Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

A METHODOLOGY FOR INPUT DATA MANAGEMENT IN DISCRETE EVENT SIMULATION PROJECTS

Anders Skoogh
Björn Johansson

Department of Product and Production Development
Chalmers University of Technology
Hörsalsvägen 7A
Gothenburg, SE-412 96, SWEDEN

ABSTRACT

Discrete event simulation (DES) projects rely heavily on high input data quality. Therefore, the input data management process is very important and, thus, consumes an extensive amount of time. To secure quality and increase rapidity in DES projects, there are well structured methodologies to follow, but a detailed guideline for how to perform the crucial process of handling input data, is missing. This paper presents such a structured methodology, including description of 13 activities and their internal connections. Having this kind of methodology available, our hypothesis is that the structured way to work increases rapidity for input data management and, consequently, also for entire DES projects. The improvement is expected to be larger in companies with low or medium experience in DES.

1 INTRODUCTION

Discrete event simulation (DES) has proved itself to be a very powerful tool for decision support in production development (Williams 1996). The tool provides possibilities to conduct precise dynamic analyses in order to improve running production or to secure smooth implementations of new products or production equipment. However, despite the promising potential, industry has not fully adopted the tool (Ericsson 2005).

One of the most conspicuous disadvantages of DES is arguably the extensive amount of time needed to perform a simulation study (Johansson, Johnsson, and Kinnander 2003). The substantial time-consumption has been especially conspicuous when applying DES in early conceptual phases of major change projects, e.g. new product introduction or implementation of new production equipment. In this kind of projects, quick responses on analyses are usually essential in order to reduce project lead-time.

Moreover, there is a broad consensus on the fact that input data management is one of the crucial parts of a simulation project, with regard to the time-consumption. Previous studies have shown that the input data phase constitutes on average 31% of the time in entire projects (Skoogh and Johansson 2007), and Trybula (1994) reported similar results, stating that the input data phase consumes 10-40%.

Fortunately, a considerable amount of research work has been performed to reduce the time-consumption for input data management. A lot of work has focused on automating the process of input data collection. For instance, Randell and Bolmsjö (2001) demonstrated a method to reduce project lead-time using database driven factory simulation. Robertson and Perera (2002) described how the Corporate Business System can be used as the simulation data source and, thus, be advantageous in order to increase speed of input data collection. On the other hand, complete automation of the entire input data management process requires well developed original data sources. However, in many cases these sources omit some necessary data for simulation, especially data needed to mimic the dynamics of the investigated system (Robertson and Perera 2002; Ho, Wu, and Tai 2004). Furthermore, small and medium sized companies do not always have continuous collection of production data.

Hence, there is also a need for structured methodologies to support more traditional working procedures throughout the input data phase. However, there are few practical guides or previous contributions using a more systematic approach (Perera and Liyanage 2000, Lehtonen and Seppälä 1997), which is a pity since the number of non-specialists working with DES tools is increasing (Hollocks 2001). Moreover, successful examples of this kind of methodologies are found by just studying the numerous guidelines of structuring entire simulation projects (Banks et al. 2004; Law 2007; Pidd 1995). These have proved to be of great support for practitioners, not least for users with low or medium experience of DES.

The aim of this paper is to contribute to the work towards more time-efficient and accurate input data management for simulation projects, by proposing a structured methodology for activities in the input data phase, e.g.

978-1-4244-2708-6/08/$25.00 ©2008 IEEE

identification, collection, analysis, and storage. A survey was performed among 15 previously completed simulation projects, in order to identify the design requirements for the methodology. Results from the survey also indicate a significant potential of utilizing an easy guide that outlines the most important steps in the input data phase of DES projects.

2 INPUT DATA MANAGEMENT

In this paper, input data management is defined as the entire process of preparing quality assured, and simulation adapted, representations of all relevant input data parameters for simulation models. This includes identifying relevant input parameters, collecting all information required to represent the parameters as appropriate simulation input, converting raw data to a quality assured representation, and documenting data for future reference and re-use. Our focus is on management of data required for model realization. However, many of the activities and descriptions are also relevant for contextual data and data needed for model validation (Pidd 2003). Moreover, the approach adopted in this paper is primarily intended for quantitative data and, thus, logical relations between model entities are presupposed to be handled in the conceptual model.

In addition to the categorization of data as contextual, required for model realization and needed for model validation (Pidd 2003), Robinson and Bhatia (1995) divide data into three other categories based on availability and collectability (Table 1). This classification is very useful to refer to when considering input data methodologies, since the three categories require significantly different approaches during collection. Firstly, category A data is already available, for instance in automated logging systems, Corporate Business Systems or just previously measured data intended for another study. Of course, this type of data is very convenient, since further work is limited to data analysis and validation. Secondly, category B data requires additional effort because it needs to be gathered during the simulation study. Finally, category C data is neither previously available nor collectable, often due to new processes or equipments in the investigated system. Estimation of category C data requires both a well designed strategy and scrupulous care, in order to maintain model quality.

