# An operational validation approach for logistic simulation models of Vanderlande Industries

Menting, M.M.T.

*Award date:*
2010

Eindhoven, May 2010

# An Operational Validation Approach for Logistic Simulation Models of Vanderlande Industries

by
Mark Menting

BSc Industrial Engineering and Management Science — TU/e 2008
Student identity number 0577081

in partial fulfilment of the requirements for the degree of

**Master of Science**
**in Operations Management and Logistics**

Supervisors:
dr.ir. R.M. Dijkman, TU/e, IS
dr.ir. H.P.G. van Ooijen, TU/e, OPAC
ir. P.J.A. Thoonen, Vanderlande Industries

TUE. Department Technology Management.
Series Master Theses Operations Management and Logistics

Subject headings: Discrete Simulation, Logistics; simulation

# *I. Abstract*

Within this Master Thesis a validation approach is developed for logistic simulation models of Vanderlande Industries, based on comparison of a simulation model to the corresponding real system. The approach is designed based on an extensive literature study and experiences acquired by performing a case study. It will be shown that real system output is required for comparison with simulation output, in order to be able to obtain a high degree of confidence in the validity of a model. As it turned out, differences should be evaluated under similar stochastic circumstances in order to isolate them from a simulation model's own randomness. Furthermore, statistical techniques often prove not to be adequate for deciding whether the model is valid or not.

This research has resulted in a schematic approach that presents the subsequent actions to be performed for operational validation, in a stepwise manner. On a high level, an assumptions document should be created, followed by validation of simulation input, validation of simulation output that is driven by real system input, and a sensitivity analysis. The case study results will be presented as a comprehensive illustration of the developed approach.

## *II. Management Summary*

This report is the result of a Master Thesis project at Vanderlande Industries. Vanderlande is dedicated to improving customers' business processes by providing automated material handling systems. Due to increasing complexity and scale of systems, Vanderlande Industries uses simulation to reduce the risk of nonperformance of their systems. Also, simulation may be requested by customers of Vanderlande Industries in order to confirm that the functionality and performance of the proposed system will meet their requirements. However, this requires a high degree of confidence in a simulation model and its results. A high level of confidence can be obtained by comparing a model's and system's behavior for several different sets of experimental conditions, which is called operational validation. Correspondingly, the purpose of this research has been to develop an operational validation approach for the logistic simulation models of Vanderlande Industries.

In order to develop this operational validation approach, the most appropriate methods have been selected by combining findings of an extensive literature study with experiences acquired by performing a case study. At a high level, the approach consists of maintaining an assumptions document, input validation, trace-driven output validation, and conducting a sensitivity analysis. Evaluation criteria have been used to develop a conceptual approach based on current literature. For input validation and trace-driven output validation, these criteria are: generality, power, objectivity, data, and effort. Additionally, for trace-driven output validation the subject of comparison is taken into account. Different criteria have been used in order to evaluate methods for conducting a sensitivity analysis, namely efficiency, effectiveness, robustness and ease of use.

From the literature study and the case study the following findings can be derived:

### General

- The logistic systems simulated at Vanderlande Industries are inherently stochastic. Because random inputs will produce random outputs it is difficult to relate observed differences to specific model characteristics. To isolate the differences in simulation models from a model's own randomness, the system and simulation model, or various model variants, should be evaluated under similar stochastic circumstances.
- As a result of dynamic, nonstationary input and correlated, nonstationary output, formal statistical techniques have turned out to be difficult to apply, because of violations of their assumptions.
- Since a simulation model is a simplification, and consequently merely an approximation of the real system, some differences between the real system and the model are to be expected. As a result, tests that evaluate whether system and model input or output are similar are expected to be false. Therefore, it is more useful to ask whether or not the differences between the model and system are significant enough to affect any conclusions derived from the model.
- Because validity relates to a sufficient level of accuracy for the intended purpose of a model, no definite criteria can be used in general to determine whether a model is valid or not; the effect of observed differences should be assessed in relation to the objective of the model.

