SENSITIVITY ANALYSIS OF OPTIMIZATION
Examining sensitivity of bottleneck optimization to input data models

Master Degree Project in Automation Engineering
One year 30 credits
Spring term 2016

Marie Ekberg

Supervisor: Amos H.C. Ng
Examiner: Anna Syberfeldt
Abstract

The aim of this thesis is to examine optimization sensitivity in SCORE to the accuracy of particular input data models used in a simulation model of a production line. The purpose is to evaluate if it is sufficient to model input data using sample mean and default distributions instead of fitted distributions. An existing production line has been modeled for the simulation study. SCORE is based on maximizing any key performance measure of the production line while simultaneously minimizing the number of improvements necessary to achieve maximum performance. The sensitivity to the input models should become apparent the more changes required. The experiments concluded that the optimization struggles to obtain convergence when fitted distribution models were used. Configuring the input parameters to the optimization might yield better optimization result. The final conclusion is that the optimization is sensitive to what input data models are used in the simulation model.

**Keywords:** simulation, optimization, input modeling, probability distribution, simulation-based constraint removal, production systems
# Table of Contents

1 Introduction .................................................................................................................. 1  
1.1 Research motivation and background ...................................................................... 1  
1.2 Research aim and objectives .................................................................................... 1  
1.3 Method....................................................................................................................... 2  
1.3.1 Experimental method ............................................................................................ 3  
1.4 Project limitations ..................................................................................................... 3  
1.4.1 All failures for each machine are modeled by one input data model .................. 3  
1.4.2 Possible issues with data quality and reliability ................................................. 4  
1.4.3 Basic simulation model validation ....................................................................... 4  
2 Background .................................................................................................................. 5  
2.1 Simulation-based optimization .................................................................................. 5  
2.1.1 Optimization ........................................................................................................ 5  
2.2 Discrete event simulation ......................................................................................... 6  
2.3 Multi-objective optimization ..................................................................................... 6  
2.3.1 Important Pareto front properties ........................................................................ 8  
2.4 Bottlenecks and system performance ...................................................................... 8  
2.4.1 Simulation-based constraint removal ................................................................... 9  
2.5 Flow production data ................................................................................................ 9  
2.6 The importance of data quality ............................................................................... 10  
2.7 The effect of variability .......................................................................................... 11  
2.8 Input probability distributions .................................................................................. 12  
2.8.1 Empirical modeling ............................................................................................... 12  
2.8.2 Data analysis and model selection ....................................................................... 13  
2.8.3 Parameter estimation .......................................................................................... 13  
2.8.4 Model validation ................................................................................................... 13  
2.8.5 Input data modeling problems ............................................................................ 14  
2.8.6 Useful probability distributions as seen in FACTS ............................................. 15  
3 Input data management ............................................................................................... 16  
3.1 Simulation models and model generation ............................................................... 16  
3.2 Data collection ......................................................................................................... 17  
3.3 Input data modeling .................................................................................................. 19  
3.3.1 Data analysis ....................................................................................................... 19  
3.3.2 Summary statistics using ExpertFit .................................................................... 19  
3.3.3 Graphical estimates in ExpertFit ......................................................................... 20  
3.3.4 Model input parameters ...................................................................................... 21  
3.4 Probability distribution models ................................................................................ 21  
3.4.1 ExpertFit .............................................................................................................. 21  
3.4.2 GDM Tool ........................................................................................................... 22  
3.4.3 Distribution fitting .............................................................................................. 22  
3.5 Updating the models ................................................................................................ 22  
3.5.1 FACTS support tool ............................................................................................ 22  
4 Experiments ................................................................................................................. 24  
4.1 Simulation experiments ............................................................................................ 24  
4.1.1 Simulation models ............................................................................................... 24  
4.1.2 Simulating manufacturing using FACTS ............................................................ 25  
4.2 Optimization and setting up SCORE ....................................................................... 25
4.2.1 Optimization input parameters ................................................................. 26
4.2.2 Estimating Weibull parameters ................................................................. 26
4.2.3 FACTS variables ...................................................................................... 27
4.2.4 Optimization objectives ............................................................................ 27

5 Analysis ...................................................................................................... 29
5.1 Validating the distribution models ............................................................... 29
5.1.1 Absolute model evaluation in ExpertFit ..................................................... 30
5.2 Simulation model validation ......................................................................... 32
5.3 Optimization Pareto fronts .......................................................................... 33
5.3.1 The effects of estimating Weibull parameters .............................................. 35
5.3.2 Searching the optimal region ..................................................................... 35

6 Conclusions ............................................................................................... 37
6.1 Summary of results ..................................................................................... 37
6.2 Discussion .................................................................................................. 37
6.3 Future work ............................................................................................... 38
6.3.1 Modeling of failures .................................................................................. 38
6.3.2 Improved data reliability .......................................................................... 38
6.3.3 Optimization input set configuration ......................................................... 39

References .................................................................................................... 40
1 Introduction

The motivation, aim and objectives of the research in this thesis are explicated in this chapter. The aim of the thesis is motivated by stating relevant observations about the subject and describing a related problem and its relevancy in the manufacturing industry. The aim of the research and corresponding objectives are then defined with respect to the problem. The chosen methodology used to evaluate the result of the study is described in detail. The limitations of the study are presented at the end of this chapter.

1.1 Research motivation and background

Simulation-based optimization has proven to be a powerful combination of the two different techniques regarding analyzing the inner workings of complex manufacturing systems (Ólafsson & Kim, 2002). The technique can be used in supporting decisions concerning possible actions of improvement to existing production lines in order to remove bottlenecks constraining the performance of the line (Ng et.al., 2014). In the manufacturing industry it is important to obtain the maximum possible performance in the production to be able to secure the position as a leading and thriving company in the highly competitive market (Bernedixen et.al., 2015). It is therefore crucial that any limiting factors in the production system impeding the performance (i.e. bottlenecks) are removed, and that any improvements are spent on the key machines first to allow maximum gain in performance to a minimum cost of changes to the system (Ng et.al., 2014).

In order for the simulation results to be justified as decision support in deciding where to invest and improve in the production line to achieve the optimal increase in performance, it is important that the simulation is producing reliable and accurate estimations of the true system. The validity of the simulation results depends on the accuracy of the model, and how accurately the input data is modeled in the simulation model. Accuracy in this sense is however relative and is reflected by the problem to be investigated by the simulation study. The particular problem at hand determines what is important to capture accurately in the simulation model regarding estimating the behavior of the true system. If system variability is important for the specific problem, it should be captured in the simulation model. It is stated by Moris et.al. (2008) that it is seldom appropriate to model a data set by using constant values, since it will affect the accuracy of the model. It is therefore interesting to investigate when it is appropriate to use sample mean combined with “default” distributions, and when it is important to use fitted probability distributions to model input data. Determining the appropriate probability distribution model to represent the variability in data is in itself a difficult and time-consuming process, especially in simulation studies of complex systems and large sets of data samples, requiring input data analysis and model validation. If it is justified to skip this step and use sample mean and default distributions instead and obtain just as valid simulation results, this is valuable information.

1.2 Research aim and objectives

According to an analysis made by Skoogh and Johansson (2007), the input data management phase in a simulation study constitutes on average 31 % of the total time. Since input data management in a simulation study (including collecting required data and input data modeling) is a time-consuming process, it is motivated to reduce the necessary time
such that the simulation and optimization results still are valid with respect to the current problem. When it comes to input data modeling in simulation studies, it has been stated that it is of great importance to model data by fitted probability distributions and to avoid modeling by sample mean (Law, 2015) and using constant values (Moris et.al., 2008) for accurate and valid results. However, modeling an input data set by fitting probability distributions is in itself a time-consuming process, which there might be no or little time for. Moreover, when there is no significant variability expressed in a particular data set, it might be sufficient to model by using sample mean (Murthy et.al., 2004).

The aim of this thesis is as follows:

- Investigate how sensitive the optimization is to different methods of modeling input data in a simulation model using simulation-based constraint removal (SCORE)

The purpose of this aim is to determine whether it is justified for some problems to model input data by sample mean and default distributions instead of using fitted probability distributions, without compromising the validity of the optimization results. Optimization sensitivity refers to how the Pareto front is affected by the accuracy of the input models used. The optimization sensitivity is estimated by perceived properties of the respective Pareto front in terms of convergence and diversity. The described aim will be accomplished by following objectives:

1. Collect data from selected production system required for the simulation study
2. Build simulation model of the selected production system
3. Analyze collected data and model input data by sample mean and fitted distributions
4. Experiments using SCORE optimization
5. Evaluate and compare Pareto fronts from the optimization results

1.3 Method

The aim of this thesis is to determine if there is any difference between the Pareto fronts of the optimization results from SCORE in terms of convergence and diversity, when input data modeling techniques of varying accuracy are used in the simulation model. Hopefully there will not be any distinct difference. The Pareto fronts will be visually interpreted to determine the sensitivity of the optimization. The input data modeling techniques considered are as follows:

1. modeling by sample mean combined with underlying default distributions in FACTS
2. modeling by fitting probability distributions using distribution fitting software
3. modeling by sample mean and fitted probability distributions

The study will be applied on an existing production system containing automated machines from a Swedish automotive manufacturer. The simulation software used in this thesis has been FACTS Analyzer (Ng et.al., 2007). The machine representations in the simulation model are described by their process time, time to repair and their availability. The process time can be modeled by either sample mean using a constant value, or modeled by a fitted probability distribution. The disturbances in the system can be modeled by using sample mean combined with an underlying default distribution in FACTS, or by using fitted
distributions. The default distributions in FACTS are based on experience and knowledge from the car manufacturing industry. It is therefore interesting whether these default distributions modeling disturbances of the production line are accurate enough to provide optimization result equal to that of a simulation model using fitted distributions. The preliminary hypothesis is that modeling process time in automated machines by sample mean and using the default distributions of the simulation software to model disturbances in the system, is accurate enough to provide desirable optimization results. This thesis will conclude whether this is true by visually comparing how the distinct input data modeling techniques affect the optimization outcome from SCORE. The result will be evaluated by visually comparing the produced Pareto fronts in terms of convergence and diversity, in order to determine the optimization sensitivity.

