Development of an Integrated Solution for Data Farming and Knowledge Discovery in Simulation Data

Jonas Genath[1*], Sören Bergmann[1], Niclas Feldkamp[1],
Sven Spieckermann[2], Stephan Stauber[2]

[1]Group for Information Technology in Production and Logistics, Ilmenau University of Technology,
Ehrenbergstraße 29, 98693 ilmenau, Germany; **jonas.genath@tu-ilmenau.de*

[2]SimPlan AG, Sophie-Scholl-Platz 6, 63452 Hanau, Germany

SNE 32(3), 2022, 121-126, DOI: 10.11128/sne.32.tn.10611
Selected ASIM SPL 2021 Postconf. Publication: 2022-01-18;
Received English version: 2022-03-14; Accepted: 2022-05-09
SNE - Simulation Notes Europe, ARGESIM Publisher Vienna
ISSN Print 2305-9974, Online 2306-0271, www.sne-journal.org

Abstract. Simulation is an established methodology for planning and evaluating manufacturing and logistics systems. In contrast to classical simulation studies, the method of knowledge discovery in simulation data uses a simulation model as a data generator (data farming). Subsequently, hidden, previously unknown and potentially useful cause-effect relationships can be uncovered on the generated data using data mining and visual analytics methods. So far, however, there was a lack of integrated, easy-to-use software solutions for the application of the data farming in operational practice. This paper presents such an integrated solution, which allows generating experiment designs, implements a method to distribute the necessary experiment runs, and provides the user with tools to analyze and visualize the result data.

Introduction

Simulation is an established tool for planning and controlling complex production and logistics systems and has proven to be an important key component, among other things, in solving challenges in the context of Industry 4.0 [8]. Traditional simulation studies are usually designed to cover a previously defined project scope or to achieve a concrete project goal through manual experimentation. This includes, for example, the optimisation of a production layout [9].

With increasing computing power and the general availability of Big Data infrastructures and cloud-based solutions, as well as considerable progress in the field of data mining, another possible application for simulation models arises: conducting a very wide range of experiments to uncover hidden, previously unknown and potentially useful cause-effect relationships. Particularly in complex systems, there may be relationships, problems or even solutions that go beyond the defined goal of a traditional simulation project and can therefore contribute to decision support. The basis for this approach is the methodology of data farming [5].

Based on data farming, Feldkamp et al. [4] developed a method named Knowledge Discovery in Simulation Data, which supplements data farming with methods from data mining and visual analytics, specifically suited for the analysis of production and logistic systems. Initial case studies have proven its potential [1, 2].

However, a broad transfer into operational practice was so far held back due to the lack of an integrated software solution that also enables non-simulation or data farming experts to conduct knowledge discovery in simulation projects.

This paper presents such an integrated solution, which initially extends the existing software solution SimAssist (cf. [13]) as a prototype. The development was carried out within the framework of the German Federal Ministry of Education and Research (FMER) project "Development of an integrated solution for data farming and knowledge discovery in simulation data (DaWiS)". The sub-aspects to be considered here are procedures of intelligent experiment design, methods for the (cloud-based) distribution of experiments as well as the selection and adaptation of suitable data mining and visual analytics methods, so that data farming or the method of knowledge discovery in simulation models according to Feldkamp [1] can be effectively applied with little training effort.

In this paper, the methods and the implemented software solution are explained using an example from the automotive industry. The actual simulation is carried out in the simulation software Siemens Plant Simulation.

The remainder of the paper is structured as follows. First, the state of research and the necessary theoretical foundations of data farming and the method for knowledge discovery in simulation data (KDS) are presented briefly. Then, in Section 2 the main part of the paper, the integrated method is presented step by step, and illustrated by a workflow example. At selected points, particular attention is paid to the technical implementation. The article ends with a conclusion and an outlook on possible extensions of the integrated solution.

