of simulation models is for operational decisions such as scheduling or dispatching (Law and Kelton 2015; Lin et al. 2001; Sabuncuoglu and Kizilisik 2003). In related works about application areas, as well as the recent solutions of simulation in production control, simulation has been typically used for off-line decision making. Consequently, effective integration into the control process of production was restrained. One of the limitations of its use in on-line decision making is the considerable amount of time spent in gathering and analysing data. In quasi real-time control (hours, minutes), however, the three key issues are data acquisition, quick response, and instantaneous feedback (Cowling and Johansson 2002). As a result, decision makers mostly apply simulation primarily for off-line decision support and not for the critical on-line decision making that may arise.

Therefore, based on previous results (Pfeiffer et al. 2016; Pfeiffer et al. 2011; Monostori et al. 2007), we propose a decision support architecture, in which the integrated, self-building simulation module can be applied for validation of the calculated manufacturing capacities, a priori recognition of due date deviations, and analysis of the effect of possible actions taken. In the research presented in the paper, special emphasis is given to the prediction and evaluation of the production on a daily, rolling time horizon (e.g., Work in process (WIP) trajectories, machine utilizations).

An important issue regarding the short-term (operational level) simulation is the automatic collection and definition of simulation input data. Therefore, as the main focus of this paper, a new reject rate classification and prediction method using logged testing data is presented.

The paper is structured as it follows. In Section 2, possible combinations of statistical learning and simulation in production control are introduced. Section 3 provides a method for production log data processing to predict reject rates through an industrial case study. Section 4 introduces simulation-based workload estimation, and the conclusions and outlook are provided in Section 5.

2 COMBINATION OF SIMULATION AND STATISTICAL LEARNING

The main goal of the framework introduced here is to provide a self-building production simulation, capable of both prospective (e.g., locate anticipated disturbances, identify trends of designated performance measures), and retrospective (e.g., gathering statistics on resources) simulation functionalities. Self-building simulation means that the simulation model is built up by means of the combination of the manufacturing execution system (MES) data as well as the knowledge extracted from the MES data (e.g., resource and execution model). In addition to the automatic model building feature, the main requirement of the solution is to minimize the response time of the experiments and to enable the quasi “real-time” applicability of the simulation (Monostori et al 2007).

The main operation modes of the simulator in the proposed architecture (Figure 1) are as follows:

- Off-line validation, sensitivity analysis and statistical modelling of the system. Evaluation of the robustness of the system against uncertainties (e.g., different control settings, thresholds and system load levels). Consequently, this scenario analysis might point out the resources or settings which can endanger the normal operation conditions. In Figure 1, off-line simulation represents the comprehensive model of the plant.
- On-line, anticipatory recognition of deviations from the planned operation conditions by running the simulation parallel to the plant activities; and by using a lookahead function supporting situation recognition (proactive operation mode, Figure 1).
- On-line analysis of the possible actions and minimization of the losses after a disturbance already occurred (reactive operation mode, Figure 1), e.g., what-if scenario analysis.

In Figure 1, *Plant* represents the underlying production system, which is generally controlled through the manufacturing execution system. Thus, dark grey arrows represent production-related data provided by the plant (e.g., resource status, job completion, or the performance measure of KPI in the current case), either gathered by the MES and stored as log data, or monitored on-line by, i.e., the simulation framework. Light grey arrows represent an interaction or information exchange, e.g., the *Decision-maker* might control the process of the production highlighted as *Reaction*) of the plant by the MES system. In a real-world application, the three main distinct operation modes follow each other during operation.

In contrast to the on-line proactive mode of the simulation, in the off-line scenario, simulation is applied in combination with the MES log data for setting up and parameterizing statistical learning prediction models (James et al. 2013) represented by *Statistical learning, Prediction and classification models* in Figure 1. Once the prediction models are set as an off-line analysis, permanent on-line simulation analysis of the manufacturing system is performed (highlighted by *Time* in the bottom right corner of Figure 1). This leads to a rolling horizon monitoring of the productions systems’ behaviour (e.g., by monitoring preselected performance measures of interest, e.g., LT of jobs) in advance by using prospective simulations. When a

1. provide reliable input for a simulation-based analysis of the system's capacity (these data are used by the robust capacity planning method);
2. directly utilize the outcomes, i.e., improve machine level robustness and performance by exploring capacity losses and proposing improvements.