1.3.1 Experimental method

The production system is modeled into a suitable representation using FACTS Analyzer, since FACTS provides good support for aggregated modeling (Ng et.al., 2007). Necessary data regarding the production line have been collected from an internal manufacturing monitoring system connected to the system. The collected data consists of several, and often large data sets, describing relevant aspects of the production system. The data sets are analyzed using a preliminary input analysis of basic descriptive statistics using the ExpertFit software (Law, 2011 b). The analysis provides valuable information about the characteristics of each data set, which is useful when evaluating the fit of a particular distribution. Based on this information, the data sets are modeled by sample mean and by an “exact” distribution fit and processed to produce input parameters in the accepted format of FACTS. Each data set has been modeled by a probability distribution using the distribution fitting software ExpertFit (Law, 2011 b) and GDM Tool (Skoogh, 2009). Before any experiments are conducted, the simulation models are validated against the true production system. Each model will then be used in experiments of SCORE optimizations. SCORE is a technique that is appropriate for comparing the possible difference between the different input data modeling techniques in the simulation models. In SCORE, the optimization searches for solutions in the model that maximizes the key performance measure of the production line, while the number of corresponding necessary improvements is minimized. If there is any apparent difference between the distinct input modeling methods, the optimization result should be more sensitive to the solutions consisting of numerous changes to the system. The Pareto fronts from the optimization will be discussed with respect to convergence and diversity in order to evaluate the optimization sensitivity to the specific input data models used in the simulation model.

1.4 Project limitations

There are some limitations that the work of this thesis is confined to. In the context of this thesis, a limitation has been defined as any constraining factor of the work done that has been noted on beforehand and its effects hence considered. The limitations are important to have in mind when the result of the experiments is presented, evaluated and used to draw the final conclusions of this work. These limitations and their consequences are presented and described in this chapter.

1.4.1 All failures for each machine are modeled by one input data model

The disturbances in the production system are modeled by time to repair and time between failures for each machine. Disturbances in the system can be caused by different types of
failures. However, the failures are treated as if they all are of the same type due to the low number of data samples available for each failure. For example, it is not distinguished between short and long stops in a machine. The available disturbance data for each machine are treated as independent and identically distributed data samples, which might not be true. Data samples coming from different types of failures might come from distinct distributions. Combining data from heterogeneous distributions into one distribution might give a false picture of how disturbances affect the particular machine, and hence compromise the validity of the simulation results.

1.4.2 Possible issues with data quality and reliability

Another potential concern is the quality of the collected input data. Collecting the necessary data for the simulation study has been confined to the manufacturing monitoring system surveying the production system. When collecting data describing the behavior of the production system to use in the simulation study, it is important that representative data and sufficient amount of data samples are collected. It has been difficult to collect a large amount of data samples spanning over a large period of time from the monitoring system due to the extensive amount of time it takes to export data. There are also some other problems related with data reliability. The monitoring system is capable of logging most of the data automatically. The exception regards some of the failure types that have to be registered manually by an operator of the current production line. If this manual step isn’t properly managed, data can be either missing or be erroneous and therefore unreliable.

1.4.3 Basic simulation model validation

The simulation model has not been properly validated against the true system. This is because the model itself has been constructed using high abstraction level and aggregated modeling. It has been outside the scope of this thesis taking these aspects into account when validating the simulation model. However, the effect of using different input data models in the simulation has been examined by running simulations and comparing each result with the plant output from the true production system. The plant output considered has been throughput, lead time and work in process. The plant output of the true production system has been acquired by consulting domain experts of the production line. This result have been used to somewhat determine the validity of the input data models in comparison to each other.
2 Background

This chapter presents relevant theory related to the problem of research that is necessary to be able to understand the core concepts of the described problem and to comprehend following discussion and conclusions of this thesis. Theory regarding the combination of simulation and optimization for finding the potential for improvement in existing production systems related to its performance is presented. The essential aspects of simulation and building simulation models are covered alongside with input data analysis and its importance for simulation studies. When nothing else is stated, all concepts presented are related to the domain of the manufacturing industry.

2.1 Simulation-based optimization

Simulation-based optimization is described by Law and McComas (2002) as probably one of the most important new simulation technologies in recent years. Applying simulation in aiding the design and analysis of manufacturing systems is one of numerous successful application areas where simulation has been found both useful and powerful (Law, 2015). The purpose of simulation is to use computers to evaluate models numerically, as opposed to analytically, where the gathered data is used to draw conclusions about the true system (Law, 2015). For numerous real-world problems it is not possible to use analytical methods due to the complexity of the problem and hence simulation is the only available option (Law, 2015). The technique combines the two separate techniques of simulation and optimization, and it has proven to be a useful combination when analyzing complex systems in the manufacturing industry (Ólafsson & Kim, 2002). It is a powerful method for finding the optimal configuration of a production system based on certain assigned objectives. Such an objective typically concern optimizing the performance of the system in terms of relevant manufacturing performance measures e.g. throughput, lead time and work in process. In order to analyze a production system using simulation-based optimization, a digital model is used to represent the system in a computer and to simulate its behavior during particular scenarios and conditions. A simulation model is different from a physical model as it is purely mathematic, defined by logical and quantitative relationships and by the input data collected from the true system (Law, 2015). The model should represent an appropriate abstraction of the production line with respect to the problem at hand, populated with data collected from the real production system. For successful simulation results with accurate estimations of the behavior of the true production system, it is important that the model is valid and the collected data is reliable and of high quality (Law, 2015).

2.1.1 Optimization

The current goal of the optimization is determined by the user, and is defined by one or several objectives regarding desirable “areas” for improvement in the production system. The optimization objectives concern maximizing or minimizing different measures of performance. In the manufacturing context this could for example concern maximizing throughput while minimizing required buffer capacity, or other relevant aspects of the production line with potential for improvement. The optimization is assigned a set of input parameters, or decision variables, from the simulation model. The decision variables are based on available attributes in the model and are selected depending on the goal of the optimization. If the objective is to maximize production throughput, it would for example be useful to add the cycle times of the machines as decision variables. The optimization will
search for near-optimal combinations of values for these decision variables with respect to current optimization objectives. In the search for improvement, each set of values will be evaluated through the simulation to estimate the performance of the production line during the given conditions. Based on the simulation result, the optimization will accordingly alter the values of the decision variables to create new conditions for the simulated production line to operate within. The optimization will continue until a sufficient number of possible near-optimal solutions have been found (Olafsson & Kim, 2002).

2.2 Discrete event simulation

Discrete-event simulation is based on simulating the operation of a system through events occurring at discrete time steps, where the occurrence of an event instantly could change the state of the modeled system (Law, 2015; Banks et al., 2014). The state can only be changed by events; in between events the system is expected to be in a steady state where only time is changing. The state of the system is modeled by any variable essential in describing its current status at all times, or is relevant for the particular study to be carried out. In the manufacturing context, variables representing the state could for instance contain information about whether a particular machine is operational or down. The triggering of events is stochastic and simulated by random variables generated by statistical probability distributions for each type of event. An example from the manufacturing application would be the event of a machine breakdown for a particular machine and failure.

The model built for the simulation has the purpose of representing the system that is to be studied, and is by definition a simplification of the true system (Banks et al., 2014). Even though the model is an abstraction, it must still have an adequate amount of detail from the true system to allow the simulation results to be used as decision support for any valid decisions regarding the real system (Banks et al., 2014). The model consists of entities representing the relevant components from the true system. In the case of modeling a production system, entities like machines and buffers are often necessary, with attributes such as machine cycle time, failure times and buffer capacity. A simulation model can either be deterministic or stochastic, which determines if the output from the simulation is deterministic or random. In the context of modeling production systems, the model is stochastic since it has several input parameters that are based on random variables such as failure time. Since random inputs create random outputs, the simulation results can only be estimations of the behavior of the real system (Banks et al., 2014). It is therefore important that the estimates are accurate.

2.3 Multi-objective optimization

The goal of applying optimization on an existing production line is often to find possible areas for improvement in the line regarding the performance of the system. The performance can for example be measured in terms of throughput, work in process, lead time or optimal buffer allocation. It is important to understand that it is misleading to use a single-objective optimization approach in achieving this, since each objective often strongly depends on other objectives in a conflicting manner. If a production system is optimized with the objective to enhance the system performance by maximizing the throughput, the optimization will greedily search for solutions where throughput is maximized without considering other important aspects relevant for the final performance of the system. As an example, consider
Little’s law (Little, 1961), where work in process (WIP), throughput (TH) and lead time (LT) in a stable system are correlated in the long term as described by Equation 1:

\[ \text{WIP} = \text{TH} \times \text{LT} \]  

(1)

Based on this equation it is concluded that throughput is correlated with work in process where high throughput results in high work in process. Maximizing throughput as a single-objective will also increase work in process, which is not desirable. This means that the objective of maximizing the throughput of the system will stand in conflict with the objective to keep work in process to a minimum. To optimize with both of these goals in mind, another approach called multi-objective optimization (MOO) (Deb, 2001) should be used. When two or more objectives are conflicting, it means that they will have different corresponding optimal solutions. An example of two conflicting objectives is shown in Figure 1. The graph illustrates possible solutions where a trade-off has been made between throughput and work in process.

Figure 1  An optimization problem where throughput (TH) is maximized while work in process (WIP) is minimized. The Pareto-optimal front clearly illustrates the conflicting nature of the two objectives.

