Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

SIMULATION BASED APPROACH TO CALCULATE UTILIZATION LIMITS IN OPTO SEMICONDUCTOR FRONTENDS

Falk Stefan Pappert
Oliver Rose

Department of Computer Science
Universität der Bundeswehr München
Werner-Heisenberg-Weg 39
Neubiberg, 85577, GERMANY

Fabian Suhrke
Jonas Mager

Osram Opto Semiconductors GmbH
Leibnizstraße 4
Regensburg, 93055, GERMANY

ABSTRACT

Capacity planning is a crucial task for economically sound production. Especially in semiconductor manufacturing, as equipment is expensive and production complex. An important part of valid capacity planning is a good understanding of equipment capabilities and characteristics and their influence on the workflow. Traditional approaches require new analysis with changing situations in the fab, which require special expertise and time. To enable a companywide standard and provide an easy to use tool, we are developing a utilization limit estimation tool. In this paper, we present our approach for a utilization target estimation system which bases its estimation on a wide range of data points created by data farming. Then, we apply a regression analysis to interpolate missing data points in order to provide fast estimates for utilization limits depending on equipment characteristics.

1 INTRODUCTION

A valid capacity planning model is a fundamental prerequisite for all operative planning activities in fab management as well as decision support for long-term business cases. This includes day-to-day business challenges like how to change the utilization limit if the product mix changed or estimating the impact of an operative improvement on the throughput of a certain equipment group. In addition, each strategic question should generally be based on a capacity model of reasonable accuracy. Typical topics are equipment procurement, capacity expansions, product ramp ups or shifting products between different sites (Robinson et al. 2003). Due to the highly competitive nature of the semiconductor industry, there are conflicting goals to balance. First of all, equipment should be utilized as much as possible as equipment and clean room space are rather expensive. Second, cycle times should be minimal to have an ideal time to market for customers and to reduce development cycles. A third goal is to reduce the inventory, e.g. work in progress (WIP), as unfinished material ties up capital.

Robinson et al. (2003) show on the one hand why accurate capacity planning is so important and on the other hand why it is so difficult within the highly sophisticated semiconductor industry. This is a result of typical wafer fab characteristics such as a wide product range, high variation of product flows combined with diverse numbers of production layers, batching, tool breakdowns, tool dedication, rework and a variating mix of single wafer and batch processes. For an appropriate planning model that has to generate reliable capacity limits, it is essential to consider these ‘capacity loss factors’.

To tackle these challenges, a co-operative project was started by OSRAM Opto Semiconductor and the Universität der Bundeswehr, München, with the aim to rise the capacity planning quality. The main objective of the project is the development of a global capacity model to estimate utilization limits for each equipment group. The goal is to provide company-wide decision support for frontend planning based

on a single standardized approach. To outperform the current static capacity planning model, which is only considering very few parameters, we include main factors influencing fab capacity. In addition to consider more capacity loss factors we generally decide that a dynamic planning approach is essential to receive sufficient utilization limits. With the improved planning model we are able to determine utilization limits for our target flow factors more accurate. Moreover, a further improvement is the classification of equipment groups by their financial value, which gives us the opportunity to minimize the overall investment in our fabs. Our combined approach of data farming and multivariate regression analysis leads to a number of advantages, i.e. reduce computation time in comparison to pure simulation based models, allows for simple changes without recalculation for a varied product mix or equipment pool, enables continuous selection for the most of our considered factor levels and is therefore very convenient to use in day to day operations. The project is still ongoing and we are currently in the process of data farming. Therefore, the results and ideas given in Section five are based on sample data and reflect our current understanding of the mathematical challenges we will be facing with the solutions we are planning.

Following the idea to plan capacity based on the operating curve (Eichholz and Schömig 2007) of a facility or work center and a target flow factor, there is still the issue of accurately fitting the operating curve to a work center in question. There are different approaches to estimating the operating curve of a production unit. Byrne (2011) presents an overview on analytical methods as well as an approach by simulating sample points. These approaches focus on estimating a single operating curve for a specific system. The goal of our work differs in two major aspects from traditional operating curve management. First of all, we are not interested in the whole operating curve but only in points where specific flow factors are exceeded because these are the target utilizations to be used in planning. A second aspect is to avoid running new analyses or simulation studies whenever significant changes to a work center happen or new work centers are created. To overcome this issue, we aim to create a tool which bases its estimation on a wide range of data points created by data farming and then using a regression analysis to interpolate missing data points. It will categorize an equipment based on equipment characteristics and then provide estimated utilization thresholds as a response.