The proposed methods are tested on a real production system, as a part of a factory producing complex electro-mechanical components for the automotive industry. On the selected flexible flow assembly line, the products require several manual and automated operations and comprehensive testing. There are more than 150 product variants produced on the line with different routings and operator control logics for the different main product families, with usually 6 – 8 setups per shifts.

After a performance analysis of the assembly line it could be stated that the planned and actual capacities are highly different and, furthermore, frequently changing from shift to shift. The main reason for that is the high variance of reject rates (also identified within one certain product code, caused by the high number of diverse parameters influencing the quality), making the testing machine a bottleneck. Due to safety criteria, the testing is done for 100% of the products, as, for instance, a rework and retest of the failed products cause unpredictable capacity requirement in the current situation. This results in a constant overriding of the production plan and schedule, causing backlogs, and in parallel a capacity loss, when operators cannot assemble parts due to insufficient testing capacity.

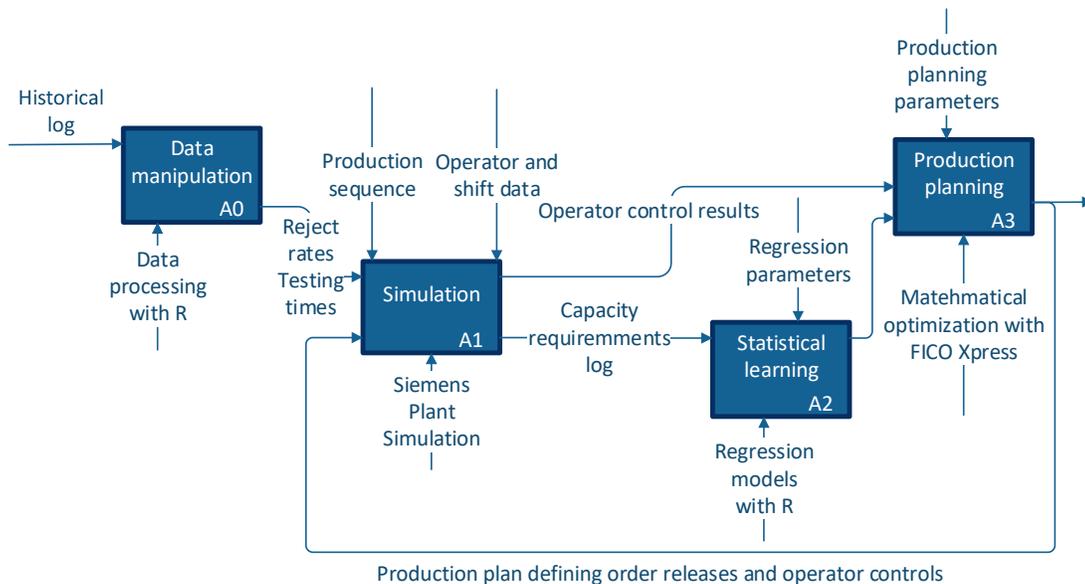


Figure 2: IDEF0 diagram of the planner and the failure analysis method.

Thus, a robust production planning method must consider in parallel both the static workload for the different products and also the stochastic behaviour of the production. The idea is to formulate shift patterns regarding different reject rate behaviours, consequently, enabling e.g., smoothing the workload of the operators by a proper mid-term planning and control policy (e.g., combine products with high expected workload and low reject rate and schedule them in one shift). As the second goal set previously, the solution should provide a better control on the reject rates and should increase the utilization of the test bench as well.

The method described in this section is responsible for the semi-automatic data manipulation and pre-processing operations, integrated into the simulation framework. The solution is a tool based on R, a statistical programming language and analysis framework.