The effect of two conflicting objectives in an optimization is that their corresponding optimal solutions will be contradicting each other. This makes it impossible to find an optimal solution with respect to both objectives (Deb, 2001). Instead, a set of trade-off solutions between the conflicting objectives will be found. Based on domain knowledge and experience, the most appropriate trade-off can be selected. For all multi-objective optimization problems, there exists a true Pareto-optimal front which contains the trade-off solutions where no single solution can be said to be the better or worse than any other with respect to all objectives in the entire search space (Deb, 2001). These solutions are called the Pareto-optimal solutions. These exact solutions cannot be found in reality, but through multi-objective optimization, solutions close to the true Pareto-optimal front can be found. These solutions are referred to as near-optimal solutions.
2.3.1 Important Pareto front properties
To yield successful optimization results, the Pareto front of near-optimal solutions should satisfy two main properties (Deb, 2001):

- Convergence
- Diversity

These properties and their relation to the true Pareto optimal front are illustrated in Figure 2. Note that these properties can be obtained independently of each other, and it is hence possible to have good convergence, but poor diversity.

![Comparison between the true Pareto-optimal front and the Pareto front of near optimal solutions.](image)

**Figure 2** Comparison between the true Pareto-optimal front and the Pareto front of near optimal solutions.

Good convergence of a Pareto front is desirable, since this property measures how close the near-optimal solutions are to the true Pareto optimal front and hence guarantees that the found solutions are close to optimal (Deb, 2001). Satisfying good diversity in a Pareto front is of equal importance, since it assures that the found solutions are well-spread in the optimal region (Deb, 2001). This implies a good set of varying near-optimal solutions, providing rich decision support and allows a decision maker to select the best possible tradeoff solution for the particular problem at hand.

### 2.4 Bottlenecks and system performance

The motivation behind applying simulation-based techniques on existing production systems is to improve their performance. In real-world systems it is not uncommon that the performance of an existing production line might be below expected levels due to several unknown constraints in the system (Bernedixen et.al., 2015). These performance issues are caused by some constraining characteristic of the production line acting as a bottleneck and significantly limiting the performance of the whole system. There are numerous different definitions of the term bottleneck, but in the scope of this thesis the definition provided by Ng et.al. (2014) will be used, where it is stated that the bottleneck of the system is where the smallest change has the greatest impact on the overall performance increase of the system.

The performance of the manufacturing system can only be improved if the bottlenecks in the system are removed (Ng et.al., 2014). The difficulty in removing bottlenecks in a system is identifying where they are located and what improvements that are necessary in order to remove them. This is a challenging task even for experienced people in the manufacturing industry (Ng et.al., 2014). Existing methods to identify bottlenecks in a system do not
provide sufficient information to support decisions concerning the improvement actions necessary to remove the bottlenecks (Ng et al., 2014). This is the reason that Ng et al. (2014) propose a multi-objective optimization approach in identifying the bottlenecks and finding the most beneficial improvements to the system. The large search space of possible combinations of improvements to the system makes it a suitable problem for optimization. Improvement in this sense refers to improving various system parameters, such as machine cycle time, availability and mean down time (Ng et al., 2014). The optimization problem will be defined by two objectives: maximizing the key performance measure of the system and simultaneously minimize the number of changes necessary to reach that performance level.

2.4.1 Simulation-based constraint removal

Simulation-based constraint removal, or SCORE, is a simulation technique for detecting bottlenecks in the order of their significance as performance constraint to the production system (Bernedixen et al., 2015). Not only the primary bottlenecks are identified; the secondary bottlenecks and other minor bottlenecks are detected as well. SCORE utilizes simulation-based optimization in order to find solutions of near-optimal improvements to the system, ridding the system of crucial bottlenecks (Bernedixen et al., 2015). These bottlenecks are the machines most sensitive to change, with respect to performance gain. The optimization problem in SCORE is defined by two objectives; maximize a key performance measure (e.g. throughput) and minimize the number of improvements needed to eliminate corresponding bottlenecks. This is described in its general form by Equation 2 and 3 (Ng et al., 2014):

\[ \text{Minimize/Maximize } f_m(x), \quad m = 1, 2, ..., M \]  
\[ \text{Subject to } g_j(x) \geq 0, \quad h_k(x) = 0 \]  
\[ j = 1, 2, ..., J \quad k = 1, 2, ..., K \]  
\[ x = (x_1, x_2, ..., x_n) \]  
\[ \text{where } x_i^L \leq x_i \leq x_i^U \text{ and } i = 1, 2, ..., n \]

As explained by Ng et al. (2014), Equation 2 represents the multiple objectives of the optimization, where x is the solution vector consisting of up to n variables. Equation 3 presents the requirements for any feasible solution: the inequality constraint \( g_j(x) \) and equality constraint \( h_k(x) \) must be satisfied (Ng et al., 2014). The lower and upper bounds of the variables are denoted as \( x_i^L \) and \( x_i^U \) respectively.

2.5 Flow production data

In order for a simulation study to be successful, it is crucial that the problem to be investigated is well-defined before the model is built. The required detail of the model depends on the questions to be answered in the study, which is stated by the problem (Law, 2015). Based on the defined problem, the model can be constructed from appropriate assumptions and characteristics of the true system and necessary system data can be collected accordingly. To simulate the flow in a production line as accurately as possible, the model equivalent of a machine must be updated with data collected from corresponding machines of the real production line. The data that is required for a particular simulation model depends on the detail of the model. For simulation studies of manufacturing systems, it is often necessary to collect data describing the state of each machine in the production
line at any given time. The below listed data types are some examples of useful data in the simulation study of a production line:

- cycle time
- availability
- repair time
- time between failures

In this context, the cycle time of a machine is considered to be the total time to complete one repetition of an intended task for a particular machine and part. One cycle includes the processing time, load and unload times and the move time of the processed part. This definition is based on how the cycle time is defined by the manufacturing monitoring system used to collect data for this study. The availability of a machine is the proportion of time that the machine is in fully operational state at any given point of time (Misra, 2008). The average availability can hence be calculated using the mean time between failures (MTBF) and the mean time to repair (MTTR) as shown in Equation 4:

\[
\text{Availability} = \frac{\text{up time}}{\text{total time}} = \frac{\text{MTBF} - \text{MTTR}}{\text{MTBF}}
\]  

The repair time is the total time of duration for a particular failure, and includes time for localizing and investigating cause of failure and the time for reparation of the identified failure, until it has been completely addressed and the machine brought to fully operational state (Kumar, 2008). The time between failures is simply the time between the initiations of two consecutive failures of the same failure type. Besides the different data concerning the machines of the production line, it can also be of interest to note the capacity of each buffer in the production line if any. In a simulation study, great care should be taken in collecting reliable and appropriate data to guarantee that the model is given the possibility to simulate the real system as accurate as possible. If the model is based on an existing system, data should be collected from the existing system itself. For an existing system, it is possible that these types of data can be collected from an automated logging system monitoring the production line.

2.6 The importance of data quality

In previous work done on discrete-event simulation, the importance of high quality input data for reliable and accurate simulation results is strongly underlined and the data collection process is stated as a crucial part of all sound simulation studies (Law, 2015; Banks et.al., 2014; Skoogh & Johansson, 2008). The quality of collected input data is important for the validity of the simulation results (Law, 2015). If the validity of the simulation results cannot be guaranteed, then it is inappropriate to use the results to draw any conclusions about the true system. Even the best of models will produce invalid results if the input data is of poor quality (Law, 2015). Input data of high quality is desired, but quality is a relative term. The appropriate level of quality depends on several aspects of the current simulation study (Skoogh & Johansson, 2008). System processes with a high degree of variability require data samples of higher quality to guarantee reliable simulation results. The significance of each input parameter determines the necessary quality in corresponding data samples to assure that the significance is accurately represented in the model. The model detail also has a key role in the necessary quality of the collected data. An exhaustive
model of great detail requires input data of higher quality than a simple model. In a detailed model, each input parameter will be more critical for the validity of the result. Based on the requirements for simulation model input data presented by Law (2015) and for data samples in general by Murthy et.al. (2004), the following properties have been identified as crucial for good quality data:

- the amount of data samples is sufficient
- the data samples are representative of the process
- the data samples have been collected during a representative time period
- the data have been collected with sufficient accuracy
- the data samples have not been significantly affected by observational errors

In reality however, it is a common problem that available data samples acquired from a system do not exhibit the desired quality that is necessary for the simulation study. It is important to identify possible problems with the collected data before it is used in the simulation and producing invalid results. The process of collecting data for a simulation model is often difficult due to several reasons and affects the quality of the data acquired. Some of the issues concern difficulties collecting data that is representative to the normal state of the system. Other problems are related to measurement errors causing biased data and errors caused by recording insufficiencies resulting in inaccurate or missing data (Law, 2015). The quality of the data should be validated to guarantee that the data is reliable.

### 2.7 The effect of variability

According to Law (2015), it is dangerous to disregard the variability of a system by replacing probability distributions by its perceived mean value. Without properly modeled variability, the model can fail to capture the delays occurring in the true system (Law, 2015). If there is significant variability in the production system, it should be captured in the model. The significance of the variability can be determined by examining some basic descriptive statistics. If the model is simplified by using the sample mean of each data set when there is significant variability, it can affect the accuracy of the model and the reliability of the simulation results (Law, 2015). To capture the variability of the system in the model, suitable probability distributions are fitted to the collected data, which then are used in the model. To achieve a good representation of the variability in the model, it is important to collect a sufficient amount of representative data samples (Skoogh & Johansson, 2008). To illustrate the effect of variability in a production system, consider the simple production line depicted in Figure 3:

![Figure 3](image)

**Figure 3** A simple transfer line.
In this example, the cycle time of each machine, M1 and M2, is 20 minutes. Without any variability in the system, this production line constitutes the perfect transfer line where the machines are fully synchronized. As a result, the throughput should be 3 products per hour and the lead time should be 2 400 seconds. If variability is introduced to the system, where the cycle time of each machine is distributed according to a normal distribution of mean 20 minutes and variance 10 minutes; how will it affect the production line? The pitfall here is to disregard the variance, despite the indication that the cycle time of the machines can vary greatly from the expected value. From simulation results using FACTS Analyzer (Ng et.al., 2007) it is noted that the throughput has decreased to 2.8 products per hour, which is only a 10 % loss. However, when examining the new lead time of the production line, it is noted that it has increased by over 110 % to 5 040 seconds! The corrupting influence of variability has no apparent effect on the throughput, but is devastating for the lead time. In reality, almost every production line have some degree of variability in its internal processes (Law, 2011 a), causing different types of disturbances and delays in the flow of production.

