KEDGE BUSINESS SCHOOL
ILS2016
Information Systems Logistics and Supply Chain
6th International Conference
université de BORDEAUX

SimData Toolbox: Simplifying the Analysis and Preparation of Input Data for Supply Chain Simulation

Martin Sutter[1], Thomas Ponsignon[2]

Abstract. This paper presents data analysis and data mining techniques to evaluate and generate input data for supply chain simulation. A toolbox called SimData has been developed and implemented by means of the simulation software AnyLogic. This toolbox includes objects for the analysis and visualization of data as well as for the generation of synthetic data. Therefore, two pattern recognition approaches are proposed: an extended Expectation-Maximization algorithm and a Peak-and-Noise detection algorithm. Examples based on real-world data of a semiconductor manufacturer are provided to show the applications of the proposed toolbox.

Keywords: Supply chain simulation, Simulation library, Synthetic data, Pattern recognition, Statistic analysis, Expectation-Maximization algorithm.

1 Introduction

Semiconductor fabrication is considered to be one of the most complex processes in today's industry with a highly competitive and volatile market, globally organized supply chains and short product life cycles, as pointed out by various authors [1-3]. This challenging environment makes semiconductor supply chain very attractive for academic research. One way to analyze and increase the performance of a semiconductor supply chain is by simulation [3]. Currently no software specialized for simulating semiconductor supply chains is available. [4] suggests a library of simulation components called SCSC-SIMLIB that is implemented with AnyLogic [5].

The input data procedure is one of the most critical and time-consuming phases in simulation projects. Major challenges, which lead to additional effort, relate to low quality data and massive manual workload to transform raw data into simulation input [6]. This paper is located in this context, providing an overview of a toolbox called SimData for analyzing appearance and structure of data serving as input for semiconductor supply chain simulations. SimData that is part of the SCSC-SIMLIB library, which is especially created for semiconductor supply chain simulations, contains objects that allow analyzing and identifying characteristics of datasets.

[1] Martin Sutter is with the Department of Supply Chain Innovation, Infineon Technologies AG, 85579, Neubiberg, Germany
martin.sutter@infineon.com
[2] Thomas Ponsignon is with the Department of Supply Chain Innovation, Infineon Technologies AG, 85579, Neubiberg, Germany
thomas.ponsignon@infineon.com

Starting with visualization tools, the paper will give an overview of various known techniques used for analyzing the distribution of data as well as identifying if possible groupings are suitable. Next, two pattern recognition algorithms are introduced, a generic approach and an application-tailored approach. Finally, the proposed techniques are applied using real-world data from a semiconductor manufacturer to show its functionality and potential.

2 SimData Toolbox for Supply Chain Simulation

2.1 Architecture of the SimData Toolbox

The pre-simulating procedure of examining and preparing the input data is an important, challenging, and time-consuming part of the whole simulation process [6, 7]. Providing tools for this purpose is a necessity and can reduce the associated effort. Fig. 1 shows the structure of the SimData Toolbox that is implemented by means of AnyLogic, while the different toolsets for visualization, statistical analysis, and pattern recognition are presented in the following sections. It is worth noticing that the functions of SimData are compatible with the objects of SCSC-SIMLIB.

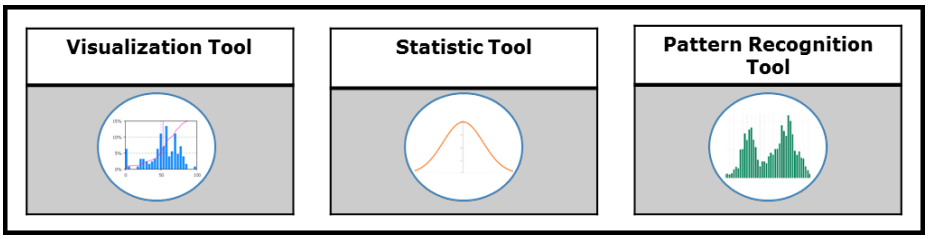

Fig. 1. Proposed SimData Toolbox

2.2 Visualization Toolset

AnyLogic already provides a variety of standard visualization tools for showing datasets, such as histograms and various charts. The Visualization Toolset provides additional objects for simplifying the plotting of histograms and graphs. As the purpose of this toolset is mainly to increase the user-friendliness of standard plotting features by means of self-programmed functions, we do not provide further details.