3.2 Proposed Statistical Methods and Algorithms Developed

The comprehensive full test of each product means checking approximately 40 features in a predefined sequence of 20 steps. If any of the steps fails, i.e., the measured value is out of the specification limits, the product will fail the complete test and will be forwarded to a rework station. A failure code will be provided in the log, referring to the failure, together with the duration of the testing, added to all the other features measured so far. E.g., an analysis of a yearly amount of logs was performed on approximately 180,000 of these log entries. The raw data source is a set of plain text files (logged daily), which must be joined together in one csv file. After performing data pre-processing, output files and charts are formulated as highlighted in Figure 3.

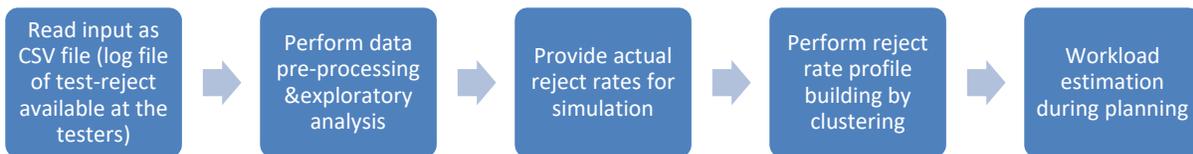


Figure 3: Main process flow of the test-reject analysis.

As part of an exploratory data analysis, the method provides an overview of the testing log data pre-processed. Figure 4a highlights a typical histogram of the failure code distribution of one selected product code. Most likely failure code 503 (6th bar from the left) will be the reason for rejecting the product, meaning an average of 120 seconds when this failure occurs during the test (square in the same bar). For this product code, the average testing time for a good product (passes all the testing steps) is 215 seconds (cross in the first bar).

In Figure 4b, the stochastic behaviors of the reject rates are highlighted. A simple linear regression does not effectively include the high spread of the reject rate, therefore, a calculation simply combining the reject rate with the batch size would mislead the capacity calculation. Regarding, e.g., product family A800 with 38 product codes (variants), the mean of the reject rates is 0.205 while the standard deviation is 0.072. All values presented here are scaled and all product names are changed in order to hide real company data.

As the outcome of the exploratory data analysis, it can be assumed that many of the product codes (within one product family) have a particular “failure behaviour”. Therefore, as the commonly applied method for an exploratory analysis, unsupervised learning is applied. Unsupervised learning is a set of statistical tools intended for the setting in which only a set of features X_1, X_2, \dots, X_p are given, measured on n observations, while an associated response variable Y is missing, thus prediction being impossible.

3.3 Computational Experiments and Validation

For each product family, a pairwise hierarchical clustering is performed in order to find subgroups both for similar product codes and failure codes. Figure 5 presents the heat maps obtained for a selected product family (data is suggested to be filtered out for product codes having at least 1,000 log data entries). Data have to be scaled and centred. Therefore, each product has the same weight. The agglomeration method applied in the tool is the complete linkage method.

In the hierarchical clustering algorithm, two dissimilarity measures are applied: the frequencies of failure codes (Figure 5a) as well as the product of test-fail time ratios and the frequencies of the failure codes (Figure 5b). The test-fail time ratio is calculated by dividing the average testing duration of a certain failure code by the total duration of a good products’ testing (full length, represented by a horizontal grey line in Figure 4a). The cutting heights of the products’ relevant trees are set to have six different groups for both heat maps. By applying the first dissimilarity measure, the results (Figure 5a) reflect anomalies of the

testing procedure. There are groups of products having similar failure profiles, which, according to the testing specialist at the company, gives the basis of a more-detailed analysis and readjustment (harmonization) of the specification limits for products with an identical function. The second dissimilarity measure results in a heat map (Figure 5b) containing also the “importance” of the failure codes, i.e., an early recognition of a failure during testing means less lost time of testing capacity.