2.8 Input probability distributions

The validity of a simulation model is strongly dependent on the quality of its input data (Law, 2015). The collected input data must accurately reflect the characteristics of the original manufacturing system, including any system randomness. Many existing production systems in the real world are to some extent affected by randomness in its internal process or from external sources (Law, 2011 a). If the original system is expressing significant variability due to sources of randomness, it is essential that this variability is represented by appropriate probability distributions in the model to assure reliable results from the simulation (Law, 2011 a). The probability distributions will be used in the simulation to generate random numbers for corresponding input variables in the model. To model system variability with sufficient accuracy through probability distributions, it is crucial that collected data samples are of high quality. The acquired data samples should capture the true randomness as accurate as possible, or it will be difficult to find a suitable probability distribution providing a good estimation of the variability in the system. Note however, that distribution models that fit the data always can be found (Law, 2011 a). The question is rather whether the distribution model is actually a valid representation of the true randomness. Using incorrect probability distribution to model the variability for a particular component in the system can give as erroneous simulation results as using the sample mean for the same data set. As long as appropriate data is available, there are two different methods frequently used when modeling variability (Law, 2011 a). The variability can be modeled by fitting a theoretical probability distribution to the data or it can be modeled by an empirical distribution (Law, 2011 a). To fit a theoretical distribution to the data, the collected data samples have to be independent and identically distributed.

2.8.1 Empirical modeling

Modeling variability by using probability distributions models is also known as empirical modeling. In empirical modeling a mathematical model is built from available data samples, and is useful when the underlying mechanisms are unknown (Murthy et.al., 2004). Such models are data dependent and needs sufficient amount of data samples to capture the randomness of the system. Empirical modeling is a method consisting of five major steps as described by Murthy et.al. (2004):
1. collect data
2. data analysis
3. model selection
4. parameter estimation
5. model validation

2.8.2 Data analysis and model selection
A preliminary data analysis should always be made before a particular distribution function is selected to model a certain data set (Law, 2011 a). The preliminary data analysis consists of various calculated sample statistics providing valuable information about the characteristics of the data. From the presented descriptive statistics, it can be decided if probabilistic and stochastic models even are required to model the data (Murthy et.al., 2004). If the data set does not have significant variability, it is sufficient to model the data by its sample mean instead of using a probability distribution (Murthy et.al., 2004). If the data set however contains significant variability, this must be captured in the model. The variability is considered significant if the range is large relative to the sample mean (Murthy et.al., 2004). The gathered information from the preliminary analysis describes the probability density function of the data, which is useful when deciding appropriate distribution family (Law, 2011 a). In addition to descriptive statistics, a histogram can be used in the preliminary analysis to graphically estimate the shape of the underlying density function (Law, 2011 a). To construct a histogram for a set of data, the interval width to use must be determined. This is nontrivial, but there exist various distribution fitting software which can aid in this process. The approximation of the shape should be smooth, which can be achieved if the selected interval width is adequately small (Law, 2011 a). Based on the gathered information from the analysis, the suitable family of distribution models can be selected to represent the data set.

2.8.3 Parameter estimation
After deciding what family of distribution models that is appropriate, the corresponding model parameters of the density function must be estimated (Law, 2011 a). There exist several different techniques for parameter estimation. Some of them are implemented in various distribution fitting software (Law, 2011 a). The accuracy of the estimated parameters is determined by the number of data samples available and the method used for estimating (Murthy et.al., 2004). For example, using graphical procedures will result in rough estimates, while using analytical methods will give more accurate estimates.

2.8.4 Model validation
Model validation is the final step in empirical modeling. Validating the model is essential to determine whether the selected distribution model, and its estimated parameters, provides a good representation of the data and its varying characteristics. The selected distribution model is valid if it is representative of the underlying distribution of the collected data (Law, 2011 a). The reasons behind a poor model fit are either of graphical or analytical nature (Murthy et.al., 2004):

- wrong distribution model selected
- correct model, but inaccurate parameter estimations
However, the model should not be more complex than for its intended purpose and its validity should be determined with this in mind. There is no such thing as a “perfect fit” (Law, 2011a). Model validation can be divided in two categories (Murthy et al., 2004): graphical methods and goodness-of-fit tests. Graphical procedures are methods in which the appropriateness of a certain model is determined subjectively by visualizing the properties of the fit. One such method is the density-histogram plot. In a density-histogram plot, the shape of the approximated density function can be visually compared with the histogram of the data samples (Law, 2011a). Graphical procedures have the advantage that they are intuitive and are easy to use, and are therefore a good starting point. When validating a model it is however necessary to use analytical and statistical methods as well, such as goodness-of-fit tests, to properly analyze the adequacy of the fit (Murthy et al., 2004).

### 2.8.5 Input data modeling problems

A potential problem when fitting a theoretical distribution to a particular data set is that sometimes it is not possible to find a distribution that provides a sufficiently accurate representation of the data (Law, 2011a). When fitting a distribution to certain data, the properties of the data should be known and understood. The difficulty of finding a fitting distribution can be due to the data set actually consisting of two or more heterogeneous populations (Law, 2011a). As an example, consider a machine in a production line. This machine can have two different types of failures: systematic failures and fatal failures. The systematic failures are distributed according to a normal distribution and the fatal failures belong to an exponential distribution. The collected data from the machine consist of the duration of each occurring failure, i.e. the repair time. The collected data has been fitted to a beta distribution as shown in Figure 4:

![Density-Histogram Plot](image)

**Figure 4** Beta distribution modeling failure data from two heterogeneous populations. The model is considered to be a bad fit.
The beta model is evaluated as a bad distribution fit. Using a density-histogram plot to graphically estimate the distribution fit, it is revealed that the data set actually consist of data samples from at least two different populations. To provide a more accurate model, the data should be separated by the type of failure and then fitted to appropriate distributions. Several input models can be therefore be necessary to model one aspect of a machine in the system.

2.8.6 Useful probability distributions as seen in FACTS
The collected input data of this thesis have been modeled by fitting probability distributions to each data set. The well-known normal-distribution is frequently used in literature to model input data of simulation studies (Law, 2015). In reality however, the normal distribution is seldom appropriate (Law, 2011 a). The distributions families frequently occurring when fitting distributions to the data sets collected from the production system used in this thesis, were lognormal, Weibull and exponential. These three distribution models and their parameters in FACTS are briefly described in Table 1:

<table>
<thead>
<tr>
<th>Probability distribution</th>
<th>Parameters in FACTS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>Mean: $e^{\mu + \sigma^2/2}$ Sigma: $\sigma$</td>
<td>The lognormal distribution can be used to model the time to complete a process (Law, 2015).</td>
</tr>
<tr>
<td>Weibull</td>
<td>Scale: $\lambda$ Shape: $k$</td>
<td>The Weibull distribution is good at modeling complex data sets and is frequently used for modeling failure in components (Murthy et.al., 2004). It is a highly flexible distribution.</td>
</tr>
<tr>
<td>Exponential</td>
<td>Mean: $\beta$</td>
<td>Applications of the exponential distribution include modeling time between independent events, and can therefore be used to model failures. The exponential distribution is a special case of the Weibull distribution (Banks et.al., 2014).</td>
</tr>
</tbody>
</table>
3 Input data management

The purpose of this chapter is to present any relevant details of the necessary preparations for the planned experiments of this thesis. These preparations include modeling the production system, collecting the required data from the actual production system and processing the collected raw data to produce suitable input data for the model in the form of sample mean and fitted probability distributions.

The majority of the work done in this thesis relates to the concept of input data management. Input data management is a term used by Skoogh and Johansson (2008) to describe the complete process of obtaining the final processed input data to be used in the model of the simulation study. This process begins by retrieving necessary data from the actual system. Once the relevant data has been gathered, the collected data samples are analyzed using elementary descriptive statistics providing valuable information about the characteristics of the data. The result from the analysis is used to accurately model the data sets and to produce suitable input parameters for the model. To determine the accuracy of the modeled data sets, the model and its input parameters must be validated as the final step of input data management (Skoogh & Johansson, 2008).

3.1 Simulation models and model generation

FACTS Analyzer (Ng et.al., 2007) has been used to model an existing production system from a real-world automotive manufacturer. FACTS is a toolset, developed at the University of Skövde, designed for analyzing and providing decision support for production systems by realizing model aggregation and simulation-based optimization in a user-friendly manner (Ng et.al., 2007; Moris et.al., 2008). The modeling process in FACTS is aided by an object library of components representing entities within the manufacturing domain, e.g. machines, buffers and other entities. The user can drag and drop the objects into a canvas and then link them together to form the production flow of the system. The motivation behind using FACTS lies in its user-friendly designed interface and support for aggregated modeling. Aggregated modeling is a method for reducing the complexity of a simulation model while maintaining the validity as a representation of the true system (Pehrsson et.al., 2014). Complexity is undesirable since it must be possible to build the model within a reasonable amount of time and to run the completed model within a practical execution time (Pehrsson et.al., 2014). The complexity of a model is strongly associated with the level of detail present in the model. Aggregated modeling can reduce complexity by using abstract objects to represent complex entities in the true system (Moris et.al., 2008). The appropriate abstraction level depends on the perspective the model must capture from the true system, but to make these simplifications requires certain knowledge (Moris et.al., 2008). These are the reasons why FACTS have been selected to model the system. The system can be modeled using a higher abstraction level and still being represented with an appropriate level of accuracy, since the goal is to investigate how the models differ when different input data models are used.