Table 1: Classification of data (Robinson and Bhatia 1995)

Category A	Available
Category B	Not available but collectable
Category C	Not available and not collectable

Existing literature on input data management focuses mainly on how to represent extensive sets of raw data in simulation models (Robinson 2004, Perera and Liyanage 2000). Hence, there is a lot of information available on how to select a proper statistical or empirical distribution. Consequently, guidelines and information about various distribution families, Maximum Likelihood Estimations (MLE) and goodness-of-fit tests are well described, for instance in Leemis (2004) and Law (2007). However, efforts to cover a wider range of issues in the input data management process, using a systematic approach, appear less frequently during a literature review (Perera and Liyanage 2000, Lehtonen and Seppälä 1997, Hatami 1990).

One of the contributions using a systematic approach is a methodology based on the Integrated computer aided manufacturing DEFinition (IDEF) (Perera and Liyanage 2000). The methodology focuses mainly on reducing required time for identification of parameters to include in the simulation model, which is arguably one of the most time-consuming activities in input data management. After investigating the system of interest for simulation, a functional model is built using pre-developed IDEF constructs. Thereafter, a required entity model is generated, which can be translated into a relational database, providing the model-builder with a structure to follow during data collection and for data storage.

Furthermore, controllability analysis (CA) has been used to increase efficiency in problem definition and data management phases of simulation projects (Lehtonen and Seppälä 1997). CA is an iterative approach intended to focus only on relevant aspects of the problem to solve. At each aggregation level, the aspect of major relevance is focused upon and further analyzed in order to pinpoint the most important factors with regard to project objectives. This structured methodology is very sound in order to identify important parameters, and facilitates the data management process by minimizing collection of data that is actually irrelevant for solving the problem. However, the methodology does not describe more detailed input data management activities like collection, preparation of raw data or data validation.

3 PROJECT INTERVIEWS

The suggested methodology is based on 15 semi-structured interviews (Denscombe 1998) where simulation practitioners contributed with their experiences from DES projects performed between 1999 and 2006. The selected projects represent a wide range of companies with regard to size, line of business, and previous DES experience.

During the interviews, the working procedures applied in each of the 15 projects were closely examined. Additionally, several issues from the projects' input data processes were addressed. The respondents shared their reflections on problems they faced related to input data management. The interviews also covered the respondents' own suggestions on important steps to make input data management more efficient in future DES projects. Specific results from the interviews are presented in Table 2 and Table 3. The 15 project members were also requested to estimate the value of a predefined methodology for input data management. On the specific question "do you think that the input data management phase in your project would have been more rapid if a structured methodology was applied?", the average response was 5.73 on a scale from 1 to 7. 1 means that the respondent totally disagrees and 7 means that he or she totally agrees. The specified reasons, indicating that there is promising potential in a more structured way of working with input data, are presented in Table 2. The reasons are arranged in decreasing order, starting with the most frequent. No explanations for disagreements were given (only one answer was below 4).

Table 2: The respondents' major explanations of why a structured methodology is assumed to increase rapidity and quality of input data management.

Major Expected Benefits
Increased awareness and focus on identifying the correct parameters, before starting the data gathering
A generally increased definition of work structure
Deciding the number of samples before starting the data gathering
Inconspicuous but important activities, such as a separate validation of input data, are highlighted
Increased focus on identifying the correct data sources and making sure that all data will be found

Additionally, the interviews brought up several interesting points about input data related problems, which likely could have been avoided using a structured methodology. These problems are summarized in Table 3.

Table 3: Problems experienced due to lack of structured methodology.

Experienced Problem	Root Cause
Made too many measurements with regard to model detail.	"Measured everything from the beginning, without specifying required accuracy".
Late additional rounds of data gathering	No rigid analysis, verifying that all data would be found.
Several attempts of raw data gathering failed.	The gathering methods were not properly chosen and clearly defined in advance.
Many iterations in data collection.	Inefficient validation process. No separate data validation.

Accordingly, the proposed methodology was developed as a combination of the 15 closely examined working procedures, the respondents' further suggestions, and the authors' experiences from more than 50 DES projects world wide.