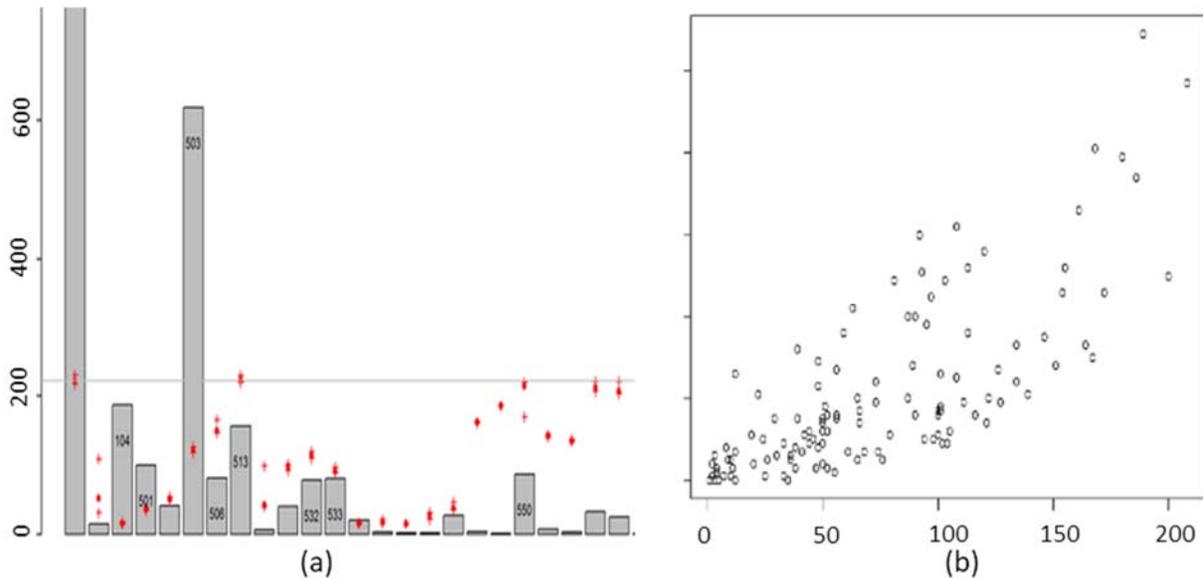


Figure 4: Characteristics of the failure code histogram for a selected product code. Grey bars represent the frequency, red squares the average time associated to the failure code and red plus signs the upper and lower quartiles. b) Stochastic behaviour of the reject rate for a selected product code. The horizontal axis represents the number of products in one batch and the vertical axis gives the rejected tests within the batch.