1 State of the Art

In data farming, a previously validated simulation model is used as a data generator to cover the largest possible spectrum of model or system behaviour (response surface) with the help of intelligent experiment design and high-performance computing [5, 10]. The "farming" metaphor expresses that the goal is to maximise the data yield of the simulation model, analogous to a farmer who cultivates his land as efficiently as possible to maximise his crop yield [11].

The research and development of improved procedures for the design of simulation experiment plans is one of the crucial prerequisites. These allow possible combinations of factor values to be comprehensively represented and at the same time guarantee a reasonable number of experiment runs to generate data [7, 12]. Especially in the context of the simulation of production and logistics systems, the selection of one of the design methods or even the selection of a suitable combination of different design methods is of great importance. To carry out the experiments, the data farming literature often refers to appropriate high-performance computing [5].

Interesting relationships can then be uncovered in the generated data with the help of various data mining or visual analytics methods [6]. This way, previously unknown relationships, problems or even solutions can possibly be identified.

Feldkamp [1] presents a selection of possible data mining methods, e.g., clustering, and the appriopate workflow for applying those methods contiguously. It is recommended that the actual analysis of the generated simulation result data and the relationships between factors and result data (key figures) is ideally supported with interactive, visual analysis. Visualisation is generally a crucial tool when an interpretation of data is required. A consistent dovetailing, as is generally recommended in the research discipline of visual analytics, between interactive visualisation, e.g., by means of interactively adaptable animations, time series diagrams, graphs, and data analysis by means of data mining methods, enables the user to incorporate the human ability to draw conclusions in the best possible way [3, 6].

In summary, the state of the art in science and technology in this context shows that the basic individual methods (data farming, intelligent experiment design, data mining and visual analytics) have reached a sufficient maturity level. Prototypical applications in the context of simulating production and logistics systems demonstrating the potential of this approach have also been published. However, it must be stated that there is yet no holistic solution for transferring the methods as a whole or at least in significant parts into a framework that can be operated by non-experts, and which also focuses on the area of simulation in production and logistics. Furthermore, there is a lack of methods for (partial) automation of the processes and for supporting non-experts in general.

2 Integrated Solution for Data Farming and Knowledge Discovery in Simulation Data

As already mentioned in the introduction, the aim of the FMER research project DaWiS is to develop a software solution, supplemented by best practice procedures, which also allows non-experts to acquire knowledge based on data farming using data mining and visual analytics methods.

For this purpose, the proven modular software SimAssist (cf. [13]) by Simplan AG, which already provides extensive assistance functions for the administration, analysis, visualisation, and documentation of result data of classic simulation projects, has been expanded.

The extensions are combined in a new module of the software call “4farm”. Corresponding components were designed and prototypically implemented for the already mentioned sub-aspects, the intelligent experiment design, the distribution of experiments as well as for the data mining and visual analytics. The basic architecture of the module can be seen in Figure 1.

It is worth mentioning that in addition to research on the methodology, substantial effort was put into the design of the user interactions during the conception and development, so that all necessary sub-aspects (from experiment design to the distribution of experiment runs to the analysis and visualisation of data) are available via an integrated interface without changing the software.

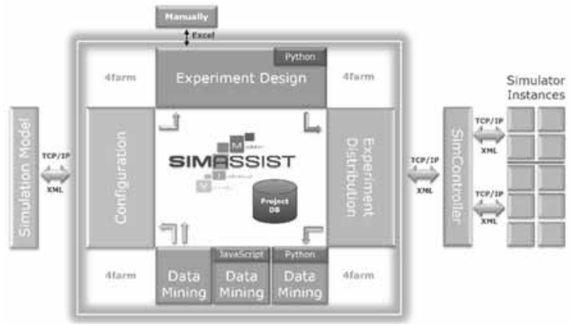

Figure 1: General architecture of the software solution (SimAssist – 4farm module) for data farming and knowledge discovery in simulation data.