### Assumptions Document

- Making the model's assumptions explicit in an assumptions document is important because the model's assumptions and input values determine whether the model is valid, and will remain valid when the real system and its environment will change.
- The assumptions document should be used to make the level of detail contained in the simulation model explicit, as well as its intended purpose.
- For Vanderlande Industries this document can have additional value; if the assumptions are conform customer specifications for instance, differences between the real system and the simulation model that can be attributed to these assumptions can be acceptable.

### Input Validation

- Statistical techniques that evaluate whether system data can be considered a random sample of a predefined distribution generally tend to be too sensitive for simulation purposes. However, the statistics can be used to compare used distributions to best fitting distributions.
- For the graphical techniques proposed experience is required to conclude whether the input is valid or not.
- In absence of the required level of experience the effect can be further evaluated within the sensitivity analysis.

### Trace-Driven Output Validation

- In order to isolate the differences between a model and the real system from a model's own randomness, output validation should be trace-driven; i.e. model output should be based on real system input.
- An increase in utilization generally leads to an increase in the difference between simulation output and system output. As such, a difficulty with output validation remains that differences are preferably evaluated per range of utilization. However, for this purpose a high amount of data is required of a relatively stable period per range of utilization, which is difficult to acquire for baggage handling systems.
- In order to acquire detailed insights in the behavior of differences between a real system and a simulation model, a high utilization range is required. Typically airports experience the most extreme capacity requirements only a few days per year, of which the dates are generally known beforehand. Data of these days would be especially appropriate for operational validation.

### Sensitivity Analysis

- A sensitivity analysis can generate insights into what causes the differences observed at the output validation.
- Furthermore, it can result in an upper boundary in what output may be reached with a simulation model when all the parameters are configured optimally, based on the level of detail it contains.
- Alternative levels of factors indicated as important by the sensitivity analysis can be used to generate new simulation results for output validation, if a more detailed evaluation of their effects is preferred.

- Differences between model variants can be isolated from a model's own randomness by using a variance reduction technique called common random numbers.

In addition to this research, some recommendations can be made related to future research:

- The research within this Master Thesis has been bounded to validation methods that could increase the degree of confidence in the simulation models. This in absence of a feedback loop from implemented systems to simulation models. Additional to this research it can be recommended to evaluate possibilities to facilitate the feedback loop itself. A large reduction of time and effort can be gained by simplifying data acquisition and processing. It would be beneficial if system responses and parameter values could be observed more directly, for instance by using BPI.
- For the different types of simulation models, with respect to their objectives, different tolerance limits can be identified that determine whether or not a model is valid.
- The operational validation approach can be used to determine which level of detail is sufficient for different types of simulation models. Dependent of the simulation requirements, a tradeoff can be made between the level of detail used in a model and the cost of performing a simulation, i.e. the complexity of the coding and the additional value of it for the simulation have to be balanced against each other.

## *III. Preface*

This thesis is the result of my graduation project for the Msc. program Operations Management and Logistics at the Eindhoven University of Technology. The assignment was performed at Vanderlande Industries, in Veghel, the Netherlands.

I would like to thank Remco Dijkman, my primary university supervisor, for his enthusiasm and support during the project, as well as the large amount of time he spent on helping me. Also, I would like to thank Henny van Ooijen, my second supervisor at the TU/e, for his useful feedback on the research design and earlier versions of the report.

I would like to thank Jeroen Goes for giving me the opportunity to conduct my master thesis project at Vanderlande Industries. Furthermore, I would like to thank Paul Thoonen, my company supervisor, for his useful input and remarks, as well as for providing me the information I required. Also, I would like to thank the other colleagues of the systems simulation department, who where always available for questions and made my stay a pleasant one.

Finally, I would like to thank my family and friends in general, and my girlfriend Margo in particular, for their overall interest and support.

Mark Menting,

May, 2010

## *Table of Contents*

# 1. Introduction

In this Master Thesis a validation approach for logistic simulation models is developed, fit to meet Vanderlande Industries' specific requirements. For this purpose findings of an extensive literature study will be combined with experiences acquired by performing a case study. The simulation department is in search of methods to increase the degree of confidence in their models. Currently at Vanderlande Industries no feedback loop exists from implemented systems to simulation models to facilitate obtaining this high level of confidence.