The tool capacity planning approach we use is based on different equipment categories which are not defined by their logistical behavior but by their financial value. This is done to optimize the use of the overall investment in the fab. Very expensive pieces of equipment are expected to run at higher utilizations were we have to accept higher flow factors. In contrast, less expensive tools are expected to handle material at much lower than target flow factors. As these categories are mainly determined by equipment and process pricing which is not necessarily influenced by logistical characteristics of these tools, we are not just looking at a single target utilization for all equipment but for a target utilization for each equipment category. By assigning tools to these different categories, we aim to achieve a trade-off between pricing and overall target flow factor. This paper is focused on discussing our approach on determining reasonable utilization limits based on equipment characteristics.

In this paper, we will first give an overview of the considered capacity loss factors and levels used in our simulation models. Afterwards, we will present the data farming and simulation used to generate a data base. In Section four we will discuss our planned nonlinear regression analysis to generate a sophisticated, fitted planning model. After a short conclusion, we will discuss our next steps in the project.

2 CAPACITY FACTORS

To identify the capacity-relevant factors for OSRAM Opto Semiconductor, we discussed the existing planning methods and reviewed the currently considered factors. During a workshop, planning and industrial engineering experts using – among others – Robinson et al. (2003) and Hopp and Spearman (2008) as a starting point, defined the relevant factors for our production systems. These factors including short descriptions and the effects due to the planning capacity are listed below in Table 1. Column three

of Table 1 shows the number of levels for each factor used for the simulation-based analysis. The level type in column four distinguishes between quantitative values (quant.) and categorical levels (cat.). The categorical levels are ‘low’, ‘medium’ and ‘high’.

Table 1: An overview of capacity relevant factors for the planning model.

Factor name	**Description / Influence on equipment capacity**	**Number of factor levels**	**Level type**
Number of equipment	The number of equal equipment in a workshop has a significant influence on the robustness of performance in the case of breakdown. Particularly one of a kind tools can cause fab wide issues if highly utilized but broken.	7	quant.
RPT	Describes the mean raw process times (RPT) of the processes performed by an equipment group. The higher the RPT are the lower the overall throughput.	6	quant.
Batching	Describes processes with a combination of more or less wafers than the common lot size. Batching can lead to higher queueing time before process start due to the time needed to wait for valid batches (see Kuik et al., 1994).	7	cat.
Setup	Describes the mean time to change tool parameters or to switch consumables for different processes (see Zhou and Egbelu, 1989).	3	quant.
Rework	Describes the mean rate at which material has to be reworked because of production quality issues. This can increase the workload of effected tools significantly.	3	quant.
Dedication	Describes the selection limitations to designated tool sets due to process specific restrictions or quality issues (see Pappert et al., 2016).	3	cat.
Maintenance	Describes the planed tool downs to maintain equipment (see McKone et al., 2001).	3	quant.
Breakdowns	Describes the unplanned tool downs (see Logendran and Talkington, 1997, and Chiu et al., 2010).	3	quant.
Product mix	Describes the percentage of different products on a work center. A high product mix results in high process variability which is associated with capacity loss.	3	cat.

To best fit the factor levels to our production system, we analyzed the fab dataset. One way to identify some natural clusters is to plot the cumulative density function (cdf). Figure 1 shows the daily rework ratio cdf plot for a certain equipment group as an example. There are three clearly visible clusters separated by a red line which could be used as factor levels for rework.