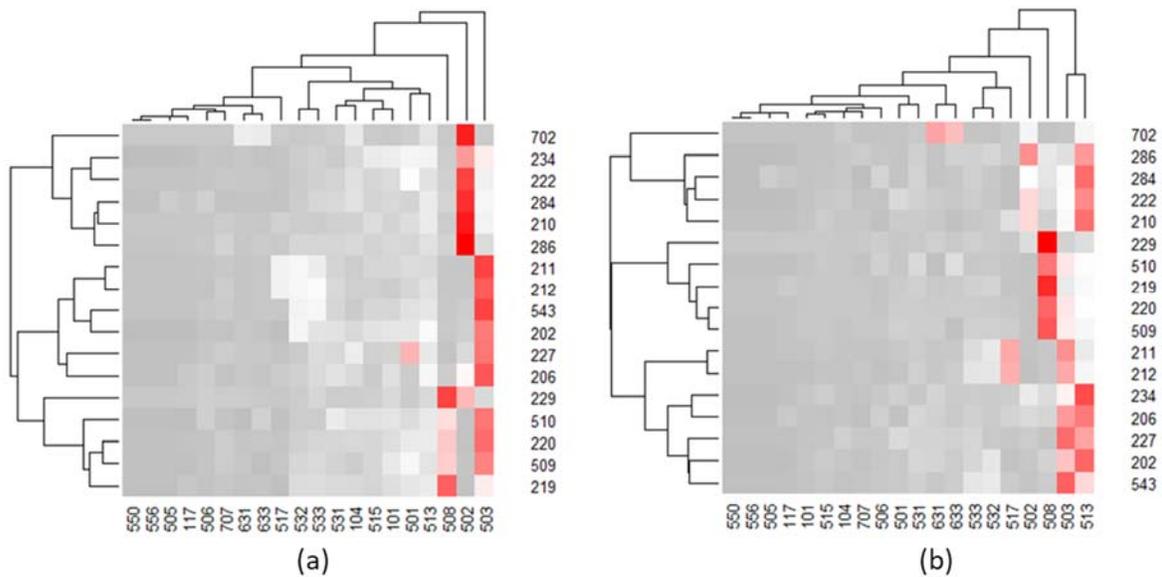


Figure 5: Heat maps (a) and (b) by hierarchical clustering product (y-axis) vs. failure codes (x-axis), applying two different dissimilarity measures. Red means a higher, while grey a lower value.

Applying the grouping of this latter clustering, it is assumed that product codes could be handled together during the planning phase, mainly having similar key features influencing (and describing) the stochastic behaviour of their capacity requirement (Table 1). The ratio of average test-fail time and average testing time is given in column five (FTr). Together with the reject rate (Rr) these two factors highly influence the total loss time portion (TLp), the time portion of testing capacity used for testing failed products. The main focus of the improvements is given to the minimization of TLp.

As an example of use, it can be stated that resulted product groups collect product codes with similar key features, e.g., group 1, 4 and 5, and finds outliers as group 3.

Table 1: Results of the grouping by the key features of the selected product family.

Prod. ID	Gr. ID	Avg. testing time [s]	Avg. test-fail time [s]	FTr	Rr	TLp
P286	1	274,84	108,06	0,411	0,17	0,076
P284	1	282,87	128,67	0,478	0,16	0,080
P222	1	271,84	97,15	0,375	0,19	0,079
P210	1	275,08	109,58	0,415	0,14	0,061
P234	2	284,73	109,74	0,416	0,13	0,053
P543	2	270,99	119,89	0,459	0,17	0,085
P227	2	280,92	108,23	0,413	0,13	0,056
P202	2	259,24	119,19	0,469	0,19	0,097
P229	3	290,31	158,28	0,591	0,25	0,152
P510	4	271,53	141,83	0,541	0,12	0,066
P509	4	273,9	136,16	0,520	0,17	0,094
P220	4	272,26	140,89	0,540	0,17	0,093
P219	4	273,57	148,64	0,565	0,16	0,093
P212	5	273,12	137,33	0,522	0,22	0,122
P211	5	272,44	132,31	0,503	0,21	0,114

4 WORKLOAD ESTIMATION WITH SIMULATION

In flexible, manually operated assembly systems (as introduced in Section 3), the prediction of estimated workload is often complicated, due to the variety of product types and the deviation of processing times. Though, stochastic or robust optimization models can be applied to cope with non-deterministic parameters, requiring high computation efforts and special solver algorithms. Additionally, diverse reject rates of the product variants and the varying rate of rework also increase the complexity of planning models.

The workload control of the selected assembly line (Figure 6) specifies the assignment of the operators with the assembly tasks. In this case, the general scheme of the assembly lines is applied, in which three main tasks are distinguished: assembly, rework, and final assembly.

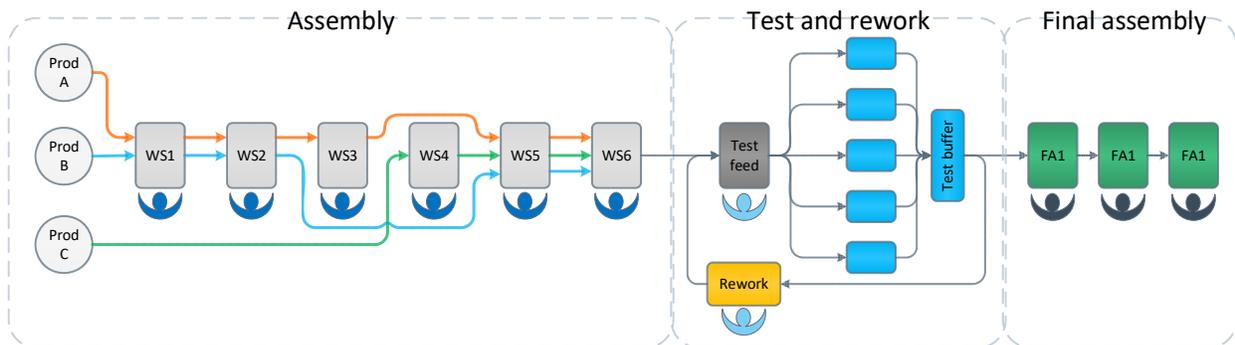


Figure 6: General process description of the assembly line under consideration.