In this chapter the problem context and the research outline will be addressed. In section 1.1 Vanderlande will be discussed as a general company, followed by the background of the project in section 1.2. The latter will be described by addressing the role of simulation models within Vanderlande Industries. Subsequently, in section 1.3, validation of simulation models will be formulated as the research area, as well as the corresponding problem definition. The resulting objective of this Master Thesis will be underlined within section 1.4. These two sections emphasize the relevance and the necessity of this research. The methodology (section 1.5) describes the related steps to perform in order to develop a solution for the defined problem. Finally, research limitations and boundaries of this research are introduced in section 1.6, as well as the structure of the report in section 1.7.

## 1.1 Description of the Company

Vanderlande is dedicated to improving customers' business processes by providing automated material handling systems. The company has sixty years of experience in the design and implementation of integrated logistics solutions of all sizes, and in providing all the required operational support, gained with customers in a broad spectrum of industries. It was founded in 1949 as Machinefabriek E. van der Lande, a general machinery and construction company manufacturing hoists, cranes and conveyor equipment for bulk materials and oil drums. A joint venture was established with Rapistan Inc and Fenner Limited in 1963. Consequently the name changed to Rapistan Lande. In 1988 the company parted from Rapistan and the name was altered to Vanderlande Industries. With the acquisition of Gambit GMBH in 1997, the company acquired knowledge of Warehouse Management Systems. Their current solutions include warehouse automation systems, baggage handling systems and end-to-end sortation systems for the parcels and postal market (Vanderlande Industries, 2009h).

Concerning integrated logistics solutions for distribution purposes, such as warehouse control systems, automated storage, order selection, sortation and consolidation, Vanderlande Industries is a leading supplier. Over 1000 distribution projects have been covered in the recent years, in segments as diverse as care, food, fashion, automotive, parts & components, retail and business-to-consumer (Vanderlande Industries, 2009a).

The company designs, builds and services leading baggage handling systems for airports of all sizes. The systems range from raw baggage to high speed tub and track systems and robotics, and provide fast, safe and robust sortation, transportation, and security screening of departure baggage, as well as

transfer and arrival baggage. More than 600 systems of all sizes have been implemented worldwide, resulting in more than 50 years of experience (Vanderlande Industries, 2009c).

For parcel handling and documents a wide range of technologies is being offered. The parcel automation systems provide end-to-end logistics solutions in depots of all sizes: from the world's largest automated sorting hub handling well over 200.000 parcels per hour to small local depots with throughputs of some thousands of parcels per day. The company is a top 3 supplier in its market, with a history of more than 250 automated sorting centers and over 30 years experience (Vanderlande Industries, 2009f).

During recent developments more and more emphasis is placed on service. Vanderlande's services cover all system- and process-related issues throughout the life cycle of material handling systems. Different service packages are being offered: from call-out response and preventive maintenance, up to a permanent, on-site maintenance presence based on a comprehensive Service Level Agreement (Vanderlande Industries, 2009g).

Vanderlande's headquarters is located in Veghel. Local Customer Centers, from which day-to-day support is provided, are based all around the world. The company slogan 'In a world of technology a belief in people' is seen as a reflection of the core values of Vanderlande Industries. These are considered to be a strong emphasis on team play, and dedication to further improving performance everyday, which led to an excellent reputation with their customers and partners (Vanderlande Industries, 2009h).

## *1.2 Simulation at Vanderlande Industries*

Due to increasing complexity and scale of systems, Vanderlande Industries uses simulation to reduce the risk of nonperformance of their systems. Customers of Vanderlande Industries may request confirmation that the functionality and performance of the proposed systems will meet their requirements as well. Prior design testing and validation are essential in order to gain confidence in future system's performance. Without it, the performance and possible shortcomings will only become apparent once the system is built and operating. Serious faults could even emerge during commissioning, requiring design changes which can mean delay and extra costs. The high investment and mission-critical role of material handling systems, and specifically baggage handling systems, mean that in most cases these risks cannot be accepted (Schipper, n.d.). For design validation queuing models are of limited value, due to the large numbers of factors which can be encountered under both normal and exceptional operating conditions (Schipper, n.d.). Therefore simulation models are being used.