2.3 Statistic Analysis Toolset for Supply Chain Simulation Data

In many simulation studies, knowing the distribution of a dataset is crucial for a good result [7]. Often it is based on assumptions or on recommendations from the literature, but finding the right distribution by analyzing real datasets leads to more precise inputs. The Statistic Toolset offers approaches to analyze the distribution of datasets.

To get a most fitting outcome, the approach presented here is a combination of already existing statistical tests. These were selected among several possible approaches (all based on hypothesis testing) with regard to their applicability for the most important and most used distributions in supply chain, the power of the tests, and the amount of data points they are able to handle. The following tests have been selected: Kolmogorov-Smirnov test (K-S) [8], Anderson-Darling test (A-D) [8], and Shapiro-Wilks/Shapiro-Francia tests (S-W) [8].

The K-S algorithm was chosen for its applicability for all kind of distributions. It is based on the difference between the cumulative distribution function and the empirical distribution function. In general K-S is less powerful than the other approaches, but due to its applicability for almost every possible distribution, it is used in the overall algorithm to get a first impression of which distributions are best fitting. For a more powerful and thus reliable test on one of the most used distributions, the Normal distribution, the S-W algorithm is used. As a last algorithm, A-D can be applied for testing for Normal, Gamma, Exponential, Weibull, and Log-Normal distributions. This algorithm is in most cases one of the most powerful hypothesis goodness of fit distribution tests.

In the overall approach the algorithm goes iteratively through all implemented distributions for all tests (if applicable). Whereas K-S is only used for an overview of the fit, S-W and A-D give a final statement of whether or not the dataset is following one of the tested distributions (to a given significance level). Since the results of S-W and A-D are more reliable, their final statement is more important than the one of K-S. Nevertheless, there exist distributions which can only be handled by K-S. A modified version of the algorithms K-S and A-D is also available to check two or more datasets for having the same distribution, which can be useful during the verification and validation processes.

Furthermore, the Statistic Toolset contains an Analysis of Variance algorithm. This is a hypothesis test to determine if a grouping of an overall population makes sense. Therefore the algorithm examines whether or not the means of the different groupings are significantly different. The detailed procedure is described in [9].

2.4 Toolset for Pattern Recognition and Generation of Synthetic Data as Input for Supply Chain Simulation

In many simulation studies, generated input data are either represented by one (expected) value for all data points or by loading a representative list of values into the model. Both concepts have disadvantages: the first method results in low accuracy, as the values in the real scenarios usually show a non-zero variance. However, using the second method often does not allow the generation and modification of data on the fly according to the user's considerations. In addition the number of values may result in a high usage of internal memory and long loading times due to a large amount of data. A common approach to overcome these limitations is to create synthetic data that serve as an input for the simulation (Fig. 2). In this paper, synthetic data are referred to as data which are obtained by analyzing the real dataset and creating out of these observations a representative approximation of the data with high accuracy. It has the benefit of first being compact and avoiding

high amounts of data, second including all patterns of the real dataset, and third being able to adjust according to the user's needs regarding the number of data points that are generated. It remains the responsibility of the user to decide which input data shall be represented with real data from the database or from synthetic data. Synthetic data are suggested to be used in case of a huge amount of real data which can be represented well by synthetic data or in case the number of the real data points does not match to the needed number of points for the simulation. Below two approaches for generating synthetic data are presented, an extended Expectation-Maximization algorithm and an additional approach based on the examination of peaks and noise.

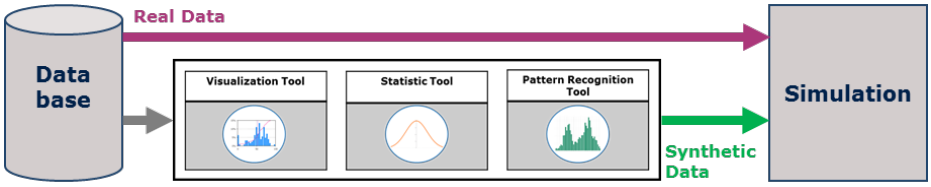

Fig. 2. Conceptual work flow of the input process for simulation

Expectation-Maximization Algorithm. For simple datasets (i.e. with one peak), the statistic tool introduced in section 2.3 is best for finding the right distribution and therefore creates the formula which is the basis for generating synthetic data. However, if as in most cases the dataset is not that simple and contains several peaks, another approach is necessary.

For the pattern recognition and the creation of a formula for generating synthetic data, an extended Expectation-Maximization algorithm (EM) is presented. The EM procedure can be found in the literature, see [10] among others. To achieve a tool, which can be handled more easily and requires few user assumptions without jeopardizing the accuracy, the implemented EM algorithm can be automated in some ways and can use additional clustering in advance (Fig. 3).