The system to be modeled is a production line provided from an automotive manufacturer. This system will be referred to as production system A. Production system A has been modeled using a concept referred to as model generation. Model generation has been implemented in a supporting software tool adapted to FACTS developed during the work of this thesis. Data from the system has been collected from a manufacturing monitoring
system connected to the production line. This data can be used to extract information for generating a rough estimate of the layout of the production line and its machines. The identification name of each machine is present in the collected data in the same order as they appear in the flow of the production line, assuming the flow does not contain alternate routes. The identification names can also be parsed by the support tool for information regarding if the machine is serial or parallelized with other machines. Using this information, it is possible to generate an initial model that can be improved later on by knowledge that cannot be obtained by the data alone. This model generation process is useful in the early stage of modeling and saves time and effort of placing all the machine objects on the canvas in FACTS. The method has been especially suitable for the modeling of system A, since it is a straightforward production line without any complex logic. Additional details have been added to the model by consulting domain experts of the line, such as buffer sizes and their locations. The finished FACTS model of production system A, containing 37 different automated machines, is shown in Figure 5. The names used in the model do not reflect the actual names used internally by the manufacturer owning the production line.

![Diagram of production system A](image)

**Figure 5** Model of production system A

### 3.2 Data collection

In order to conduct the experiments of this thesis using SCORE, the required data to accurately simulate the performance of each system in different conditions must be collected from the real production system and updated to the corresponding simulation model in the accepted format of the simulation software. Data collection is mentioned by Banks et.al. (2014) as one of the most crucial and challenging processes in a simulation study due to its importance for accurate results. If data from the production line is available directly from a manual or automatic manufacturing monitoring system, the remaining process concerns analyzing the data, producing suitable input parameters and validation (Skoogh & Johansson, 2008). In this case, the necessary data have been possible to acquire from a monitoring system connected to the production line, or been able to calculate using the available data. The role of the manufacturing monitoring system in the data collection process is illustrated in Figure 6.
The data directly available from the manufacturing monitoring system is following machine specific properties:

- identification name
- cycle time
- timestamp for occurred failures
- duration of each failure (i.e. repair time)
- MTTR (mean for the specified time period of the export)
- MTBF (mean for the specified time period of the export)

The monitoring system is capable of automatically registering the cycle times of each machine continuously and logging the majority of the occurring failures. For each occurring failure, the monitoring system logs the current time when the failure is detected as a timestamp and the duration of the failure. The failures not being automatically managed by the monitoring system needs to be manually updated to the system by an operator of the production line. This manual step is a possible source of error with respect to data reliability. Buffer capacities are not available from the monitoring system, but have been collected from domain experts of the corresponding production line. The data can be retrieved from the monitoring system as an export in the format of Excel sheets. Each data entry is neatly presented in columns of its data type and ordered in rows by the identification name of the corresponding machine. The raw data collected from the monitoring system consists of data samples logged over a particular period of time. It is however a time-consuming process to export data over long periods of time due to the large data sets created from the production system. The data used in this thesis have been collected from a time period of a week regarding process times, and a complete month regarding failure data. This export contained a large collection of data samples to process and analyze.
3.3 Input data modeling

The main process of input data management is the input data modeling phase. An overview of how input data modeling fits into input data management and how it has been managed in this thesis is shown in Figure 7.

![Figure 7 Overview of the input data management process](image)

The exported data from the manufacturing monitoring system were analyzed using elementary descriptive statistics supplied from the distribution fitting software ExpertFit (Law, 2011 b). The analysis of each data set provided valuable information about the characteristics of the data; information which later can be used to validate the fitted distribution models. The data have been modeled by using sample mean combined with default distributions and by fitted probability distributions. Two different distribution fitting software have been used to model the data sets: ExpertFit (Law, 2011 b) and GDM Tool (Skoogh, 2009). The input models based on sample mean have been created using the FACTS support tool (previously used to parse the exported data and generate the simulation model for production system A), which calculated sample mean from the raw data. All of these input models were then updated to the corresponding simulation model in FACTS using the FACTS support tool.

3.3.1 Data analysis

Before any data sets are modeled, a preliminary statistical analysis should be made in order to fully understand the essential properties of the data (Banks et.al., 2014). This is in order to guarantee input model validity. The validity of an input model is not automatically assured just by using some probability distribution fitting software, since it is always possible to find a fitting probability distribution; it might just not be that good of a fit (Banks et.al., 2014). The characteristics of each data set in the collected data have therefore been analyzed in order to create accurate input models. This concerns both types of input models: input models based on sample mean and probability distributions. The sample statistics provides valuable information regarding important data properties that should be captured by the input model for it to become a valid representation.

3.3.2 Summary statistics using ExpertFit

Each collected data set has been analyzed using the ExpertFit software (Law, 2011 b) and its summary statistics option. The summary statistics used are presented in Table 2:
The calculated sample statistics describe several properties of the collected data. The relation between the sample mean and the median can expose information about the symmetry in the data set. If the estimates of the mean and median for a particular data set are close to equal, it might be an indication of a symmetric data set (Law, 2011 a). Skewness is another important measure of symmetry and can be interpreted as the degree of asymmetry in a data set. When the data is perfectly symmetric, the skewness is 0 (Law, 2011 a). The most important characteristic about each data set is the expressed variability. The measure of variability is important since this information is valuable when the input model is validated and when the results from the experiments are analyzed. If the variability in a particular data set is high, then it might provide an explanation behind inaccuracy problems with the input models. Modeling a data set by its sample mean and default distributions can be inappropriate for such a data set, but also when fitting a certain probability distribution since the high variability might be an indication of data samples from two or more heterogeneous distributions. Every manufacturing system has some degree of process variability. If there is no significant variability, it is sufficient to model the data set by its sample mean (Murthy et al., 2004). If the preliminary analysis of a particular data set however displays significant variability, it is crucial for the validity of the simulation results that this is characteristic is modeled by an appropriate probability distribution.

### Graphical estimates in ExpertFit

The calculated sample statistics have also been complemented with graphical estimates of the shape of the underlying probability distribution, since the true characteristics of the data cannot be described by sample statistics alone. For example, the common rule of thumb regarding the relation between mean and median, where it is stated that the mean is more sensitive to skewness and hence located beyond the median in the long tail, is actually frequently inaccurate (Hippel, 2005). Some examples provided by Hippel (2005) are when...
dealing with multimodal distributions and distributions with one heavy tail and one long tail. Examining an estimate of the shape of the underlying distribution illuminates such erroneous assumptions about the data and is therefore an important complement used in this work. Histograms created in ExpertFit have been used to estimate the shape of underlying distributions of each data set.

3.3.4 Model input parameters

Due to the particular types of data registered by the manufacturing monitoring system and its format, it has been necessary to process the data to produce suitable input parameters corresponding to the model of the simulation study. This processing has been done using a software support tool for FACTS developed during this work and the GDM Tool (Skoogh, 2009) developed at the Chalmers University in Gothenburg.

The machine entities of the simulation model in FACTS have attributes for process time, availability and mean time to repair (MTTR). These attributes are determined as the minimum amount of information necessary for simulating the approximated behavior of the true production system being represented. The data export from the manufacturing monitoring system does not contain any data of the availability of the machines, but it is possible to calculate the availability for a particular machine using MTBF and MTTR according to Equation 4 presented earlier in section 2.5. The disturbances in the system can also be modeled by probability distributions by using the attributes interval (time between failures) and duration instead of availability and MTTR. If probability distributions should be used to model the input data, the relevant input parameters for the particular distribution are needed for each machine property. Suitable probability distributions are found by fitting distributions to the provided data sets of raw data samples. The manufacturing monitoring system however, provides only raw data for cycle time and repair time. But since the monitoring system registers the timestamp for when each failure occurs, the time between each consecutive failure can be calculated and used when fitting a probability distribution for modeling the interval of failures. This has been aided by the GDM tool software.

3.4 Probability distribution models

The process of modeling input data by probability distributions has been an activity within input data management and consisted of several steps: model selection, parameter estimation and model validation. This process has been aided by using distribution fitting software for automatically fitting probability distributions to each of the collected data sets and validating the fitted models. It would have been impractical to accomplish this process manually due to the vast number of collected data sets. The selected distribution fitting software used in this thesis has been ExpertFit (Law, 2011 b) and GDM Tool (Skoogh, 2009).

3.4.1 ExpertFit

The ExpertFit software have been used to analyze each data set using basic descriptive statistics and then both ExpertFit and GDM Tool have been used to produce corresponding input parameters of fitted distributions that can be updated to the model in FACTS. The ExpertFit software will fit several distributions to the applied data set, and then rank the fitted models based on some criteria determining the quality of the fitted model. It is then up to the user to select one of the proposed distribution models, guided by the available information about each model. ExpertFit has been selected since the designers behind it are recognized for their experience of simulation and input data modeling.
3.4.2 GDM Tool

The use of GDM Tool is motivated by its concept of general data management which means it is not dependent on a specific input data format (Skoogh, 2009). This advantage of GDM Tool allows the user to process the input into a suitable format. As stated earlier, the manufacturing monitoring system does not register the time between failures explicitly; only the timestamp for when a failure of a machine occurs. Taking advantage of the strength of GDM Tool, this information can be used to calculate time between failures by subtracting rows of consecutive timestamps with each other in the input data and create a new column containing the result. After processing the data into a suitable format, the GDM Tool can be used to fit distributions to selected data sets. GDM Tool will then present the user with the fitted distributions it considers to be the absolute best fit for each data set.