Furthermore, as many technical details as possible are hidden from the end user, especially regarding the experiment design, the data mining and visual analytics methods, necessary settings are anticipated based on best practices. If this is not possible, the necessary settings are transferred to an intuitively understandable technical application level. This is done by asking for the necessary parameters when using the methods. In each case, the user is offered lists with selection options.

The corresponding notes on the use as well as the advantages and disadvantages of the individual variants are stored in the software in the form of information texts or decision trees or in a similar way. The user can thus focus completely on the simulation model and objective of the simulation study.

Due to the rapid development of research in the field of data mining and visual analytics, but also due to the large number of possibilities regarding experiment design, another requirement placed on the software is that new methods, including visualisations, can easily be added in the future via a standardised mechanism.

2.1 Workflow Example – Supplying a Car Production Line with Batteries

In the context of this paper, selected methods as well as the implemented software solution - especially the excerpts of the user interfaces - are presented based on an example scenario. Among other use cases, this example was used in the DaWiS project starting with the requirements analysis until the final demonstration of the methods and the software.

The workflow example includes a typical logistical problem in which the supply of a running car production with two different types of batteries as well as the disposal of the stackable empty load carriers is considered. The delivery of the batteries in load carriers and the collection of the empty load carriers is done by truck at an unloading dock. The actual handling of the batteries is done by forklift trucks (Figure 2).

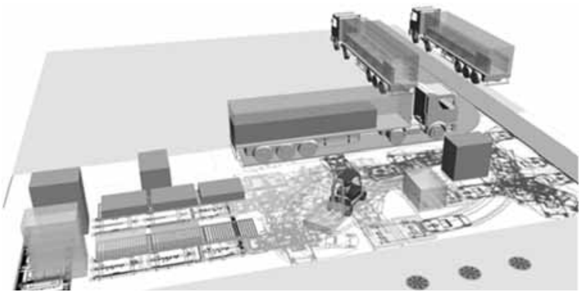

Figure 2: Screenshot of the model of a loading/unloading of batteries to supply a running production.

The variable input parameters (factors) of the model include the ratio of battery types A and B in the total production programme of the vehicle assembly (which itself is not part of the simulation model under consideration). Further factors of the simulation model are the total production volume, the cycles in which trucks deliver new batteries and the size of the buffers for full loads. In addition, three different scenarios are defined, each of which differs in the forklift variants used (5t or 8t forklifts) and the number of forklifts (1-2 forklifts). To analyse the results, 31 result parameters are stored, including the forklift utilisation or the downtimes of the connected assembly due to missing batteries.

The model of the production and logistics system was created using the Siemens Plant Simulation software. Here, modules were developed which enable the setting of the factor values (variables to manipulate) and the readout of the pre-definable result data (XML format).

With the help of these modules, models can be enriched with meta-information about the factors and result parameters as well as their data types and, if applicable, existing value ranges. This information can be evaluated by the integrated solution and presented to the user and used as a basis for the concrete experiment design.

2.2 Experiment Design

As described earlier, generating suitable experiment designs is the first major challenge in the process of data farming and thus also in the method for knowledge discovery in simulation data that builds on it. Currently, five different experiment design methods have been selected and implemented: the full factorial design, the 2^k-design, the central composite design, and the Latin Hypercube Sampling (LHS) as well as a design in which an LHS can be crossed with another factor or design.

It should be noted that experiment design methods sometimes require method-specific parameters in addition to the parameters describing the factors, i.e., the names of the factors, the data type and value range of the factor. For example, in case of the LHS, the number of experiments must be specified. All design methods are implemented as Python scripts. The scripts use a uniform library for XML-based data import and export and can contain corresponding meta-information as comments at the beginning.

The information about the factors can be set manually or – as indicated in the previous section – read out from the simulation model. The design method-specific parameters are queried from the user in SimAssist. Which parameters are queried and how the interface is designed in the SimAssist 4farm module is defined in the meta information of the respective script.

Adding further experiment design methods is possible at any time without restarting the software. To do this, the methods only must be made known as Python scripts annotated with the addressed meta information by copying them into a defined directory (so-called hot deployment). The corresponding selection option and the interfaces are generated ad hoc and can be used immediately.