The workload control takes the operator headcount as an input and specifies the operator-task assignments, considering that more tasks can be assigned to a single operator. The assignments are many-to-many type ones, which means that an operator might perform several tasks and a given task can be assigned to more than one operator. In order to determine the proper workload controls for each product and possible headcount, a discrete-event simulation (DES) model of the assembly line is applied (highlighted as A1 *Simulation* in Figure 2). Even though state-of-the-art assembly systems are usually equipped with advanced sensor networks, the real workload of the operators is hard to monitor. The DES model of the assembly line can provide reliable results about the workloads, and several various control policies can be evaluated. The main advantage of using simulation for such purposes is its capability to represent the stochastic nature of important processes, i.e., using the reject rate profiles for a certain product mix to be processed. In order to select the proper workload control, several random generated operator control scenarios are analysed (Table 2). The main outputs of the simulation analysis are the utilization of the operators and the performance of the line. By this way, the proper controls can be selected for each product mix and operator headcount (Table 3).

Table 2: Operator control scenarios defined for the simulation experiment analysing the workload under fluctuating reject rates per product mix. A: assembly, R: rework and testing, F: final assembly.

#	OP1	OP2	OP3	OP4	OP5	OP6	SUM
1	A	A	FA	FA	R		5
2	A	A	FA	FA	R	A	6
3	A	A	FA	FA	R + A		5
4	A	A	FA	FA + A	R		5
5	A	A + R	FA	FA			4
6	A	A	FA	FA	R	A	6
7	A	A	FA + A	FA + A	R		5

Table 3: Results of the simulation analysis of estimated workload and operator utilization on a selected product mix.

#	OP1	OP2	OP3	OP4	OP5	OP6	THP	Virtual Takt	Takt	Performance	Avg	Dev
1	100	100	58	41	37		96100	1,50	1	67%	67	30,7
2	100	100	80	69	55	100	144130	1,00	1	100%	84	18,9
3	100	100	71	55	100		121630	1,18	1	84%	85	20,8
4	100	100	99	100	49		127860	1,13	1	89%	89	22,6
5	100	100	52	33	0		81064	1,78	1	56%	71	43,4
6	100	100	80	69	55	100	144130	1,00	1	100%	84	18,9
7	94	94	100	100	48		124418	1,16	1	86%	87	22,0

5 CONCLUSIONS AND FUTURE WORK

Concluding, “failure profiles” were formulated by using the grouping, providing expected test-fail times and frequencies. Consequently, a cluster analysis could be applied as an input for the simulation. As a final goal in the future, it could be applied in the robust production planning method. The ongoing work also includes creating reject rate profiles based on the clusters and a comprehensive time-series analysis. Thus, regression models could be applied for predicting expected reject rates based on the recent system status.

Future work in this direction includes the extension of the methods for dynamic, “real-time” reject rate estimation and prediction, based on continuous data streaming. In this case the main problem is the

unpredictable behaviour of rejected tests within one corresponding lot (Figure 7), making real-time reject rate calculation difficult.