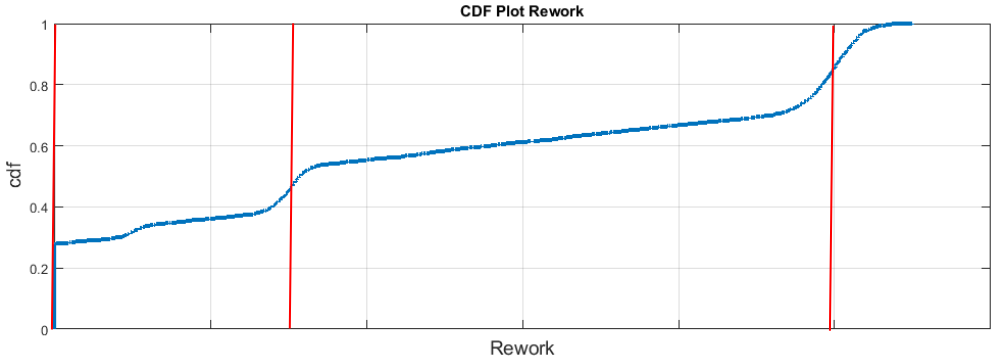

Figure 1: A cumulative density plot of daily-rework ratio for a certain equipment group.

A more mathematical approach is to apply a cluster analysis to find all density points. For instance, with a k-means clustering we were able to scale the amount of factor levels and, as a consequence, the duration of the overall simulation time. However, with simulation data on hand we aim to interpolate all unknown data points with a regression analysis. Thus, we only need to simulate a sufficient number of supporting points.

3 DATA FARMING AND SIMULATION

The goal of our data farming is to find reasonable equipment utilization limits for each factor level combination with respect to given target flow factors (FF). The use of other key performance indicators (KPI) is under discussion and may be included in the future. The general idea here is to generate models for each design point and run them at different utilization points. This is done until we find the threshold utilization for each equipment category where the FFs are still within our bounds while the next higher utilization point (minimal step size is 1%) is violating our FF limits. Once we have found these thresholds for all categories the next design point is evaluated. Thereby, we create the base data for the following regressing model.

Throughput and performance with regard to FF is heavily depended on the factor levels. Equipment which is processing batches consisting of hundreds of wafers will be able to handle much more material before reaching similar flow factor levels than single wafer tools or tools which only process a few wafers at a time (see Hopp and Spearman, 2008). We therefore start the analysis of every design point by performing a static capacity analysis for the system at hand. This static capacity analysis provides results about fixed capacity losses mainly caused by breakdowns and utilization-based capacity uses depending mostly on equipment characteristics like batching and processing time. This allows us to calculate necessary arrival rates to achieve specific utilization points.

After this preparation step, we search among valid utilization points to find the utilization thresholds for all defined equipment categories. As computation time is a very critical factor in data farming, we use a search strategy akin to binary search to reduce the number of evaluated utilization points. To further reduce computational cost, simulation results gained during the search for one category are evaluated with respect to other categories before starting any new simulation runs for them. As our model contains some stochastic influences a number of repetitions has to be performed for each utilization point.

The simulation software used is a factory simulation package developed at the Universität der Bundeswehr with the purpose of simulating and evaluating complex job shops, with a special focus on supporting very complex equipment characteristics and material control strategies.