Simulation is a widespread technique for the exploration, design and optimization of complex systems. Simulation is the "replication of a dynamic process in a model, in order to arrive at conclusions that can be transferred to reality" (Nyhuis et al., 2005, p. 2). It offers facilities to model real systems by means of computer programs and to analyze and describe their behavior by changing the simulation parameters. By using simulations it is possible to evaluate the effects of system load variants, interferences, changes of routing rules, structure variants or resource alternatives sufficiently accurately (Nyhuis et al., 2005).

In terms of Vanderlande Industries this implies that customers can be confident that their future systems will meet their expectations in terms of productivity and trouble-free performance over a wide range of loads and other operational conditions. In line with the findings, the system design, configuration and dimensioning can be adjusted as necessary to ensure that the required capacity levels will be met in practice. Also the design robustness can be matched to the expected operating and emergency conditions to whatever extent is desired (Schipper, n.d.).

As well as for the purpose of design validation, simulation is also used to evaluate existing systems; for example in case of proposed design and configuration changes, capacity extensions or changed operational demands such as security requirements. In these situations, simulation can allow alternatives to be evaluated accurately, acting as a decision support tool to allow the right choices to be made to achieve the specified performance levels. A further benefit is that simulation can allow testing of possible future operational scenarios and their impact on logistics systems (Schipper, n.d.).

## *1.3 Validation of Simulation Models*

In the previous section the benefits of applying simulation models have been discussed. However, in order to benefit from the opportunities discussed, a sufficient level of accuracy of the simulation models is required. Assessing the accuracy of simulation models is part of the validation of simulation models, which is defined as "the determination whether an executable simulation model is an accurate representation of the real system" (Goossenaerts & Pels, 2006, p. 87). Its aim is to assess whether the model can be substituted for the real system for the particular objectives of the study (Goossenaerts & Pels, 2006).

Validation can be used to increase the degree of confidence that the events inferred from the model will occur under the conditions assumed. Currently, at Vanderlande Industries validity is addressed by comparing a simulation model to similar models that have turned out to be valid, as well as by applying logistic formulas based on queuing theory for instance. These validation methods are performed before drawing conclusions of the model's results. Even though models have been validated, in practice, after implementation of the real system, significant dissimilarities have been found in certain cases. Though research indicates that the mismatch between simulated performance and actual performance is rarely caused by the simulation engineer, the simulation department is, in concordance with the company slogan "improve every day", in search of methods to increase the degree of confidence in their models.

"To obtain a high degree of confidence in a simulation model and its results, comparisons of the model's and system's output behaviors for several different sets of experimental conditions are usually required" (Sargent, 2008, p. 163). Currently at Vanderlande Industries no feedback loop exists from implemented systems to simulation models to obtain this high level of confidence.

In line with the finding of Sargent (2008) and the absence of a feedback loop, the purpose of this Master Thesis project is the development of an ex post validation approach for logistic simulation models of Vanderlande Industries, based on comparison of the simulation model to the real system. Ex post refers to the fact that both model data and system data are required for the comparison, but systems are generally not built before simulation results have been finalized. This implies that the approach cannot

be seen as a substitute of the validation methods currently applied. However, it does provide the opportunity to assess whether current validation methods are sufficiently capable of evaluating the extent to which simulation models are an accurate representation of the corresponding real system. Moreover, when a high degree of confidence in a simulation model is acquired, the model could be used to aid the system optimization process, which is especially relevant within the distribution market. In case of proposed design and configuration changes, capacity extensions or changed operational demands this information is also of importance. An additional possible benefit of having more exact information of the accuracy of simulation models is that it can generate insights in what causes the found differences; certain assumptions may appear to be problematic or specific parts of code could turn out to be generally inaccurate. This knowledge is valuable for (similar) future simulation projects.