4 PROPOSED METHODOLOGY

The proposed methodology follows distinct activities which are shown in Figure 1. The proposed methodology for input data management does fit well into the previously frequently cited works of Banks et al. (2004), Law (2007), Pegden, Shannon and Sadowski (1995), which all show methodologies on how to perform a DES project. In these methodologies, the input data management part represents a smaller portion of a full project. That smaller portion is described in more detail below. Figure 1 shows a scheme of the proposed input data management methodology.

4.1 Identify and Define Relevant Parameters

The first step while preparing input data for simulation models is to identify which parameters are necessary to include in the model. This might appear a simple task, but due to problems like high system complexity and selecting an appropriate level of detail according to the problem definition and objectives, one should not underestimate the required effort (Perera and Liyanage 2000). It is of great importance to closely investigate the system, for example by practice or pre-observation sessions, and detailed interviews with process experts. Preferably, the identification of data is performed in close connection to the development of a conceptual model (Robinson and Bhatia 1995).

Moreover, to support the identification process, there are models and methodologies, which help to select an appropriate level of detail (Lehtonen and Seppälä 1997) and to decide which parameters are usually needed to model specific entities or processes. Core Manufacturing Simulation Data (CMSD) is one such effort, driven by the Simulation Interoperability Standards Organization (SISO) (Lee et al. 2007). The previously introduced IDEF-based methodology, developed by Perera and Liyanage (2000), also include functionality to connect specific parameters to entities in the conceptual model.

Finally, the activity does not just include identification of relevant parameters. All parameters also need to be defined with regard to how they shall be measured and represented in the model. For instance, in many cases it is not obvious how to define a machine's cycle time. Does it start when the product is taken from the material handling device into the production cell, or should the measurement start when the machine actually starts processing the part? According to our interviews, lack of parameter definition has caused confusion during input data management in several of the 15 studied projects. To avoid the problem, system experts should be involved to explain how the company usually defines and measures different parameters.

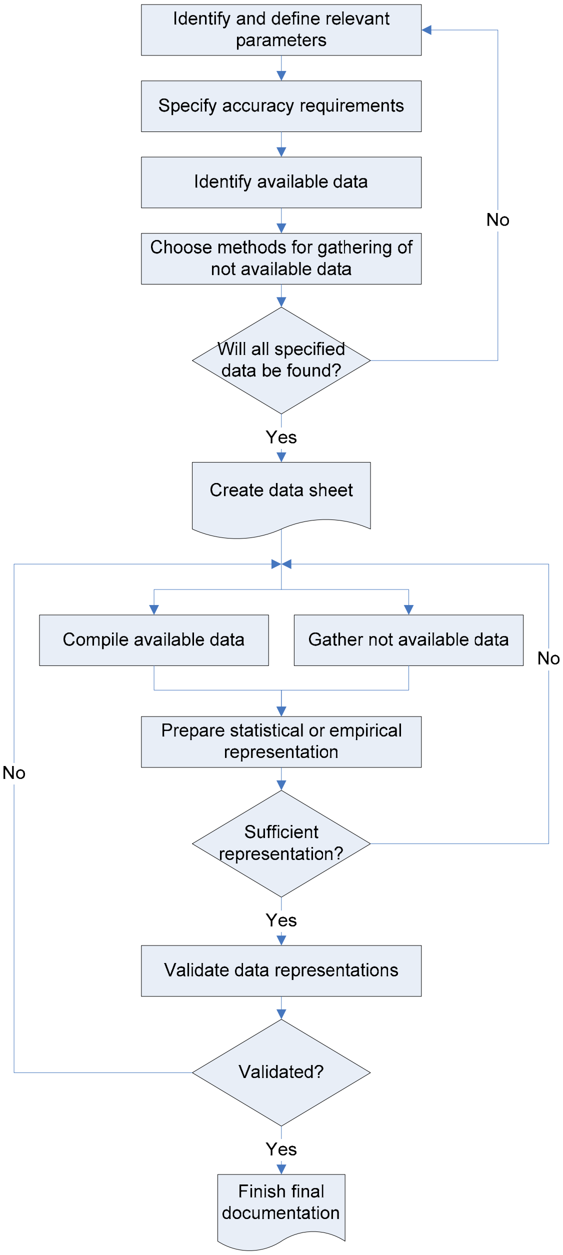
Identify and define relevant parameters
Specify accuracy requirements
Identify available data
Choose methods for gathering of not available data
Will all specified data be found?
No
Yes
Create data sheet
Compile available data
Gather not available data
No
Prepare statistical or empirical representation
No
Sufficient representation?
Yes
Validate data representations
Validated?
Yes
Finish final documentation

Figure 1: Proposed methodology for increased precision and rapidity in input data management.