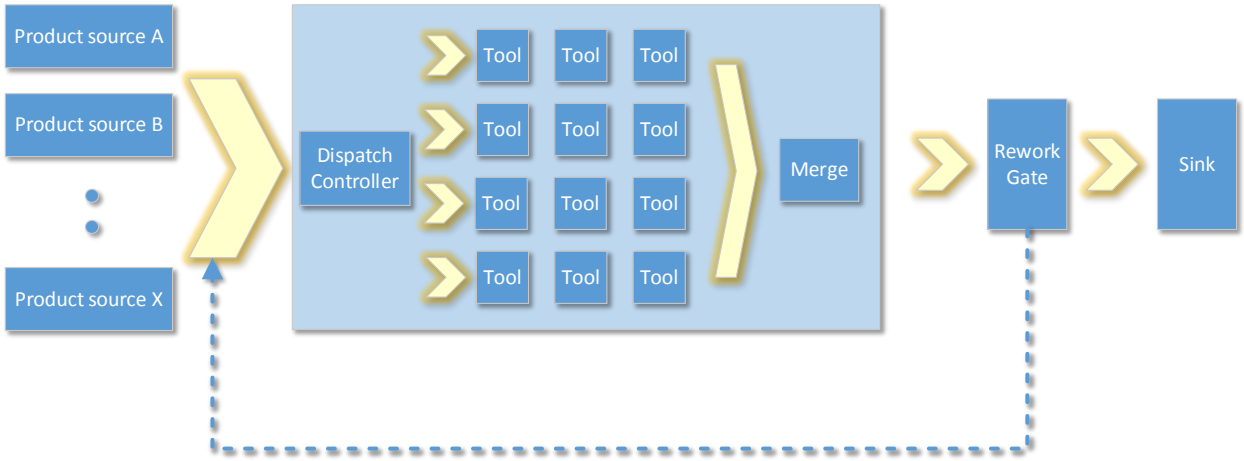

Figure 2: The structure of the simulation model.

The models are generated automatically based on given factor levels and target utilizations. The general structure of the principal model we are using is shown in Figure 2. It consist of a number of sources, one for each product, generating the stream of arriving material, according to the desired product mix. An equipment group with a number of tools is the central component of the simulation. The Dispatch Controller considers the relevant factors for dispatching, e. g. batching, setups and dedication. Other factors like breakdowns and maintenance are handled directly within the tools.

Based on the level of rework required in a particular setting, a rework gate is introduced which is routing part of the processed material back to the machine buffer for another processing run. Dispatching, batching, and setup avoidance strategies are currently kept the same for all runs.

With this large number of data points (>210000), computation time becomes a very important issue. Depending on the design points, replications for a single utilization point are between 20ms and 90s on a desktop PC. The longest simulations are the ones which feature large numbers of tools with large batch capacities. They handle a lot more lots and therefore events during the same simulated time period. Besides the sheer number of runs utilization plays a major role in determining simulation run time. Higher target utilization simulation runs typically take longer. Our static capacity analyses and search strategy help considerably in reducing the number of runs and in avoiding highly utilized runs. Although a desktop PC is sufficient for running smaller batches of test sets and to create sample data, the actual data farming runs are performed on a computing server with 30 cores to reduce waiting time for results.

A big issue for all simulation projects is verification and validation (V&V), which is especially difficult with respect to data farming. It is simply impossible to evaluate the simulation results of each and every design point manually, especially considering that there are a number of utilization points and replications for each of them. We therefore have a two way approach to V&V in this project. First of all, we borrowed from software engineering where during the development of the simulation software itself numerous unit tests were created and are constantly run and evaluated to ensure that changes to the system do not influence previously correct behavior. In addition, unit tests are also used to check against oddities which came up during the development for this specific project. For example, this includes testing for result limitations which are based on the factor levels we allow. E.g., the breakdown factor levels do not include tools having no breakdowns at all and therefore all simulation runs should have

unscheduled downtimes. Both test case bases are extended continuously throughout the project. The second major part of our V&V strategy involves a panel of experts in industrial engineering and production planning coming from different work backgrounds. These experts are confronted with sample simulation results and asked to validate them based on their experience. Furthermore, a set of key equipment groups was chosen by the expert panel. For these key equipment groups, the experts were asked to assign parameter sets reflecting our factor levels and to evaluate the simulation results against real equipment data.

4 REGRESSION ANALYSIS

To reduce the number of simulation runs and to be able to interpolate unknown data points, we apply a multivariate regression analysis. With simulation data on hand, the main goal is to generate a smooth curve with minimal difference to the simulated data points $x_{i,*}$ without outliers. The independent variables $x_{*,j}$ influence the dependent variable U_{lim} (Utilization limit) as shown in Equation (1)

$$U_{i,lim} = f(x_{i,1}, x_{i,2}, \dots, x_{i,n}) \tag{1}$$

In the current project phase, there are only some first test data sets. Therefore, we are not able to show the results of our regression analysis yet. However, we will explain our planned approach with its steps and methods of nonlinear regression analysis in this chapter.

As a result of nonlinear dependencies within our simulation model, i.e., batch size vs. cycle time (see Hopp and Spearman, 2008), we have to use nonlinear regression models. In this chapter, we give a short overview on the techniques of robust nonlinear regression and modeling.