## *1.4 Research Objective*

From the problem definition it can be derived that the objective of this Master Thesis is to develop an ex post validation approach for logistic simulation models of Vanderlande Industries, based on comparison of the simulation model to the real system. This approach should be capable of giving detailed information about the accuracy of a simulation model. In case a mismatch exists, the approach should provide insights in what causes the observed differences. Furthermore, the approach should be able to handle models of various levels of detail.

Related to this objective a main research question can be identified, which can be divided into two sub questions:

- How to validate the logistic simulation models of Vanderlande Industries such that a high degree of confidence can be obtained in the model and its results?
    1. Which methods can be selected based on literature, to compare logistic simulation models to corresponding real systems?
        - 1.1 Which type of methods can be found in literature to compare logistic simulation models to corresponding real systems?
        - 1.2 Which evaluation criteria can be identified for the different types?
        - 1.3 Which techniques can be found in literature per type?
        - 1.4 Which techniques should be selected for the validation approach, based on the evaluation criteria?
    2. Which methods should be selected for comparing logistic simulation models of Vanderlande Industries to corresponding real systems, based on findings of the case study?

## 1.5 Methodology

The type of methodology that should be used depends on the research objective. Verschuren & Doorewaard (1995) differentiate between two main types of research: theoretical research and practical research.

Theoretical research, also called fundamental research, aims to solve problems in the formulation of theories. It can be divided into theory development and theory testing. Motives for conducting theory development are hiatuses/blind spots/gaps in current theory, or that current theory cannot be generalized. The purpose of theory testing is to test existing theories and if necessary adapt or optimize them (Verschuren & Doorewaard, 1995).

Practical research, on the other hand, contributes to an intervention to solve a practical situation or problem. In order to perform practical research efficiently, an intervention cycle should be carried out (Table 1). The first step of the intervention cycle proposed by Verschuren & Doorewaard (1995) is problem recognition; the problem should be clearly defined. The second step involves a diagnosis; the main causes and effects of the problem should be clearly defined. Subsequently a realistic solution should be developed that will solve the problem as defined in the previous steps. The fourth step is to implement the proposed solution. Finally, the proposed intervention should be evaluated. The lasts steps may involve an iterative process; evaluation of the implemented intervention may result in alterations, which on turn may result in changes in implantation and its evaluation.

| Step | Description |
|---|---|
| 1 | Problem Recognition |
| 2 | Diagnosis |
| 3 | Development |
| 4 | Implementation |
| 5 | Evaluation |

**Table 1: The various steps of the intervention cycle (Verschuren & Doorewaard, 1995)**

The research type applied in this Master Thesis is practical research. Hence the intervention cycle should be performed. The first two steps have already been handled in section 1.3. The recognized problem is that the simulation department is in search of methods to assess the accuracy of their simulation models more exactly than is currently the case. It is diagnosed that a comparison between a simulation model's output and the corresponding system's output is required to solve this problem. At Vanderlande Industries no feedback loop currently exists from implemented systems to simulation models, in order to give a reliable indication of the validity of the model. The initial development of a solution to overcome this problem will be based on an extensive literature study. Subsequently, this solution will be implemented for a single case. The evaluation of the implementation may lead to changes in the designed solution. Thus, some iteration may take place. Consequently, the following steps may be identified in relation to Table 1:

3.1 Identify relevant criteria for comparing validation techniques
3.2 Perform a literature review in order to create an overview of available methods
3.3 Develop a conceptual approach based on an evaluation of the various discussed methods grounded on the identified criteria
4 Perform a case study
5 Update the conceptual approach based on findings of the case study

Using a case study to test theories developed in advance is called a confirmatory case study. An explanatory case study, on the other hand, uses a case to deduce a theory (Johnston et al., 1999). "The case study is a research strategy which focuses on understanding the dynamics presented within a single setting" (Eisenhardt, 1989, p. 534). This setting can involve either single or multiple cases, and numerous levels of analysis. Cooper & Schindler (2003, p. 150) state that, compared to other studies, "case studies place more emphasis on a full contextual analysis of fewer events or conditions and their interrelations." It is harder to generalize findings because case studies generally rely on data of a single case. Nevertheless, an emphasis on detail can provide valuable insight for problem solving, evaluation, and strategy (Cooper & Schindler, 2003). This detailed information may be essential for a successful solution. Therefore the case study is applied within this research to develop a practical ex post operational validation approach. Besides ensuring that practicalities are taken into account, the case study serves as an illustration of the ex post operational validation approach.