4.2 Specify Accuracy Requirements

For quality reasons it is usually good to collect as much raw data as possible in order to generate good representations for simulation parameters. To be efficient, however, it is worthwhile to differentiate between the demand of accuracy for each parameter. System entities which do not significantly affect model performance can be paid less attention than critical ones. Thus, possible bottlenecks and narrow sectors in the system require high accuracy data with high validity.

During input data management, before the model is developed, system knowledge is, again, the only possible source to determine how important certain data is for model performance. For example, if a sequenced production chain has one resource which has significantly longer processing time compared to other resources upstream and downstream, it is of more importance to have accurate data on this resource, since it will likely control the output frequency. However, later during the project it is also possible to create an experiment to analyze input data sensitivity, which will verify (or disprove) that the collected data is sufficient for the model validity. This is a very powerful complement to early decisions on accuracy requirements. It is recommended to do sensitivity analyses for all borderline cases.

Another factor that affects the efforts and number of samples required for a specific parameter is process variability, which is also possible to predict using process knowledge. Process attributes expected to show a constant behavior, e.g. conveyor speed and cycle times for automated stations, need only enough samples to ensure that no unexpected variability is present. On the other hand, factors with high variability require more samples to succeed with a good representation. For instance, one of the most common input data types is breakdown data. Data describing Time To Repair (TTR) and Time To Failure (TTF) is often highly variable and hard to find on short notice, if no historical data has been collected over a long time period. Perrica et al. (2008) present one example on accuracy and estimation for both TTF and TTR. They also recommend that at least 230 samples should be collected to estimate the probability functions of TTR and TBF. This number of samples is a good rule of thumb for collection of all types of variable parameters, if possible.

4.3 Identify Available Data

In order to save time during input data management, it is important to take advantage of all previously collected data, the available category A data. Nowadays, continuous data collection is increasing significantly in industry, which is good for simulation projects. Unfortunately, simulation aspects have usually been ignored during the specification process of the collection systems and databases. Hence, companies incorrectly believe and claim that they have all data necessary for simulation, but on closer investigation they do not. Therefore, it is of crucial importance to go through all sources of available data to make sure that required data is possible to extract and that it is measured in a suitable way. This problem was experienced in several of the projects included in this study. To learn and to understand the structure of all sources is many times a very time-consuming task (Skoogh and Johansson 2007).

Available data can be found at many different places and in various formats. A significant share of available data stems from collection systems, automated or manual, which are used to follow up certain aspects of the system. For example order handling system, maintenance systems, staffing systems and other databases. Other common sources of data are systems used by other functions in the company, such as Enterprise Resource Planning (ERP) systems, Material Planning Systems (MPS) and Manufacturing Execution Systems (MES). Data can also be found in materials from previous analysis efforts, for instance frequency studies, Lean efforts, quality related projects or other system design processes.

The result of this activity is a list of sources for parameters classified as category A data. Moreover, if some sources are computerized, instructions on how to extract each parameter shall be included.

4.4 Choose Methods for Gathering of Not Available Data

If, as in most simulation studies, some or all data is not available beforehand (categories B and C), it needs to be either gathered or estimated. In this activity, gathering and estimation methods for data in these categories are defined and chosen. The choices will be the basis for the evaluation and decision in activity 4.5 (that all data will be found and that data collection can start). Note that no data collection starts in this activity, however, the choices will define how the actual gathering will be performed, later in activity 4.8.

To gather data for a DES model can be done in many ways. The most common, and probably easiest way, is to use a stopwatch and start to walk along the product flow, measuring parameters for each and every step of the process. At each process step, and for each different product, measurements are made for all parameters identified in activity 4.1. This method is rather chaotic but swift to conduct. Care needs to be taken when considering where and when a process ends and another starts. Hence, it is important to adhere to the parameter definitions, also established in activity 4.1. Moreover, buffer capacities and conveyor speeds can also be collected this way. If more than one person will collect the data, make sure that exactly the same way of measuring is used. Other examples of manual gathering methods are frequency studies and video analyses.

Time studies on a more detailed level are preferable if the project is to deliver more accurate results, or if all entities in the model do not yet exist in reality. It will, however, require more time spent during the input data phase of the project. MTM (Methods-Time Measurement 1973), SAM (Sequence-based Activity and Method analysis) (Johansson and Kinnander 2004) and DFA (Design for Assembly) (Boothroyd and Dewhurst 1989) studies can be used, for manual and automatic operations, during modification of existing or design of new assembly systems. For other new systems, it is recommended to use more process oriented simulation or emulation tools in order to create good quality input data for DES models. Cycle times can for instance be extracted from tools for offline programming of robots, PLC emulation, or code generation for NC-Machines.