3.4.3 Distribution fitting

The data sets have been fed to each of the distribution fitting software for input data modeling. Using two different distribution fitting software is motivated as a validation step of the generated probability distributions. The use of distribution fitting software to automatically fit distributions to data and find the “best” fit is no guarantee that the actually most suitable fit for a particular data set is found (Gupta & Parzen, 2004). Gupta and Parzen (2004) discuss several issues related to ranking fitted distribution models using goodness-of-fit tests. Ranking different fitted distribution models by using p-values can be problematic due to small differences (which can be caused by randomness in the data); the ranking can therefore be questionable (Gupta & Parzen, 2004). Another related issue is that the highest ranked model often is over fitted (Gupta & Parzen, 2004), meaning the model cannot handle data samples not originating from the original data set. This opposes the entire purpose of using the model as a representation of the reality. The questionable ranking leaves the analyst to consult the preliminary analysis of the data to decide if the number one ranked model is the best for the specific problem. This is why the analysis using descriptive statistics and graphical estimates is so important in this work.

3.5 Updating the models

The parameters of the validated input data models have been updated to corresponding FACTS simulation models using the previously discussed FACTS support tool. Updating the input parameters to every machine entity in the simulation models would be both error-prone and unnecessary time-consuming. The production system referred to as system A have several representing simulation models. The simulation models only differ in how the input data of the machines are modeled and are exactly the same in all other aspects. All of these models need to be updated with input data from the real production system.

3.5.1 FACTS support tool

The FACTS support tool is capable of updating a FACTS model with input data from Excel sheets (since the manufacturing monitoring system exports data to Excel). A FACTS model is saved as an xml-file, which makes them easy to manage and update with new data. The input data does not have to be in a specific format for the software to recognize the data; instead the user can freely map each column in the Excel file of a particular input data type to the corresponding data types in FACTS. The data entries will then be interpreted as specified by this mapping. An overview of the support tool is presented in Figure 8.
The specific FACTS model file to be updated is loaded in the support tool together with corresponding input data files in Excel. The support tool will then parse the available data based on how it is configured to interpret the input data file with respect to FACTS. The support tool will automatically try to find data columns containing process times, repair times and time between failures. If the software cannot find appropriate data or if it makes the wrong assumptions about the data, the user can modify this mapping.

When the data has been parsed, each of the machines present in the input data file will automatically be mapped to a suitable candidate in the model file. The mapping is based on the resemblance between names used in the input and model files determined by the edit distance measure (Wagner & Fischer, 1974). The associated input data will then be updated to the mapped model entity accordingly. If the input data contains parameters from fitted distribution models, these will be updated to the model in the accepted format in FACTS. But if the input data contains raw data, the sample mean for each data set will be calculated and then updated to the model. ExpertFit and FACTS do not use the same parameters for describing some of the available distribution models. This problem has been managed by the FACTS support tool, which automatically translated the input parameters from ExpertFit to corresponding parameters in FACTS. Concerning the lognormal models, ExpertFit uses the parameters shape $\sigma$ and scale $e^\mu$, while FACTS uses the mean value instead of the scale parameter. The mean value can be calculated from the scale parameter as seen in Table 1. This calculation has automatically been managed by the support tool.
4 Experiments

This chapter describes the approach of the conducted experiments in FACTS and their outcome. The experiments have consisted of simulations and optimizations using SCORE. Several different simulation models representing production system A have been used in the experiments, distinguished by the different input data models used. The purpose of the experiments has been to validate the models and examine how the optimization result in SCORE is affected by the different input models used in the simulation model with respect to convergence and diversity of the Pareto fronts.

4.1 Simulation experiments

The simulation experiments have been carried out using FACTS Analyzer (Ng et.al., 2007). The purpose of the simulation experiments has been to validate the simulation models against the true system, including the input data models used for the machines. It is essential that the simulation model is validated against the true system which it represents if the result should be used to make any assumptions or decisions affecting the true system. If the model is not validated, the reliability and accuracy of the simulation results should be questioned (Banks et.al., 2014). It is however both time-consuming and difficult to apply every recommended validation technique (Banks et.al., 2014), and hence for this thesis it has been deemed sufficient to simply compare simulation output with the actual output from the corresponding production system, since the aim of this thesis is to investigate optimization sensitivity to input data models.

4.1.1 Simulation models

Six different simulation models have been used to represent each production system, distinguished by the particular input data models used to model the machines. Initially there were only three simulation models:

- Model A1: sample mean and FACTS default distributions
- Model A2: distributions fitted by ExpertFit
- Model A3: distributions fitted by GDM Tool

Based on the poor simulation result from Model A3, the decision to create an additional model was made; a simulation model using distributions from GDM Tool, except to model the process times where sample mean will be used instead. The motivation behind this decision will be explained in the coming chapter. Concerns regarding how the parameters of the Weibull models were translated to optimization input motivated the creation of another model; Model A5. It is based on Model A2, but the Weibull models have been replaced by lognormal and exponential distributions. Due to unsatisfactory Pareto fronts from the optimization of both Model A2 and A5; a final model was created based on Model A5, but using sample mean for process times instead. These additional models are denoted as:

- Model A4: based on Model A3, but using sample mean for process times
- Model A5: based on Model A2, but replacing Weibull models with lognormal and exponential distributions
- Model A6: based on Model A5, but using sample mean for process times
4.1.2 Simulating manufacturing using FACTS

Simulating manufacturing systems using FACTS is a straightforward process, and could be initiated once the simulation models had been updated with corresponding input data models and their respective parameters. The simulation can be configured by the following available simulation settings in FACTS: start time, simulation horizon, warm-up time and the number of simulation replications. The number of replications is particularly important due to the simulation model being stochastic, since the input data models used are based on random variables. In order to obtain an accurate estimate of the true system output mean, several simulation runs is required through replications (Hoad et.al., 2007). For each replication, the simulation is rerun with altered random number streams to produce different results (Hoad et.al., 2007). Each replication will produce different results independently of each other (Law, 2015). The replication setting controls how many replications the final result will be based on and should be of a sufficient number assuring an accurate estimate within reasonable execution time (Hoad et.al., 2007). The recommendation is to use 3 to 5 replications as a rule of thumb (Hoad et.al., 2007). The result from these experiments is based on the mean value of 5 replications, since it is the default setting in FACTS and within the recommendations.

4.2 Optimization and setting up SCORE

The main experiments of this thesis have been the SCORE optimizations. The optimization is based on the genetic algorithm NSGA-II (Deb et.al., 2002), and each optimization has consisted of 15,000 evaluations. The primary objective of the optimization has been to maximize throughput by finding the necessary improvements to the system which removes the critical constraints. The secondary objective has been to minimize the number of required changes to accomplish the maximum improvement in performance of the system. This will force the optimization to find the bottlenecks in the system and the solutions to how they should be removed. A change refers to an improvement to any aspect (process time etc.) of an entity in the system and is defined by a set containing the original value and the improved value as determined by a percentage of the original value. The set offers the optimization a choice of using a machine as it is or improving a certain aspect of it. The effect of a particular choice is then evaluated during the optimization. This optimization problem is defined as seen in Bernedixen et.al. (2015):

\[
\begin{align*}
\text{Maximize:} & \quad \text{Throughput} \\
\text{Minimize:} & \quad \sum_{i=1}^{n} l_i \\
\text{Subject to:} & \quad x_i \in \{ \text{original\_value}, \text{improved\_value} \} \\
& \quad l_i \in \{ 1, 0 \} \text{ where } \begin{cases} 
\quad l_i = 0 & \text{if } x_i = \text{original\_value} \\
\quad l_i = 1 & \text{if } x_i = \text{improved\_value} 
\end{cases} \\
& \quad i \in \{ 1, 2, \ldots, n \}
\end{align*}
\]

By minimizing the number of changes in Equation 6, the optimization strives to find the bottlenecks having the greatest impact on the system, meaning that the input data models, which ultimately model the behavior of the machines, will presumably have greater effect on the optimization result the more changes required. SCORE has therefore been used in order to examine how different input data models affect the optimization result in comparison; the
optimization sensitivity to different input data models with respect to convergence and diversity of the Pareto front. To accomplish this, the optimization needs input parameters, defined objectives and possibly constraints determining which solutions are feasible solutions. All of these aspects will be explained in coming sections. FACTS Analyzer has built-in support for configuring SCORE optimizations (Bernedixen et.al., 2015). This option has been used for the simulation model with sample mean and underlying default distributions to model input data. Setting up SCORE for probability distribution models is however not as straightforward, as will be explained further on, and has therefore been automatized in the FACTS support tool.

4.2.1 Optimization input parameters
The optimization needs input parameters representing the properties of the machines which are possible to change in order to improve the performance of the production system. One input parameter for each machine property is needed. The input parameters define a set representing an improvement from the original value to an improved value by some percentage of the original value. The selected improvement factor in this work has been 10 %. When modeling input data by sample mean, the configuration of input parameters can be managed automatically by FACTS (Bernedixen et.al., 2015). However, for the simulation models using only fitted distributions, creating the input parameters for the optimization is not as straightforward and cannot be automatically set up by FACTS. Instead, the FACTS support tool has been used to aid in this step. The improvement sets depends on the particular parameters of the distribution model. To represent an improvement, the distribution model mean is calculated and the improvement is applied to the mean value. Using the improved mean value, the new corresponding input parameters were estimated. This method is appropriate for distributions described by one single parameter in FACTS, or as in the case of the lognormal models; described by the mean value (the variance is not improved). The method is however not suitable for the Weibull models, since those are defined by two parameters: shape k and scale \( \lambda \) (refer to Table 1).