In the literature, it is often dealt with Gaussian mixtures as it is a convenient and easy to understand example. Real datasets however can often not be described as a Gaussian mixture, or even a mixture containing only distributions of one single type. To overcome this obstacle, the EM algorithm used in this toolset can handle all possible mixtures of the most important and most used distributions in supply chain simulation (e.g. Normal-, Gamma-, Exponential-, Weibull and Log-Normal-distribution). This variety makes the already complex initial guess of the distribution mixture, i.e. starting sub-distribution parameters and weights, even harder. However the extended algorithm used has the advantage to go iteratively through all possible distribution mixtures. It stops after examining all mixtures and suggests the one which fits the dataset the best. This comes unfortunately with a high computation time and it is therefore recommended only up to an amount of five sub-distributions.

As pointed by [11], the initialization step, i.e. the choice of the initial parameters, in the algorithm is crucial for the speed of the convergence. To automate this process a clustering algorithm is often suggested to get partitions of the data and therewith approximate the initial parameters based on these partitions. Often, the K-Means algorithm is proposed, and therefore it is also used in this extended approach. With

some of its known modifications K-Means is able to not only cluster the dataset, but also to estimate the number of clusters (i.e. sub-distributions). It decreases the user input to nearly nothing and, together with the iterative mixture check introduced before, automates and improves the whole combined algorithm significantly. The implemented tool is designed to use either the automated process (with or without clustering) or to take user-assumed input parameters and stopping levels.

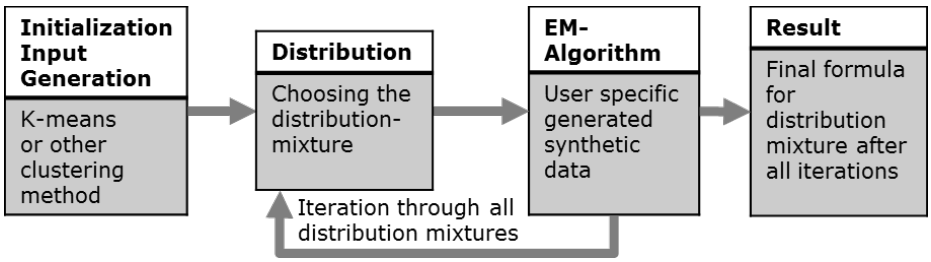

Fig. 3. Scheme of the extended Expectation-Maximization algorithm

Peak-and-Noise Detection Approach. While the EM algorithm is suitable for nearly all kinds of distributed data, there are a few datasets (see sub-section 2.5) where it does not give any good result. For one of these groups of datasets, an additional approach has been developed based on the detection of high peaks in a dataset and approximating the remaining noise.

The basic concept of this approach is to transfer the dataset into a histogram, which allows the algorithm to classify the information into peaks and noise to describe them separately and according to their characteristics. Using this information one can formulate a descriptor that gives a good and very compact approximation of the information in the dataset. In the initial state the complete dataset B is considered to be noise. A threshold value that separates the peaks from the noise is introduced. It depends on the number of elements in the noise dataset, i.e. *cardinality*, that is defined as follows:

$$\text{threshold(cardinality)} = c * \text{cardinality},$$

with c representing a percent value of the total number of elements in the dataset. By adjusting c the user can set a threshold from which a data point in a bin represents either a peak or noise. Once a peak is detected, the respective data points have to be removed, the threshold is adjusted to the new cardinality, and the algorithm starts a new iteration. This is repeated until no further peak is detected in the noise dataset. Due to the fact that a bin, which contains a peak, may also contain noise, it is convenient to replace the peak by the average of the neighboring bins.

Pseudocode of the Peak-and-Noise detection approach:

```
Create histogram
For each b∈B;
  If frequency>threshold;
    p↔b;
    replace data stored in b;
```

```
    update threshold;
    restart (at first b);
Normalize each b and p.
```

As p of P, with P denoting the set of detected peaks, usually contains very little information, it is convenient to write the respective information directly in the descriptor. B however contains many data points and thus it is suitable to give an approximation of the noise data.