However, many times when the system does not yet exist (category C data), no information at all is available and parameter values have to rely on estimates. Robinson (2004) gives three options to support this kind of guesswork: discussions with subject matter experts like machine vendors or in-house production engineers, review of historical data from similar systems in the same, or another, organization and, finally, for some processes there are standardized data available that is previously measured and stored in process libraries.

Another difficult situation for data gathering is when humans are involved. Humans will not act logically in all cases and are much more unpredictable than other parts of a system. Even though breakdowns of machines and other resources are unpredictable, they still tend to follow a distribution which can be modelled using random numbers to generate a failure. Gathering of data at manual stations also needs to be carefully planned in order to avoid the Hawthorne effect (Landsberger 1958) and to avoid annoying operators, which can jeopardize further cooperation.

4.5 Will All Specified Data Be Found?

It is necessary to check that all parameters will be possible to find with regard to the outcome of previous activities, e.g. available data, possible gathering methods and required number of samples. Hence, the decision is not a straightforward yes or no decision; aspects on enough data points, data accuracy and data quality have to be considered. If mistakes are made in this step, there is a risk to suffer from them, for example in activity 4.10, since too few data points in this step will give a bad estimate on the probability function, or in 4.12, since low quality data could be invalid.

If all data will be possible to find, one can proceed with a limited risk of future unnecessary iterations due to problems in the data collection process. On the other hand, if some parameters turn out to be impossible to collect, the accuracy requirements or the relevance of the parameter must be reevaluated.

4.6 Create Data Sheet

A data sheet needs to be established in order to maintain coherence in the data collection process. All raw data, as well as all analyzed data, should reside at the same place, usually a spreadsheet or, in large projects, a database. Unfortunately, many project teams try to save time by storing raw data in temporary spreadsheets and analyzed data directly in a simulation spreadsheet interface. Usually this approach gives the opposite effect due to lack of structure and loss of data, since information stored in the interface runs the risk of being overwritten.

To use pre-defined data structures such as CMSD (Lee et al. 2007), is an efficient way to design appropriate data sheets. The CMSD data structure is based on a Unified Modeling Language (UML) scheme, from which an eXtensible Markup Language (XML) instance document can be generated, in order to store specific data for a model or a system. Many models can reside their data in the same XML document, if desired.

4.7 Compile Available Data

In this activity, all data in category A is collected or extracted from the sources of available data, identified in step 4.3. Category A data can be found as raw data, for instance automatically measured cycle times or time-stamps stating start and stop times of breakdowns. However, category A data can also be previously analyzed and ready to use in simulation models, either as a result of previous projects or because the same data is used by other functions in the organization.

Previously analyzed data is ready to await data validation in step 4.11, but raw data requires a lot more efforts. Based on the number of samples for all parameters, decided as a result of the accuracy requirements in step 4.2, a sufficient amount of data points are extracted from the sources, usually databases. Thereafter, additional calculations are often needed to convert the samples into a suitable form. For instance, to obtain TTR information from the breakdown time-stamps exemplified in the paragraph above, the stop times need to be subtracted from the start times. Moreover, a majority of cases requires some kind of filtering process, for example to exclude incorrectly measured samples or data points from shifts that do not represent a normal system state. The final result of this activity is sets of raw data points (for instance 230 individual cycle times from an assembly station) ready to analyze in order to prepare a statistical or empirical distribution in step 4.9. For pre-analyzed data, further preparation is usually not required and can, thus, be finally reported in the data sheet.

4.8 Gather Not Available Data

This activity includes measurements of previously unavailable production data, but also estimation of performance for future equipment. Hence, it will change data from being category B or C (Robinson and Bhatia 1995) to become category A data. The input to this activity is which parameters to measure (from activity 4.1), how many samples to gather for category B parameters (from activity 4.2) and which gathering methods to use (from activity 4.4).

For category B data, the activity might consume quite some time, since data gathering often equals manual work. If the system to be modeled has a high frequency of products, it might be quicker. However, if cycle times are long, data gathering is surely a time-consuming process due to the fact that more than 200 samples are often preferable (Perrica et al. 2008). To gather category C data, on the other hand, is usually less time-consuming if the assumptions are based on information from process experts. However, if the assumptions are based on historical data from similar processes, gathering of category C data can also be rather time-consuming.

The result of this activity is, in conformity with activity 4.7, sets of raw data ready to analyze and prepare for simulation in step 4.9. For category C data, the results are often given on a form that is already suitable for simulation and can be finally reported in the data sheet.