The utilization limit is defined as the minimal utilization which violates at least one KPI limit. The critical flow factors vary for basic, common and major equipment. Therefore, we calculate separate parameter sets for each equipment category. The categorical variables, e.g., dedication and product mix, are coded as dummy variables. For model selection it is necessary to obtain an overview on the dataset. Therefore different plots are created to find a promising model. First, a scatter plot is used to identify interactions. Second, a plot which holds all independent variables but one constant helps to identify the function family of each variable.

The model selection process has to be repeated until the results of the nonlinear regression are satisfying. For further details about this we refer to Royston and Sauerbrei (2008). A sample set of plots for a dataset with the same factor levels is shown in Figure 3. Some outliers are marked with a circle. The expected dependencies are clearly visible. For high utilization, the flow factor, daily-going-rate and the maximum Work-In-Progress (maxWIP) increase significantly.

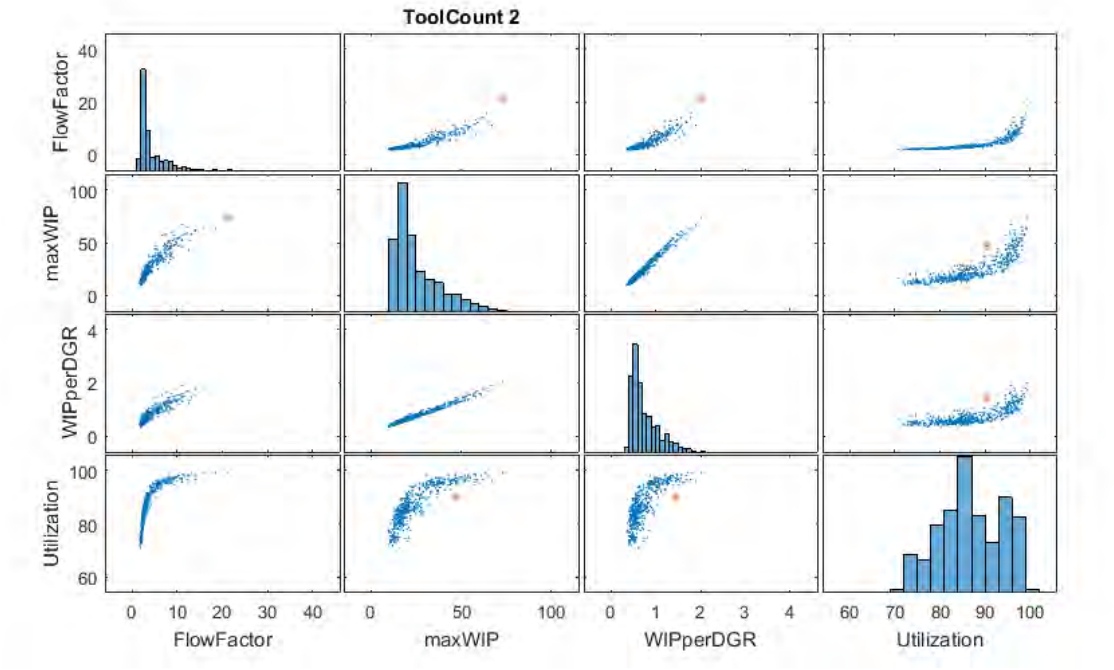

Figure 3: A scatter matrix plot of some simulated logistical key figures.

Møller et al. (2005) discussed why it is important to use robust mathematical techniques for data analysis. Therefore we apply a ROUT (robust regression followed by outlier identification) method described in Motulsky and Brown (2006) for a robust nonlinear regression and to identify outlier.

Due to the possibility of rare event combinations creating extreme results during simulation runs, we have to consider fortifying our further analysis against outliers. For planning proposal purposes, we do not want to calibrate the model for an unlikely scenario, even though the main goal is to fit the simulation data as good as possible. For this reason, we use a robust regression analysis.

The robust nonlinear regression is based on a discussion in Numerical Recipes (Press, et al. 2007) that the variation around the curve follows a Lorentzian distribution rather than a Gaussian distribution. The difference is that a Lorentzian distribution has wide tails. So, outliers are fairly common and have little impact on the fit.