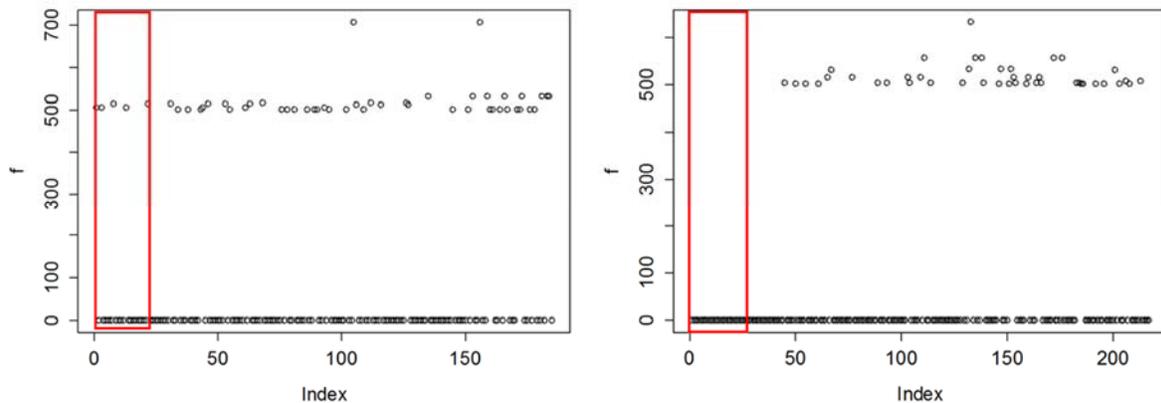


Figure 7: Distribution of test rejects within two selected lots, based on log data analysis.

ACKNOWLEDGEMENT

This research has been partially supported by the GINOP-2.3.2-15-2016-00002 grant on an "Industry 4.0 research and innovation centre of excellence" as well as by the European Commission through the H2020 project EPIC (<https://www.centre-epic.eu/>) under grant No. 739592.

REFERENCES

- Bagchi, S., C. Chen-Ritzo, S. T. Shikalgar, and M. Toner. 2008. "A Full-factory Simulator as a Daily Decision-Support Tool for 300mm Wafer Fabrication Productivity". In *Proceedings of the 2008 Winter Simulation Conference*, edited by S. J. Mason et. al., 2021-2029. Piscataway, New Jersey: IEEE.
- Banks, J. 1998. *Handbook of Simulation: Principles, Methodology, Advances, Application and Practice*. New York: John Wiley & Sons.
- Cowling, P. and M. Johansson. 2002. "Using Real Time Information for Effective Dynamic Scheduling". *European Journal of Operational Research* 139(2):230-244.
- Honkomp, S.J., L. Mockus, and G.V. Reklaitis. 1999. "A Framework for Schedule Evaluation with Processing Uncertainty". *Computers and Chemical Engineering* 23(4-5):595-609.
- James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning*. New York: Springer.
- Kim, Y., D. Lee, J. Kim, and H. Roh. 1998. "A Simulation Study on Lot Release Control, Mask Scheduling, Batch Scheduling in Semiconductor Wafer Fab Facilities". *Journal of Manufacturing Systems* 17(2):107-117.
- Law, A.M. and D. Kelton. 2015. *Simulation Modelling and Analysis*. 5th ed. New York: McGraw-Hill.
- Lin, J.T., F. Wang, and P. Yen. 2001. "Simulation Analysis of Dispatching Rules for an Automated Interbay Material Handling System in Wafer Fab". *International Journal of Production Research* 39(6):1221-1238.
- Monostori, L., B. Kádár, A. Pfeiffer, and D. Karnok. 2007. "Solution Approaches. to Real-time Control of Customized Mass Production". *CIRP Annals – Manufacturing Technology* 56(1):431-434.
- Pfeiffer, A. 2007. *Novel Methods for Decision Support in Production Planning and Control*. Ph.D. Thesis, Budapest University of Technology and Economics.