When selecting the experiment design methods, the user is supported on the one hand by textual help for the individual methods. On the other hand, assistance is available in the form of a decision tree (Figure 3), in which a design method is suggested by answering simple questions.

In the example scenario mentioned here, a crossed LHS design with 15,000 experiments (5000 LHS * 3 scenarios) was used due to the different scenarios (green/dashed path in Figure 3).

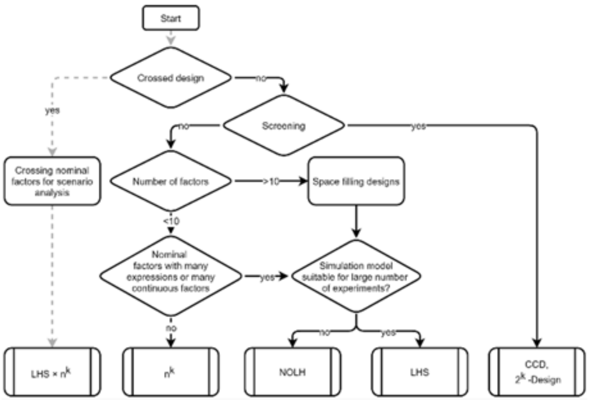
Start
Crossed design
no
yes
Screening
no
yes
Crossing nominal factors for scenario analysis
Number of factors
>10
Space filling designs
<10
Nominal factors with many expressions or many continuous factors
yes
Simulation model suitable for large number of experiments?
no
no
yes
LHS × n^k
n^k
NOLH
LHS
CCD, 2^k -Design

Figure 3: Flowchart for selecting the appropriate experiment design methods [1].

2.3 Experiment Distribution

Due to the large number of simulation runs, it is usually not practical to carry out the experiments on a single computing instance (single computer or single processor core on a single computer). Instead, it is desirable to distribute individual experiments across different computing instances. The technology used in the DaWiS project is based on a software component from a preliminary project of Simplan AG, the so-called SimController. This was adapted in such a way that it is now possible to distribute experiments, i.e., models and the concrete factor values in the form of XML data via a central instance to self-registering client instances using TCP/IP. The individual instances continuously report their status.

After running an experiment, the defined characteristic values (and used factor values) are reported back to the central instance and stored in SimAssist in the form of a SQLite database. This data can then be used and analysed very easily.

2.4 Analysis of the Result Data – Data Mining and Visual Analytics

Analogous to the experiment design methods, the number of possible data mining methods and visualisation methods is very large. In the method for knowledge discovery in simulation data, Feldkamp [1] analyses different groups of methods and provides an assessment them with regards to their benefit for knowledge discovery.

Based on this research, methods of descriptive statistics, correlation analysis, clustering (k-means and gaussian mixture) as well as regression analysis and the formation of classification trees were classified as most valuable for practical application and implemented in the prototype. In addition, there are suitable visualisations such as heat maps for correlation analysis or parallel coordinate and scatter diagrams for the evaluation of clustered data. The technical implementation here is analogous to the implementation of the experiment design methods, i.e., each of the methods is implemented as an annotated Python script and the data exchange with the script is again carried out via XML format. Here, too, it is thus easy to implement further data mining methods and visualisations, which are immediately available to the user via generic dialogues from the 4farm module for knowledge discovery.

The data analyses within the workflow example have not yet been completed. However, initial findings are emerging. For example, it turns out that assuming the current demand for batteries, every scenario leads to a secure supply of production. However, even with a moderately increasing proportion of battery electric vehicles, scenarios with two forklifts, at least one of which is an 8t forklift, work better, especially if one battery type is in demand significantly more often.