For ordinary regression, the parameters p are optimal if Equation (2) is minimal

$$\sum_i \left[y_i - f\left(x_{i,1}, x_{i,2}, \ldots, x_{i,n}, p_0, p_1, \ldots, p_n\right)\right]^2 \rightarrow min \tag{2}$$

Unlike (2) the ROUT method has the goal to minimize (3)

$$\sum_i ln\left[1 + \left(\frac{y_i - f\left(x_{i,1}, x_{i,2}, \ldots, x_{i,n}, p_0, p_1, \ldots, p_n\right)}{RSDR}\right)^2\right] \rightarrow min \tag{3}$$

With RSDR defined as (4)

$$RSDR = P68\frac{N}{N-k} \tag{4}$$

With N being the number of data points, k the number of parameters, and P68 the 68.27 percentile of the absolute values of the residuals. For further details of the algorithm, we refer to Motulsky and Brown (2006).

An example of robust vs. non-robust regression is shown in Figure 4. The test model is described in Equation (5)

$$y = -e^{ax} + ln(\mathrm{b}x) + \mathrm{c}x^2 + eps,\ x\,\epsilon[0.1;2],\ eps \sim N(0,1) \tag{5}$$

with $\mathrm{a} = 2, \mathrm{b} = 3, \mathrm{c} = 2$. In Figure 4, this model is illustrated with a lot of additional noise/outliers. The robust regression captures the underlying parameters quite well. The least square fit, however, does not match the model adequately.

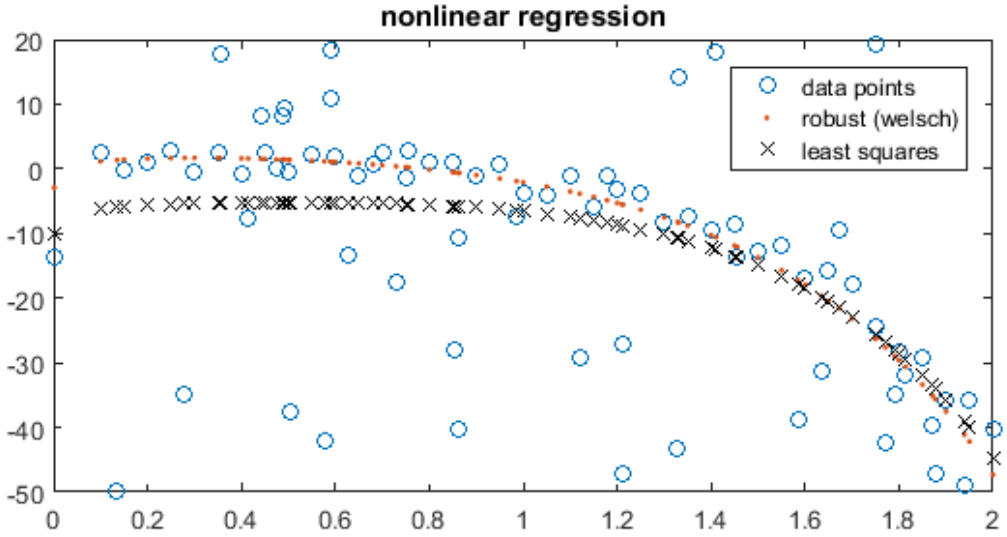

Figure 4: A comparison between robust vs non-robust regression of $\boldsymbol{y = -e^{2x} + ln(3x) + 2x^2}$.

As a next step, we have to define whether any parameter is constricted in any way. As mentioned earlier, we are only interested in a good and robust fit of our simulation data. Thus, our design of parameters is not constricted.

The intervals for our quantitative variables are based on production data and therefore there is no reason to extrapolate values considering our broad design of minimal and maximal factor levels. The codomain of U_{lim} is between: $0\% \leq U_{lim} = f(x_{i,1}, x_{i,2}, \dots, x_{i,n}, p_1, p_2, \dots, p_m) \leq 100\%$. Therefore, we design our model and parameter to stay in the codomain range for every point x.