4.9 Prepare Statistical or Empirical Representation

The actual data collection in activities 4.7 and 4.8 results, as stated, either in already pre-analyzed data and/or in sets of raw data, which need a way to be represented in simulation models. For constant data, the analysis part is usually not very arduous, but data describing variability require some more efforts. The variability needs to be represented, generally using one of the following four options (Robinson 2004): traces, empirical distributions, bootstrapping, or statistical distributions. Of course all four options have pros and cons, which should be evaluated before a choice is made. However, three of the four options are quite straightforward using basic mathematics, but the fourth alternative, input modeling with statistical distributions, requires more attention.

Moreover, the statistical representation is a very popular way to describe variability when possible and, hence, much research adopting this approach is available. Fortunately, there are numerous tools supporting the process of input modeling. ExpertFit® and Stat::Fit® are two examples and many of the commercial simulation software packages also hold functionality in input modeling. For those who have no access to one of these tools, it is necessary to do it the hard way. Leemis (2004) gives a good description of "manual" input modeling including the following steps:

- Assess sample independency
- Chose one or more distribution families to evaluate
- Estimate parameters, for example using MLE
- Assess model adequacy using a goodness-of-fit test
- Visualize the model adequacy using P-P or Q-Q plots

The result of this activity is that the data sheet is completed with data representations that are ready to use in the simulation model.

4.10 Sufficient Representation?

The decision whether the representations delivered by activity 4.9 are sufficiently adequate is not always easy to make. At best, a chosen statistical distribution can be mathematically justified by passing a goodness-of-fit test, usually at level $\alpha = 0.05$ (Perrica et al. 2008). However, especially for large number of samples, goodness-of-fit tests are very conservative and it is almost impossible for any representation to pass. Hence, it is very important for the simulation engineer to decide the required level of significance according to the accuracy requirements specified for each parameter. For this reason, graphical comparison of the representation and the original data might be preferable in some situations. Later during the simulation project, a sensitivity analysis can be made on representations with weaker correspondence to the real-world data. In this way, critical parameters are identified and additional investigations on data accuracy can be required for these factors.

If representations are insufficient according to the accuracy requirements, additional data collection and analysis are needed. Other solutions are to change the representation of variability (se activity 4.9) or in worst case to reconsider the accuracy requirement for a specific parameter and consequently also for the entire simulation model.

4.11 Validate Data Representations

Data validation is an important activity to make sure that all raw data is correctly measured and filtered, and that calculations and analyses during the preparation process are properly performed. The activity is very difficult according to Sargent (2005), who states that "there is not much that can be done to ensure that the data is correct". One reason is that data in itself rather often is a part of the validation procedure. Nevertheless, to ensure face validity and to stick to good procedures along the entire data collection process is a good start.

Face validity can be achieved by cooperating with process experts during the entire input data management phase and also setting up a final check towards the end, for instance using more structured interviews. Moreover, in addition to face validity there are other methods to validate the data before using it in the simulation model. One example is to evaluate data with regard to production follow-ups; e.g., breakdown data can be compared to previously performed measurements on equipment availability. Sargent (2005) also mentions "comparison to other models" as a technique to validate entire simulation models. The technique can also be applied in data validation by comparisons to known results or to data in other, previously validated models, including similar equipment.

Since the model always will be a simplified representation of the real system, it is of great importance to understand what data is crucial for model performance. As in activity 4.2, it is of course more important to make thorough validations on crucial parameters than on parameters of lesser importance. To finally make sure that no mistakes are made in the process of differing between central and non-central parameters, a sensitivity analysis can be performed once the model is built.

Finally, it is important to notice that validation of the data will be done once more during model building, since the data will be a part of the model validation later, given that a project methodology such as those described in Law (2007) and Banks et al. (2004) is followed. Still, a good data validation is a very efficient way to reduce the need for late additional iterations of data collection, since possible mistakes are detected as early as possible. It is also easier to pinpoint the root cause of a failed separate data validation than in a complete model validation.

4.12 Validated?

If the data validation succeeds for all parameters during activity 4.11, the representations are ready to use in the simulation model. Due to previously mentioned difficulties with data validation, the project team should remember that data can still be the problem causing a failed model validation later during the project. However, data validation is a good start that prevents many unnecessary future iterations of data collection.

On the other hand, if the data validation fails for one or more of the parameters, there will be a need to step back and identify the cause of the problem. Many times the problems stem from miscalculations in the analysis and preparation activity (4.9) but sometimes further gathering or extraction of raw data cannot be avoided. On rare occasions one might need to go all the way back and reevaluate the chosen gathering methods.