4.2.2 Estimating Weibull parameters
The improvement factor has been applied to the mean value of the Weibull model. The original mean value is calculated by Equation 9 (Rocha et.al., 2012), and is improved by applying a 10 % improvement factor (increasing or decreasing the original value). Translating from the improved mean value and back to corresponding parameters shape k and scale \( \lambda \) has been managed by Equation 8 and 9. This method of estimating Weibull parameters based on the mean value and variance is presented in the work of Rocha et.al. (2012). The equation is an empirical method and a special case of the moment method.

\[
k = \left( \frac{\bar{x}}{\lambda} \right)^{-1.086} \tag{8}
\]

\[
\bar{x} = \lambda \Gamma(1 + \frac{1}{k}) \tag{9}
\]

The empirical method has proven suitable for estimating the parameters for Weibull models describing wind speed (Rocha et.al., 2012). Its applicability for data sets regarding automated machines in production systems is uncertain. Therefore, an additional model were created, Model E, where the Weibull models were replaced by either exponential models or lognormal models depending on their quality of fit.
4.2.3 FACTS variables

Every machine property must be represented by only one optimization input to reflect one single change of the respective property. This is problematic for the Weibull models, since two parameters are changed as an effect of the improvement. The optimization must be forced to improve both input parameters at once and not independently of each other. To solve this issue, the concept of variables has been used in FACTS. A variable in FACTS is highly flexible and can be assigned to any machine property as it is or in a formula and then used as input in the optimization. The shape parameters of each Weibull model have been assigned a corresponding variable instead, responsible for its value. The shape variables have then been used to create input parameters for the optimization. To force the scale parameter to be affected respectively, it is defined in terms of the shape parameters as seen in Equation 10, using a ternary conditional operator:

\[
\text{variableShape} = \text{originalShape} ? \text{originalScale} : \text{improvedScale}
\]

If the shape parameter of a particular Weibull model has its original value, the respective scale parameter will use its original value. If the shape parameter has been improved the scale parameter will be assigned its corresponding improved value.

4.2.4 Optimization objectives

The objectives of the optimization is as previously mentioned maximizing throughput while minimizing the number of improvements required to increase the performance of the system accordingly. Throughput is defined by a formula internally in FACTS and can simply be
added as an objective. The number of improvements must however be counted for each machine property and then summated. To count the number of currently applied improvements, one output variable were added for each possible improvement. Since one improvement is considered to be the change from an original value to an improved value, the output variable can be 0 or 1 reflecting no improvement or that improvement has been applied as seen in Equation 11.

\[
\text{improvement} = \begin{cases} 
0 & \text{if } |\text{currentValue} - \text{originalValue}| \leq 1E - 10 \\
1 & \text{if } |\text{currentValue} - \text{originalValue}| > 1E - 10
\end{cases} \quad (11)
\]

These output variables are summed together as described in Equation 6 in section 4.2, to get the total number of improvements currently applied to the system. This summation is then added as an objective, where the number of improvements is minimized.
5 Analysis

The result from the simulation experiments and the optimizations using SCORE is presented and analyzed in this chapter in order to provide material for the conclusion in next chapter. The simulation result is briefly analyzed with the purpose to validate the simulation models and their respective input data models. The findings of the validation can provide useful information regarding the optimization result. The Pareto fronts from the optimizations are visually examined and compared in order to investigate the optimization sensitivity to particular input data models.

5.1 Validating the distribution models

The result from ExpertFit and GDM Tool differed in favored distribution families when fitting probability distributions to the data sets from the machines. The proportion that each distribution family constitutes for each property of the machines is summarized in the bar charts of Figure 10 and 11 for ExpertFit and GDM Tool respectively.

![Figure 10](image1.png) The proportion for each distribution family in Model A2

![Figure 11](image2.png) The proportion for each distribution family in Model A3
The distinction between ExpertFit and GDM Tool concerning preferable distribution families for particular aspects of the machines are curious. The ExpertFit software has ranked its lognormal models higher than other fitted models to describe the majority of the data sets. On the other hand, the GDM Tool software favors the Weibull distribution to model the properties of the machines. This difference is particularly noticeable in how the two distribution fitting software chooses to model the data sets of process times from the machines. ExpertFit ranks its lognormal models as the best fit for every data set, while GDM Tool considers its fitted Weibull models to be the best fit (with the exception of a few exponential models). Putting these differences aside, it is interesting that the same types of distribution families were frequently used to describe the machine aspects of the production line, regardless of the specific distribution fitting software used.

5.1.1 Absolute model evaluation in ExpertFit

The quality of fit for each fitted distribution model has been validated using ExpertFit. ExpertFit provides an absolute evaluation of the quality of a particular fitted distribution model (Law, 2011b). The absolute evaluation is a straightforward method which quickly offers an indication of the quality of fit. If the absolute evaluation is indeterminate, more sophisticated goodness-of-fit tests are recommended (Law, 2011b). The absolute evaluation in ExpertFit is defined by following measures explained in Table 3.

<table>
<thead>
<tr>
<th>Absolute evaluation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>The model provides an adequate fit and is good enough to be used in a simulation model.</td>
</tr>
<tr>
<td>Indeterminate</td>
<td>Additional statistics should be consulted to determine if the particular model can be used in the simulation.</td>
</tr>
<tr>
<td>Bad</td>
<td>It is strongly advised to use an empirical distribution to model the data instead of the fitted distribution model.</td>
</tr>
</tbody>
</table>

The models evaluated as bad or indeterminate fits, have been examined further by consulting graphical validation methods, such as density-histogram plots and distribution-function-differences plot available in ExpertFit (Law, 2011b). The result of the absolute evaluation is summarized in Figure 12 and 13 for ExpertFit and GDM Tool respectively. The bar charts summarize the absolute evaluation of each model fit in proportion to all other models for a particular machine property. Both of the two distribution fitting software had difficulties in finding good distribution models describing the process time of the machines. Every fitted distribution model has been evaluated as a bad fit, and ExpertFit recommends them being replaced by empirical distributions instead. To understand why the data sets for machine process times were particularly difficult to find appropriate distribution models for, the data analysis of those data sets were consulted. From the data analysis it is noted that the data sets containing process time has much greater positive skew than any other data set for
any of the other machine properties. This was investigated further by examining corresponding histogram plots. It was then discovered that many of the process time data sets contain data samples with extreme values; values of unreasonable long process times. These extreme data samples are few in proportion to the other data samples, which might be the problem when trying to fit a suitable distribution model to the data set. These data samples should have been treated as outliers since they are most likely to be garbage data, and should have been removed from the data set before fitting the distribution models. The data sets exhibit properties that are unreasonable for automated machines.

**Figure 12** Validation of ExpertFit distributions

**Figure 13** Validation GDM Tool distributions
5.2 Simulation model validation

The implications of the quality of fit for the distribution models are examined by simulating each of the simulation models. In Table 4 to 9 the result of the simulations of all simulation models described in section 4.1.1 is summarized by relative mean and the standard deviation. The result is based on the mean value from 5 simulation replications. The mean value is presented relative to the corresponding value of the true production system to avoid revealing the plant output of the existing production system. The true system value reference is 100.

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Simulation results for Model A1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plant output</td>
<td>Relative mean</td>
</tr>
<tr>
<td>Throughput</td>
<td>127</td>
</tr>
<tr>
<td>Lead time</td>
<td>88</td>
</tr>
<tr>
<td>WIP</td>
<td>110</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Simulation results for Model A2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plant output</td>
<td>Relative mean</td>
</tr>
<tr>
<td>Throughput</td>
<td>117</td>
</tr>
<tr>
<td>Lead time</td>
<td>65</td>
</tr>
<tr>
<td>WIP</td>
<td>75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6</th>
<th>Simulation results for Model A3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plant output</td>
<td>Relative mean</td>
</tr>
<tr>
<td>Throughput</td>
<td>0.02</td>
</tr>
<tr>
<td>Lead time</td>
<td>362</td>
</tr>
<tr>
<td>WIP</td>
<td>27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 7</th>
<th>Simulation results for Model A4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plant output</td>
<td>Relative mean</td>
</tr>
<tr>
<td>Throughput</td>
<td>134</td>
</tr>
<tr>
<td>Lead time</td>
<td>31</td>
</tr>
<tr>
<td>WIP</td>
<td>41</td>
</tr>
</tbody>
</table>
Table 8  Simulation results for Model A5

<table>
<thead>
<tr>
<th>Plant output</th>
<th>Relative mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput</td>
<td>118</td>
<td>0.73</td>
</tr>
<tr>
<td>Lead time</td>
<td>62</td>
<td>4426.72</td>
</tr>
<tr>
<td>WIP</td>
<td>71</td>
<td>13.59</td>
</tr>
</tbody>
</table>

Table 9  Simulation results for Model A6

<table>
<thead>
<tr>
<th>Plant output</th>
<th>Relative mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput</td>
<td>131</td>
<td>0.46</td>
</tr>
<tr>
<td>Lead time</td>
<td>68</td>
<td>2545</td>
</tr>
<tr>
<td>WIP</td>
<td>86</td>
<td>8.99</td>
</tr>
</tbody>
</table>

The output from the simulation consists of the performance measures throughput, lead time and work in process. The system throughput is the production rate per hour; lead time is the total time from initiation of production to delivery for a particular product; and work in process is the number of products in the system. All of these aspects are important when describing the output performance of a production system. It is noted that the overall performance of the true system is overestimated by the simulation models when the simulation results are compared with the actual performance of the corresponding production system. Model A1, A2, A4, A5 and A6 is overestimating the actual throughput of the system, while underestimating the true lead time of the system as a consequence. The only exception is model A3, which instead underestimates the true throughput (only 0.02 % of the actual throughput). This has disastrous consequences on both lead time and work in process. Model A3 is therefore considered as an invalid simulation model, and will not be used in the optimization. The poor performance of Model A3 is a strong indication that some of its input data models are invalid; presumably the models for the machine process times judging by the previous distribution model evaluation in ExpertFit. This motivated the creation of model A4. Model A4 is based on Model A3, but models process times by sample mean instead of using distributions, since the fitted distributions for process time had the worst quality of fit as evaluated by ExpertFit. Model A4 yielded better simulation results than the previous Model A3.