Nonlinear regression models need an initial value for each parameter to start with. For small models, it is easy to estimate reasonable initial values fairly close to the real solution. For models with a lot of parameters, this can be difficult. However, Motulsky and Christopoulos (2004) point out that rough estimates for initial values are sufficient. With initial values on hand, the regression is performed and the result are reviewed. For further information we refer to Motulsky and Brown (2006).

In Figure 5, a robust nonlinear robust regression for flow factor vs. utilization is plotted with the factor levels 0% and 8% rework. As expected, the function family for different factor levels remains the same. As a regression model we used a rational function for each rework factor level.

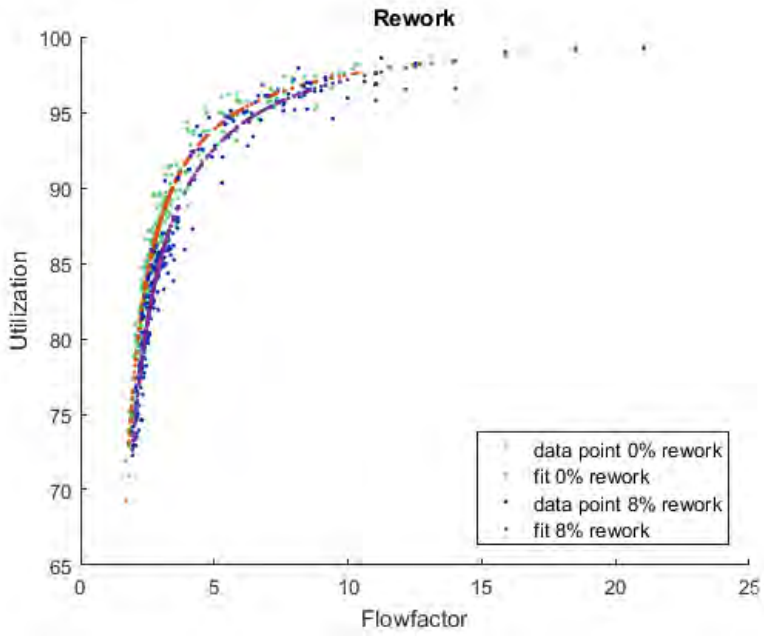

Figure 5: Some simulation data of flow factor vs utilization for two different rework level.

5 CONCLUSION

In this paper, we presented an approach to dynamically estimate utilization limits for work centers based on different equipment categories and factors influencing equipment capacity. In a first step, we use data farming for each factor level combination to find a set of data points. Afterwards, with the help of a multivariate regression analysis, we interpolate unknown level combinations to create an estimation tool for capacity planning. This gives OSRAM Opto Semiconductors the opportunity to determine sufficient utilization limits in a fast and convenient way without recalculation when the equipment pool or product mix changes.

After the implementation of the system, we are able to use mathematical analysis to have a better understanding about the factor interactions. With curve sketching it is possible to calculate the global optimum and return the gap of the optimal utilization limit for different factor levels. With a calculated gap on hand, the capacity potentials are well-known and experts can intervene at the right spots. Furthermore, we plan to apply a principal component analysis (PCA) to gain further knowledge about the influence of all considered factors on the utilization limit and their interactions.

In addition, we are looking into additional factors, e.g., time bound sequences which also have a high impact on the capacity limit. Together with factors like arrival rates, WIP levels, and dispatcher policies this will enable the system to identify operative potentials.

ACKNOWLEDGEMENTS

We would like to thank Anna Holm for her help with the visualization of the simulation data and Dr. Thomas Frey for the many fruitful discussions on the topic.

REFERENCES

Byrne, Néill M. "A Framework for Generating Operational Characteristic Curves for Semiconductor Manufacturing Systems Using Flexible and Reusable Discrete Event Simulations." Diss. Dublin City University, 2012.

Chiu, Y.S. P., F.T. Cheng, and H.H. Chang. 2010. "Remarks on the Optimization Process of a Manufacturing System with Stochastic Breakdown and Rework." *Applied Mathematics Letters* 23: 1152-1155.

Eichhorn D., Schömig A. 2007. "Betriebskennlinien-Management als Performancemessungs- und -planungskonzept bei komplexen Produktionsprozessen." In *Operations Research Proceedings 2006*, edited by K.H. Waldmann and U.M. Stocker. Springer, Berlin, Heidelberg.