4.13 Finish Final Documentation

Documentation is a continuous process throughout the entire input data management phase, starting already in the first activity where parameters are identified and defined. Much of the information to document should already be available in the data sheet, including selected parameters, raw data, and finally chosen simulation representations. However, there are often things of importance for future referencing and reuse, which are not in the data sheet. For example, the sources of data, the gathering methods, the validation results, and all assumptions made during the input data process are all of great importance for maintaining future data validity. The final result of this activity is a data report and the completed data sheet. Both of them go into the final documentation of the entire simulation project.

5 CONCLUSION

The purpose of this study is to present a structured methodology for the input data management process in DES projects. The intention is to cover all aspects of the process, including identification, collection, and preparation of input data for simulation models. As a result, this paper proposes a structured methodology including 13 activities and their internal connections. During a review of 15 previously performed simulation projects within industry, a lack of a clear mode of operation for handling input data was identified. Moreover, the results show that a more structured way to work holds significant potential to increase both rapidity and quality in the input data phase of DES projects.

Similar methodologies to the one presented in this study already exist for other parts of DES projects. For instance, Sargent (2005) outlines a set of activities for verification and validation of simulation models. Furthermore, even more known methodologies are available on a macro level, describing efficient ways to perform entire simulation projects, see for example Banks et al. (2004) and Law (2007). Simulation practitioners seem to find this kind of methodologies very helpful in their daily work, especially those who have not previously been involved in an extensive number of DES projects.

However, it is important to highlight that there are previous contributions explaining detailed methods for collection and analysis of simulation data. For instance, Leemis (2004) and Law (2007) describe the process of input modeling, mainly from a statistical perspective. Moreover, Perera and Liyanage (2000) presents a methodology for rapid identification of input parameters. Hence, our work does not intend to give any contributions on this more detailed level. Instead, we focus on linking all activities within input data management in an efficient way.

Some of the projects evaluated in this study are performed in companies with limited experience of DES. Additionally, many of the project members do not work with simulation on a daily basis as their only work assignment. The authors suppose that the profit of using the methodology is largest in such circumstances, since more experienced organizations and simulation engineers continuously discover and document efficient working procedures in an iterative manner. Still, there is always a risk of following an old route and, thus, the proposed methodology can be of value for these organizations as well.

6 FUTURE RESEARCH

For future work, we will validate the proposed methodology and evaluate its impact on data quality and rapidity in input data management. Skoogh and Johansson (2007) have measured the total time-consumption in the input data phase of DES projects that did not follow any structured way of working during their data collection and preparation. We will introduce the proposed methodology in several simulation projects starting in upcoming years and measure the time-consumption. Consequently, the impact of the new methodology will be quantified.

In parallel to the evaluation of the proposed methodology, our research group also works with development of a generic data management tool (the GDM-Tool). This work focuses on improving efficiency in companies that have advanced far into the implementation of well designed computer applications for logging and storage of production data. The tool is configurable to both standardized and custom made data sources and automates many of the time-consuming activities discussed in this paper, for instance data extraction and statistical analysis.

ACKNOWLEDGMENTS

The funding for this research is granted by VINNOVA (Swedish Agency for Innovation Systems, integrates research and development in technology, transport and working life.), and SSF (Swedish Foundation for Strategic Research). The authors would also like to thank all members of the DES projects included in the study. They shared a significant amount of time by taking part in the interviews. Edward Williams (University of Michigan and PMC, Dearborn, Michigan, U.S.A.) has kindly provided valuable suggestions for enhancing the presentation of this paper.

REFERENCES

Banks, J., J. S. Carson, B. L. Nelson, and D. M. Nicol. 2004. *Discrete-Event System Simulation*. 4th ed. Upper Saddle River, New Jersey: Prentice-Hall Incorporated

Boothroyd, G., and P. Dewhurst. 1989. *Product Design For Assembly*. New York: McGraw-Hill, Inc.

Denscombe, M. 1998. *The good research guide: for small-scale social research projects*. Buckingham: Open University Press.

Ericsson, U. 2005. *Diffusion of Discrete Event Simulation in Swedish Industry*. Doctorial dissertation, Department of Materials and Manufacturing Technology, Chalmers University of Technology, Gothenburg, Sweden.

Hatami, S. 1990. Data requirements for analysis of manufacturing systems using computer simulation. In *Proceedings of the 1990 Winter Simulation Conference*, ed. O. Balci, R. P. Sadowski, and R. E. Nance, 632–635. New Orleans, Louisiana.

Ho, C-F., W-H. Wu, and Y-M. Tai. 2004. Strategies for the adaptation of ERP systems. *Industrial Management & Data Systems* 104:234-251.

Hollocks, B. W. 2001. Discrete-event simulation: an inquiry into user practice. *Simulation Practice and Theory* 8:451-471.