3 Conclusion

The paper presented an integrated software solution developed for knowledge discovery in simulation data. For this purpose, the need for and the requirements of such a solution were derived and the essential sub-aspects of the method and its user-friendly prototypical implementation were examined in more detail. Further development steps include the implementation of additional experiment design and data mining methods as well as additional visualisations. Moreover, further tests in real-world use cases are necessary, especially to validate the implemented interfaces and file formats. Finally, research on further (partial) automation of the data mining methods, e.g., by means of meta-learning to determine suitable hyperparameters, or the use of methods for robustness analysis, is conceivable.

References

[1] Feldkamp N. *Wissensentdeckung im Kontext der Produktionssimulation*. 2020. Universitätsverlag Ilmenau, Ilmenau.

[2] Feldkamp N, Bergmann, Strassburger S. Knowledge Discovery in Simulation Data. 2020. *ACM Trans. Model. Comput. Simul.* 30, 4, 1–25.

[3] Feldkamp N, Bergmann S, Strassburger S. Visualization and Interaction for Knowledge Discovery in Simulation Data. 2020. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*. Proceedings of the Annual Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences, 1340–1349. DOI 10.24251/HICSS.2020.165

[4] Feldkamp N, Bergmann S, Straßburger S, Schulze T. Data Farming im Kontext von Produktion und Logistik. 2017. In *Simulation in Produktion und Logistik 2017*. kassel university press, Kassel, 169–178.

[5] Horne GE, Meyer T. Data farming and defense applications. 2010. In *MODSIM World Conference and Expo*. Langley Research Center, Hampton, VA, 74–82.

[6] Keim DA, Mansmann F, Schneidewind J, Thomas J, Ziegler H. Visual Analytics: Scope and Challenges. 2008. In *Visual Data Mining*, Simoff SJ, Böhlen MH, Mazeika A, Eds. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 76–90 DOI 10.1007/978-3-540-71080-6_6

[7] Kleijnen JP, Sanchez SM, Lucas TW, Cioppa T M. State-of-the-Art Review: A User's Guide to the Brave New World of Designing Simulation Experiments. 2005. *INFORMS Journal on Computing* 17, 3, 263–289.

[8] Krückhans B, Meier H. Industrie 4.0 – Handlungsfelder der Digitalen Fabrik zur Optimierung der Ressourceneffizienz in der Produktion. 2013. In *Proceeding der 15. ASIM Fachtagung Simulation in Produktion und Logistik 2013*. *Entscheidungsunterstützung von der Planung bis zur Steuerung* 147. HNI-Verlagsschriftenreihe, Paderborn, 31–40.

[9] Law AM. How to conduct a successful simulation study. 2003. In *Proceedings of the 2003 Winter Simulation Conference*. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, 66–70. DOI 10.1109/WSC.2003.1261409

[10] Sanchez SM. Work Smarter, Not Harder: Guidelines for Designing Simulation Experiments. 2007. In *Proceedings of the 2007 Winter Simulation Conference. December 9 - 12, 2007, Washington, DC, U.S.A.* IEEE, Piscataway, N.J., 84–94. DOI 10.1109/WSC.2007.4419591

[11] Sanchez SM. Simulation Experiments: Better Data, Not Just Big Data. 2014. In *Proceedings of the 2014 Winter Simulation Conference*. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, 805–816.

[12] Sanchez SM, Wan H. Better than a petaflop: The power of efficient experimental design. 2009. In *Proceedings of the 2009 Winter Simulation Conference (WSC 2009)*. IEEE Inc, Piscataway, N.J., 60–74. DOI 10.1109/WSC.2009.5429316

[13] Sokoll K, Clausing M. Methoden und Werkzeuge der Simulationsassistenz. 2020. In *Ablaufsimulation in der Automobilindustrie*, Mayer G, Pöge C, Spieckermann S, Wenzel S, Eds. Springer Vieweg, Berlin, 349–364. DOI 10.1007/978-3-662-59388-2_24.

Acknowledgement

The DaWiS research project was funded by the German Federal Ministry of Education and Research. The Funding started in January 2020 and lasted for 2 years.

Federal Ministry
of Education
and Research