5.3 Optimization Pareto fronts

The obtained Pareto fronts from the optimizations are presented in the scatter plot diagrams of Figure 14 and 15. Note that the figures regarding throughput levels are presented as a percentage of the actual throughput of the existing system as not to reveal the characteristics of the real production system. The true system value reference is 100. The optimization has been configured with an improvement factor of 10 % for SCORE. The optimization result for model A1, A2, A4 and A6 are shown in Figure 14. These Pareto fronts are considered to be the most interesting to compare due to their clear visual differences. The simulation models related to ExpertFit are shown separately in Figure 15.
Figure 14 The Pareto fronts from SCORE of Model A1, A2, A4 and A5

Figure 15 Pareto front comparison between Model A2, A5 and A6
The distinct visual appearance of the Pareto fronts generated by the optimizations in SCORE indicates that the optimization had fundamentally different conditions when searching the optimal region of suitable solutions for each simulation model. The Pareto front generated during the optimization of Model A1 exhibits the desirable properties of a satisfactory front with respect to convergence and diversity. This front seems to be close to the true Pareto optimal front as determined by its shape, expressing the conflict between maximum throughput and minimum amount of corresponding improvements as suitable tradeoffs. It has a satisfactory coverage of the optimal region; even though the number of solutions decreases the more improvements are applied to the corresponding system. In comparison, the visual appearance of this particular Pareto front stands in contrast with the other Pareto fronts (refer to the Pareto fronts of Model A2 and A4 in Figure 14). The optimization seems to have difficulties in searching the optimal region regarding the simulation models using fitted probability distributions. These Pareto fronts consist of fewer found solutions, resulting in poor diversity. The effect is worsened by each improvement applied to the system represented in the model. The fronts also exhibit unsatisfactory convergence; their shapes reveal that the optimization has troubles finding near-optimal tradeoffs between maximizing throughput and minimizing required improvements. Instead, the optimization provides poor tradeoffs, where several improvements implemented in the system offers little in return regarding increase in performance. The Pareto front for Model A6 is however something of an exception in this comparison. This simulation model uses sample mean to model process times while using fitted distributions from ExpertFit to model the failures of the machines. Its corresponding Pareto front exhibits better convergence and diversity than the fronts from Model A2 and A4. The front resembles the Pareto front for Model A1, but since Model A6 overestimates the performance of the true system with higher percentage than any other simulation model, it also offers higher performance gain for each implemented improvement.

5.3.1 The effects of estimating Weibull parameters
Model A5 were created due to concerns regarding uncertainties of the quality of the estimation method for estimating new Weibull parameters based on the respective improved mean value in Model A2. Based on the visual appearance of the Pareto fronts from the optimization of these models, the optimization of Model A2 and A5 roughly produced the same near-optimal solutions. Replacing the Weibull models in Model A2 with lognormal and exponential distributions does not seem to have affected the conditions for the optimization in searching the optimal region. The estimation method and the special configuration of FACTS variables has worked surprisingly well for this problem.

5.3.2 Searching the optimal region
The extent of the difficulties for the optimization searching the optimal region, when optimizing the models using fitted distribution models, is illustrated in Figure 16 to 18 for Model A1, A4 and A5 respectively. Compared with how the optimization has searched the optimal region during the optimization of Model A1; Model A4 and Model A5 clearly have distinct gaps in the search area of this region. The optimization has been searching a smaller area of the optimal region during these optimizations, and has had problems focusing in desirable areas; the orange colored points illustrate where the optimization has been focused, while the blue points show solutions of less interest. This might explain why the respective Pareto front of the models using fitted probability distributions has such poor convergence and diversity; but it does not provide any explanation behind the actual cause of this behavior.
Figure 16 Searching the optimal region for Model A1

Figure 17 Searching the optimal region for Model A4

Figure 18 Searching the optimal region for Model A5
6 Conclusions

The purpose of this chapter is to summarize the findings of the analysis and discuss the results. The material from the analysis and the results from the experiments are discussed in order to answer the research question and arrive at some final conclusions regarding the aim of this thesis.

6.1 Summary of results

The SCORE optimizations of the different simulation models resulted in evidently distinguishable Pareto fronts, judging by their visual appearance in the scatter plot diagrams. The distinct Pareto fronts indicated the different conditions available for the optimization while optimizing the simulation models distinguished by their input data models. These Pareto fronts have been examined and compared with respect to convergence and diversity by visually investigating the fronts, due to their apparent differences. The Pareto front for Model A1 had desirable properties with respect to both convergence and diversity. The simulation models using fitted distributions did not generate as satisfactory Pareto fronts. The optimization seemed to have troubles with searching the optimal region concerning these models, since convergence and diversity were poor of their respective fronts in relation to that of Model A1. The problems with the distributions might be related with the models for the machine process times, since all of these models from both ExpertFit and GDM Tool were evaluated as bad distribution fits. Replacing them with sample mean generated better results for Model A6, which resulted in a proper front. Based on the appearances of the different Pareto fronts, the optimization is clearly sensitive to what input data models are used in the simulation model.

6.2 Discussion

The preliminary assumption that using fitted probability distributions to model input data in a simulation model would provide satisfactory Pareto fronts; and that modeling by sample mean and FACTS default distributions would be accurate enough for desirable optimization results; is not clearly reflected in the results of this thesis. The optimization demonstrates sensitivity to the particular input data models used in the simulation models by producing Pareto fronts of evidently varying properties regarding convergence and diversity, but not as expected. Despite the preliminary assumption; Model A1 exhibits the best Pareto front with respect to both convergence and diversity judging by the visual appearance of the front. The Pareto fronts for Model A2 and A4 exhibit both poor convergence and poor diversity in comparison to Model A1, with far fewer solutions and a less distinct front. It seems that the optimization has troubles searching the optimal region and finding widespread solutions providing satisfactory trade-offs for Model A2, A4 and A5; the simulation models containing distributions. These Pareto fronts are indicating that the optimization has difficulties converging with the true Pareto-optimal front, and that the problem seems to be worsened the more improvements applied to the system. This is likely due to the introduced variance of the input data models using fitted distributions.

It is problematic drawing any conclusions of this work regarding optimization sensitivity to input data models used in the simulation model, since several of the fitted distribution models used to model the data sets were evaluated as bad. The data sets of machine process times were especially difficult to find suitable distribution models to, and as a consequence
all of these distribution models were evaluated as bad by ExpertFit. Invalid distribution models will of course affect the simulation in undesirable ways. Using the wrong distribution can be as disastrous as using sample mean to model data sets when the variability is significant (Law, 2011a). It is possible that the optimization sensitivity is caused by bad distribution models, which was worsened the more changes applied to the system. Replacing the fitted distribution models describing process times in Model A2 and A3 with sample mean to create Model A6 and A4 respectively, resulted in remarkable different optimization results. Even though the actual performance of the true production system were overestimated in the simulation by both models, it is interesting that the process times had such an impact on the result. It can be concluded that it is not just using any distribution model; they must be properly validated or the result can be disastrous as shown with the modeling of machine process times. The particular distribution family seems to be less important; it is the quality of the estimated parameters that really matters. Replacing the Weibull models of Model A2 with lognormal and exponential models in Model A5 did not seriously affect the simulation result or the Pareto front of the optimization.

Besides the problems of finding appropriate distribution models to model the collected data sets; there is another issue of how the optimization has been configured. The optimization depends on the defined sets of its input parameters describing the original value and its corresponding improved value. These input data sets constitute the choice offered to the optimization and determines the decision space where the optimization searches for solutions. The poor convergence and diversity of the Pareto fronts from the models using distribution models, and the poor coverage of the optimal region when searching for optimal solutions, might indicate that a set defining a choice between two values is not sufficient for the optimization when dealing with input data models of distributions.

### 6.3 Future work

For future reference, there are several possible improvements to be made to this work which could result in better results with respect to the aim of this thesis. In future work, the work done in this thesis could be extended by removing some of the known limitations presented in the beginning of this thesis. These have been set on purpose to limit the scope of the thesis, and some of these could simply be removed to allow a thorough research of the subject.

#### 6.3.1 Modeling of failures

The failures of the machines have been modeled as one type of failure, irrespectively of the characteristics of the occurring failures. Some of the failures cause short stops and others are responsible for long durations of failures. Modeling these as separate failures by separating the data samples and treating them as coming from heterogeneous populations, could result in more accurate simulation results.

#### 6.3.2 Improved data reliability

The data sets of machine process times were collected from the manufacturing monitoring system from a time period of one single week. The short time period is problematic since it might not be representative of the normal state of the machines. However, it was not possible to acquire data from a longer period of time due to the time-consuming process of exporting process time data from the monitoring system. The number of data samples produced during a longer period was not manageable within a reasonable amount of time for
this thesis. As an effect of this problem, the process time data sets were also the data sets most difficult to find suitable distribution models for. Every single distribution model was evaluated as a bad fit. From the data analysis it was concluded that these particular data sets had higher positive skew than any other data set. Examining the data revealed that a small proportion of each data set consisted of data samples greatly deviating from the sample mean. These data samples were most likely garbage data, since their values were unreasonable with respect to the behavior of the automated machines. Acquiring data sets of higher quality from a more representative period and removing garbage data could possibly result in more accurate results in a future work.

6.3.3 Optimization input set configuration
The SCORE optimizations have been configured with input parameters consisting of input sets with only two possible values; the original value and the respective improved value. An input set consisting of only two values might not be sufficient for the optimization when the simulation models uses fitted probability distributions to model input data. Replacing these binary input sets with multiple choice sets of several alternatives would possibly improve how the optimization searches the optimal region. This area is certainly interesting for possible future work.
References