- Pfeiffer, A., B. Kádár, L. Monostori, Z. Vén, and G. Popovics. 2011. "Co-analysing Situations and Production Control Rules in a Large-scale Manufacturing Environment". In *Proceedings of the 44th CIRP International Conference on Manufacturing Systems - New Worlds of Manufacturing*, 3-5.
- Pfeiffer, A., D. Gyulai, B. Kádár, and L. Monostori. 2016. "Manufacturing Lead Time Estimation With the Combination of Simulation and Statistical Learning Methods". *Procedia CIRP* 41:75-80.
- Rabelo, L., M. Helal, A. Jones, J. Min, Y.J. Son, and A. Deshmukh. 2003. "A Hybrid Approach to Manufacturing Enterprise Simulation". In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick et. al., 1125-1133. Piscataway, New Jersey: IEEE.
- Rose, O. 2007. "Improved Simple Simulation Models for Semiconductor Wafer Factories". *Proceedings of the 2007 Winter Simulation Conference*, edited by J. Tew et al., 1708-1712. Piscataway, New Jersey: IEEE.
- Sabuncuoglu, I. and O. M. Kizilisik. 2003. "Reactive Scheduling in a Dynamic and Stochastic FMS Environment". *International Journal of Production Research* 41(17):4211-4231.
- Sivakumar, A.I. and C.S. Chong. 2001. "A Simulation Based Analysis of Cycle Time Distribution, and Throughput in Semiconductor Backend Manufacturing". *Computers in Industry* 45(1):59-78.
- Watt, D.G. 1998. "Integrating Simulation-based Scheduling with MES in a Semiconductor Fab". in *Proceedings of the 1998 Winter Simulation Conference*. edited by D. J. Medeiros et al., 1713-1715. Piscataway, New Jersey: IEEE.

AUTHOR BIOGRAHPIES

ANDRÁS PFEIFFER earned his PhD in 2008 at the Budapest University of Technology and Economics. Currently he is a senior research fellow at the Laboratory on Engineering and Management Intelligence of the Research Institute for Computer Science and Control (SZTAKI) of the Hungarian Academy of Sciences (MTA), deputy director of SZTAKI, and industrial systems expert at EPIC InnoLabs Ltd. His current interest includes decision support in production planning and control, simulation and emulation modeling of complex production systems, self-building simulation systems, statistical analysis, and statistical learning methods.

DÁVID GYULAI is graduated from Budapest University of Technology and Economics as a Mechatronics Engineer in 2012. He wrote his master thesis in design and planning of reconfigurable production systems. Currently, he is a research associate at the Laboratory on Engineering and Management Intelligence of the Research Institute for Computer Science and Control (SZTAKI) of the Hungarian Academy of Sciences (MTA) and industrial data analytics expert at EPIC InnoLabs Ltd. He is also a PhD candidate at Budapest University of Technology and Economics focusing on capacity and production planning topics. He is interested in digital enterprise technologies, production and capacity planning, statistical analysis, and statistical learning methods.

ÁDÁM SZALLER is graduated from Budapest University of Technology and Economics as a mechatronics engineer in 2017. His master thesis is about production system layout planning considering stochastic effects. Currently, he is working as a systems engineer at the Laboratory on Engineering and Management Intelligence of the Research Institute for Computer Science and Control (SZTAKI) of the Hungarian Academy of Sciences (MTA) and the EPIC Centre of Excellence in Production Informatics and Control at SZTAKI. He is also a PhD student at the Budapest University of Technology and Economics. His research is about resource allocation in distributed production systems. His range of interest is simulation of manufacturing systems, resource sharing in agent-based systems, statistical learning methods, and digital enterprise technologies.

PROF. LÁSZLÓ MONOSTORI is the director of the Research Institute for Computer Science and Control (SZTAKI) of the Hungarian Academy of Sciences (MTA), and director of the EPIC Center of

Excellence in Production Informatics and Control. He is also full time professor at the Department of Manufacturing Science & Technology at the Budapest University of Technology and Economics. He is a Fellow and former President of the International Academy for Production Engineering (CIRP) and Full Member of the European Academy of Industrial Management (AIM), International Federation of Automatic Control (IFAC). He is Editor-in-Chief of the CIRP Journal of Manufacturing Science and Technology; Associate Editor of Computers in Industry, as well as the Measurement, and member of the editorial boards of other international scientific periodicals. For his research achievements published in more than 370 publications resulted in about 1,800 independent citations and for his development activities – among others – the Dennis Gabor Prize was given to him in 2004. Prof. Monostori is a member of the Hungarian Academy of Sciences and member of the Hungarian Academy of Engineering.