2.5 Applying the Pattern Recognition Approaches to Generate Synthetic Data

E-M algorithm on Customer Forecast Accuracy Dataset. As an example the EM algorithm tool is used to evaluate a dataset based on customer forecast accuracy measurements obtained from a semiconductor manufacturer (i.e. 127 values, as shown in Fig. 5). Forecasts provided by customers are an important source of information for generating sales predictions and production plans [12]. However, customer forecasts are not always reliable as they are non-binding demand. The forecast accuracy has a significant impact on the supply chain performance and it is studied with simulation.

As an initial guess, four sub-distributions are predicted. K-Means finds the four clusters around 1, 29, 53 and 74. With this result, the initial sub-distribution parameters can be approximated and can serve as input for the extended EM.

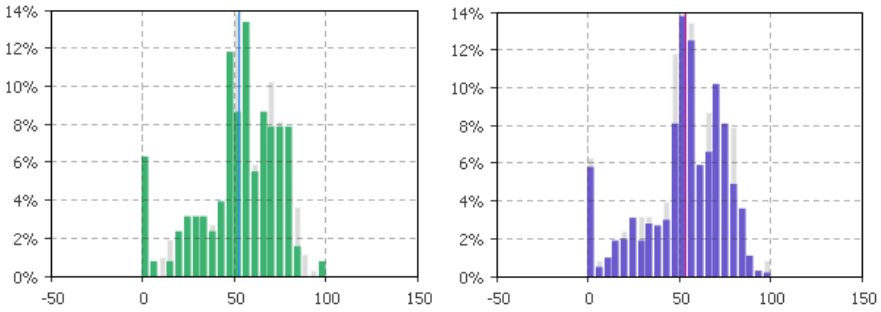

Fig. 5 and 6. Histograms of real data (left) and synthetic data (right) respectively

The derived formula of the distribution of the dataset according to EM is a weighted mixture of an Exponential-, a Normal-, a Weibull and a Log-Normal distribution with specific parameters. The descriptor of above example is as follows:

[0.07*E(1000)+0.15*N(26.24, 108.36)+0.43*W(54.70, 10.1)+0.35*LN(4.29, 0.01)],

with E(λ), an Exponential distribution, N(μ,σ), a Normal distribution with mean μ and standard deviation σ, W(λ,β), a Weibull distribution with scale λ and shape β, LN(μ,σ), a Log-Normal distribution with log-mean μ and log-standard deviation σ and 0.07, 0.15, 0.43, 0.35 being the respective weights. Generating 1000 values of synthetic data out of this formula results in the dataset shown in Fig. 6. By applying statistical tests for comparing distributions (see Sub-section 2.3), we confirm that both datasets have the same distribution at a significance level of 0.01.

Peak-and-Noise Approach on Lot-Size Dataset. Showing the need of the special Peak- and-Noise detection approach, it is used to create realistic lot-size input data for semiconductor supply chain simulation for the manufacturing level called assembly [13]. The initial dataset originates from a backend fabrication site of a semiconductor manufacturer. Figure 7 shows a transformation of the dataset into a histogram using a bin width of 10. The dataset shows a sub-structure with two sharp peaks.

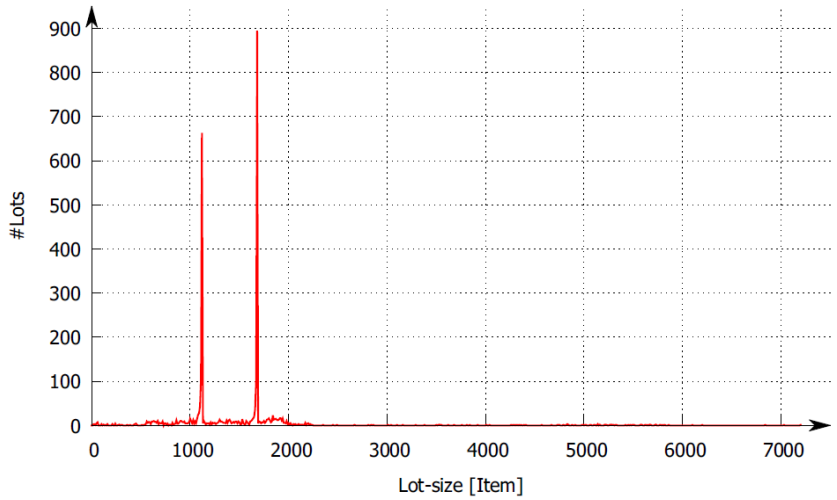

Fig. 7. Histogram of lot-size data

Applying the algorithm on this dataset results in 19.3% of the data at a lot-size equal to 1120 and 26.6% of the data at a lot-size equal to 1680. The noise was approximated by a four-parameter Weibull distribution [14]. The descriptor is as follows:

$$[W(11.6, 5460.3, -3890.6, 4.9); (1120, 0.193); (1680, 0.266); 10],$$

with the first entry being the description of the noise, a four parameter Weibull distribution, the next two entries being the description of the two peaks and the last entry being the bin width. With the help of above descriptor, synthetic data can be generated to accurately represent lot-size input data for simulating the backend fabrication site under consideration.