Frosch Møller, S., J. von Frese, and R. Bro. 2005. "Robust Methods for Multivariate Data Analysis." *Journal of Chemometrics* 19: 549-563.

Hopp, W. J., and M. L. Spearman. 2008. *Factory Physics*, 3rd ed. Illinois: Waveland Press, Inc.

Kuik, R., M. Salomon, and L. N. Van Wassenhove. 1994. "Batching Decisions: Structure and Models." *European Journal of Operational Research* 75: 243-263.

Logendran, R., and D. Talkington. 1997. "Analysis of Cellular and Functional Manufacturing Systems in the Presence of Machine Breakdown." *International Journal of Production Economics* 53: 239-256.

McKone, K. E., R. G. Schroeder, and K. O. Cua. 2001. "The Impact of Total Productive Maintenance Practices on Manufacturing Performance." *Journal of Operations Management* 19: 39-58.

Motulsky, H. J., R. E. Brown. 2006. "Detecting Outliers when Fitting Data with Nonlinear Regression – a new Method Based on Robust Nonlinear Regression and the False Discovery Rate." *BMC Bioinformatics* 7: 123.

Motulsky, H. J., and A. Christopoulos. 2004. *Fitting Models to Biological Data Using Linear and Nonlinear Regression: A Practical Guide to Curve Fitting*, 1st ed. 17. New York: Oxford University Press.

Pappert, F. S., T. Zhang, J. Mager, F. Suhrke, and O. Rose. 2016. "Impact of Time Bound Constraints and Batching on Metallization in an Opto-Semiconductor Fab." In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka and S. E. Chick, 2947-2957. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Press, W. H., S. A. Teukolsky, and W. T. Vetterling, and B. P. Flannery. 2007. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, 3rd ed. New York: Cambridge University Press.

Robinson, J., and J. Fowler, and E. Neacy. 2003. "Capacity Loss Factors in Semiconductor Manufacturing." FabTime Inc. http://www.fabtime.com/abs_CapPlan.shtml.

Royston, P., and W. Sauerbrei. 2008. *Multivariable Model-Building*, 1st ed. Mississauga: John Wiley & Sons Canada Ltd.

Zhou, C., and P. J. Egbelu. 1989. "Scheduling in a Manufacturing Shop with Sequence-Dependent Setups." *Robotics and Computer-Integrated Manufacturing* 5: 73-81.

AUTHOR BIOGRAPHIES

FALK STEFAN PAPPERT is Research Assistant and PhD student at Universität der Bundeswehr as a member of the scientific staff of Prof. Dr. Oliver Rose at the Chair of Modeling and Simulation. His focus is on conceptual modelling approaches to simulation-based scheduling and optimization of production systems. He has received his M.S. degree in Computer Science from Dresden University of Technology. He is a member of GI. His email address is falk.pappert@unibw.de

FABIAN SUHRKE is Senior Engineer at OSRAM Opto Semiconductors. He is the project leader for fab simulation and real-time dispatching in frontend Regensburg. Furthermore he is responsible for the logistical concept of global epi-steering within the frontends. He holds an M.S. degree in mathematics from OTH Regensburg. His email address is fabian.suhrke@osram-os.com

JONAS MAGER is Industrial Engineer at OSRAM Opto Semiconductors in Regensburg. He is focusing on projects to increase fab performance with simulation models and is working on data analyses at the frontend production Regensburg. He holds a M.S. degree in industrial engineering and a B.S. degree in electrical engineering from OTH Regensburg. His email address is jonas.mager@osram-os.com

OLIVER ROSE holds the Chair for Modeling and Simulation at the Department of Computer Science of the Universität der Bundeswehr Munich, Germany. He received a M.S. degree in applied mathematics and a Ph.D. degree in computer science from Würzburg University, Germany. His research focuses on the operational modeling, analysis and material flow control of complex manufacturing facilities, in particular, semiconductor factories. He is a member of INFORMS Simulation Society, ASIM, and GI. His email address is oliver.rose@unibw.de.