Johansson, B., J. Johnsson, and A. Kinnander. 2003. Information structure to support discrete event simulation in manufacturing systems. In *Proceedings of the 2003 Winter Simulation Conference*, ed. S. Chick, P. J. Sánchez, D. Ferrin and D. J. Morrice, 1290–1295. New Orleans, Louisiana.

Johansson, B., and A. Kinnander. 2004. *Produktivitetsförbättring av manuella monteringsoperationer* (in Swedish). Chalmers University of Technology report ISSN 1651-0984, Internapport 004:25.

Landsberger, H. A. 1958. *Hawthorne Revisited*. Ithaca: Cornell University Press.

Law, A. M. 2007. *Simulation modeling & analysis.* 4th ed. New York: McGraw-Hill, Inc.

Lee, Y. T., S. Leong, F. Riddick, M. Johansson, and B. Johansson. 2007. A Pilot Implementation of the Core Manufacturing Simulation Data Information Model. In *Proceedings of the Simulation Interoperability Standards Organization 2007 Fall Simulation Interoperability Workshop*. Orlando, Florida: Simulation Interoperability Standards Organization, Inc.

Leemis, L. 2004. Building credible input models. In *Proceedings of the 2004 Winter Simulation Conference*, ed. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 29–40. Washington, D.C.

Lehtonen, J-M., and U. Seppälä. 1997. A methodology for data gathering and analysis in a logistics simulation project. *Integrated Manufacturing Systems* 8:351-358.

MTMA. 1973. *Methods-Time Measurement*. MTM Association for Standards and Research, Fairlawn, New Jersey.

Pegden, C. D., R. E. Shannon, and R. P. Sadowski. 1995. *Introduction to simulation Using SIMAN*. 2nd ed. New York: McGraw-Hill.

Perera, T., and K. Liyanage. 2000. Methodology for rapid identification of input data in the simulation of manufacturing systems. *Simulation Practice and Theory* 7:645-656.

Perrica, G., C. Fantuzzi, A. Grassi, G. Goldoni, and F. Raimondi. 2008. Time to Failure and Time to Repair Profiles Identification. In *Proceedings of the 5th FOODSIM conference*. Dublin, Ireland

Pidd, M. 1995. *Computer simulation in management science*. 2nd ed. Chichester: John Wiley & Sons.

Pidd, M. 2003. *Tools for Thinking: Modelling in Management Science*. 2nd ed. Chichester: John Wiley & Sons

Randell, L. G., and G. S. Bolmsjö. 2001. Database driven factory simulation: a proof-of-concept demonstrator. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 977-983. Arlington, Virginia.

Robertson, N., and T. Perera. 2002. Automated data collection for simulation?. *Simulation Practice and Theory* 9:349-364.

Robinson, S. 2004. *Simulation: The Practice of Model Development and Use*. Chichester: John Wiley & Sons

Robinson, S., and V. Bhatia. 1995. Secrets of successful simulation projects. In *Proceedings of the 1995 Winter Simulation Conference*, ed. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman, 61–67. Arlington, Virginia.

Sargent, R. G. 2005. Verification and validation of simulation models. In *Proceedings of the 2005 Winter Simulation Conference*, ed. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 130–143. Orlando, Florida

Skoogh, A., and B. Johansson. 2007. Time-consumption analysis of input data activities in discrete event simulation projects. In *Proceedings of the 2007 Swedish Production Symposium*. Gothenburg, Sweden.

Trybula, W. 1994. Building simulation models without data. *In 1994 IEEE International Conferance on Systems, Man, and Cybernetics. Humans, Information and Technology*, 1:209-214. IEEE

Williams, E. J. 1996. Making Simulation a Corporate Norm. In *Proceedings of the 1996 Summer Computer Simulation Conference*, ed. V. W. Ingalls, J. Cynamon, and A. V. Saylor, 627–632. New Orleans, Louisiana.

AUTHOR BIOGRAPHIES

ANDERS SKOOGH is a PhD student in the field of Discrete Event Simulation at the Department of Product and Production Development, Chalmers University of Technology, Sweden. In 2005 he obtained his M.Sc. degree in Automation and Mechatronics from the same university. Anders has industrial experience of Discrete Event Simulation from his former employment as logistics developer at Volvo Car Corporation. His email is <Anders.Skoogh@chalmers.se>.

BJÖRN JOHANSSON is an assistant professor at Product and Production Development, Chalmers University of Technology, currently also a guest researcher at National Institute of Standards and Technology in Gaithersburg, Maryland, USA. His research interest is in the area of discrete event simulation for manufacturing industries. Modular modeling methodologies, environmental effects modeling, software development, user interfaces, and input data architectures are examples of interests. His email address is <Bjorn.Johansson@chalmers.se>.