3 Conclusion and Future Research

In this paper, the SimData Toolbox was presented that is composed of three toolsets, which contain approaches for data analysis and for the generation of synthetic data. The proposed approaches involve various enhancements and combinations of already existing data analysis techniques, such as distribution goodness of fit methods and the Expectation-Maximization algorithm. The toolbox is an extension of a simulation library for semiconductor supply chains that is implemented with AnyLogic. Examples were provided to show the application of the proposed toolbox.

Further research on the pattern recognition approaches, especially for the extended EM algorithm, includes the testing of the algorithms on further real-world datasets and simulation applications as well as to improve their technical functionality, accuracy and computational speed. Furthermore, we intend to improve the initial clustering algorithms with more accurate and faster schemes. As a next step, more commonly known statistical techniques should be applied to problems in semiconductor supply chain simulation to complete the already existing toolsets.

Acknowledgements. Special thanks go to Esther Wissel for her work on pattern recognition, Michael Wiedemann for implementing the Peak-and-Noise detection algorithm, Peter Benilov for implementing the graph plotting functions, and Sebastian Eirich and Hans Ehm for their constructive inputs.

References

1. Chien, C.-F., Dauzère-Pérès, S., Ehm, H., Fowler, J., Jiang, Z., Krishnaswamy, S., Lee, T., Mönch, L., Uzsoy, R. (2011) 'Modelling and Analysis of Semiconductor Manufacturing in a Shrinking World: Challenges and Successes', *European Journal of Industrial Engineering*, Vol. 5, No. 3, pp.254–271.
2. Ehm, H., Ponsignon, T., Kaufmann, T. (2011) 'The Global Supply Chain is our New Fab: Integration and Automation Challenges', *Proceedings of the 22nd Annual IEEE/SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, pp.1-6.
3. Mönch, L, Fowler, J., Mason, S. (2013) *Production Planning and Control for Semiconductor Wafer Fabrication Facilities: Modeling, Analysis, and Systems*, Springer Science + Business Media, New York.
4. Yuan, J. and Ponsignon, T. (2014) 'Towards a Semiconductor Supply Chain Simulation Library (SCSC-SIMLIB)', *Proceedings of the Winter Simulation Conference*, pp.2522-2532.
5. AnyLogic (2015) 'Use of Simulation', *http://www.anylogic.com/use-of-simulation*.
6. Barlas, P. (2014) 'Towards Automated Simulation Input Data: An Open Source Tool to Enhance the Input Data Phase in Discrete Event Simulation', *Proceedings of the Winter Simulation Conference*, pp.4007-4008 .
7. Law, A.M. (2007) *Simulation Modeling and Analysis*, 4th ed., McGraw-Hill, New York.
8. Jantschi, L. and Bolboaca, S.D. (2009) 'Distribution Fitting 2. Pearson-Fisher, Kolmogorov-Smirnov, Anderson-Darling, Wilks-Shapiro, Cramer-von-Misses and Jarque-Bera Statistics', *Horticulture*, Vol. 66, No. 2, pp.691-697.
9. Terrell, G.R. (1999) *Mathematical Statistics*, Springer, New York.
10. Dempster, A.P., Laird, N.M., Rubin, D.B. (1977) 'Maximum Likelihood from Incomplete Data via the EM Algorithm', *Journal of the Royal Statistical Society*, Series B (Methodological), Vol. 39, No. 1, pp.1-38.
11. Ng, S.K., Krishnen, T., McLachlan, G.J. (2012) 'The EM Algorithm', *Handbook of Computational Statistics: Concepts and Methods*, Springer, New York.
12. Roundy, R.O. (2001) 'Report on Practices Related to Demand Forecasting for Semiconductor Products', *Survey of Semiconductor Research Council (SRC) Member Companies*, pp.1-20.
13. May, G. and Spanos, C. (2006) *Fundamentals of Semiconductor Manufacturing and Process Control*, John Wiley & Sons, Inc., Hoboken.
14. Rinne, H. (2008) *The Weibull Distribution: A Handbook*, CRC Press.