## *1.6 Research Scope*

Analogous to the research objective the focus of this Master Thesis lies on validation methods that compare system and simulation data. Facilitating the development of the feedback loop itself is out of the scope of this research.

Due to time constraints the research is limited to a single case. Furthermore, the boundaries of the validation methods that will be considered should be addressed. Validation of simulation models is a very broad subject. In literature different validation frameworks can be distinguished (Gass, 1983; Balci, 1998; Sargent, 2008). In this report the framework of Sargent (2008) is adopted, because it is most widely adopted in literature. The framework can be observed in Figure 1.

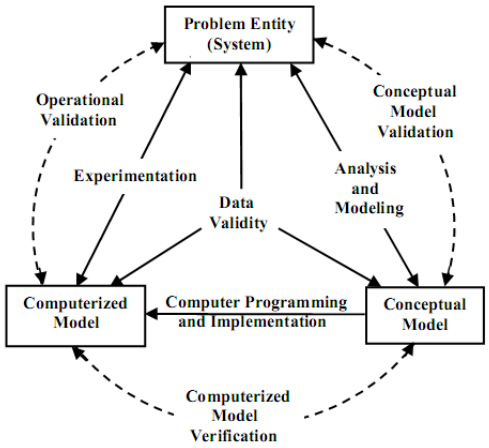



Figure 1: The modeling process (Sargent, 2008)

The framework proposed by Sargent (2008) involves various entities and validation types. A conceptual model is the "mathematical/logical/verbal representation (mimic) of the system developed for the objectives of a particular study" (Sargent, 2008, p. 159). It is developed by modeling the system, which involves making assumptions. The simulation model is an implementation of the conceptual model at a computer system, such that experiments can be conducted (Sargent, 2008).

Conceptual model validation is defined as "determining that the theories and assumptions underlying the conceptual model are correct and that the model representation of the problem entity is "reasonable" for the intended purpose of the model" (Sargent, 2008, p. 159). Data validity is defined as "ensuring that the data necessary for model building, model evaluation and testing, and conducting the model experiments to solve the problem are adequate and correct" (Sargent, 2008, p. 159). Verification is defined as "insuring that the computer program of the computerized model and its implementation are correct" (Slesinger et al., in Sargent, 2008, p. 157). Operational validation is defined as "determining that the model's output behavior has sufficient accuracy for the model's intended purpose over the domain of the model's intended applicability" (Sargent, 2008, p. 159).

As indicated by the research objective, the validation approach proposed in this report will focus upon comparison of the simulation model to the real system, which is related to operational validity. Although the focus is on a specific type of validity, any deficiencies found may be caused by what was developed in any of the steps that are involved in developing the simulation model including developing the system's theories or having invalid data, since the simulation model is used in operational validation (Sargent, 2008, p. 159).

## *1.7 Report Structure*

This chapter has provided insight into the problem background, the research area, the problem and the research objective. Furthermore, the methodology has been addressed that will be used to solve the problem. The report structure is based on this methodology.

The remainder of this report has been structured as follows. In chapter 2 relevant literature is evaluated and a first selection of appropriate methods for the validation approach is made (research sub question 1). The chapter starts with identifying various high level steps that can be applied to compare the simulation model to the real system. Within the following subsections evaluation criteria and available methods will be presented per high level step. The list of methods will be narrowed down by rating the various methods based on the evaluation criteria. In chapter 3 the initial design of chapter 2 will be applied within the case study and the resulting practical findings will be presented (research sub question 2). This evaluation leads to the schematic design presented in chapter 4, which addresses the main research question. The design is illustrated in chapter 5 by the case study. Finally, chapter 6 contains conclusions and recommendations for the application of the designed approach and further research.

## 2. Approach Development

As discussed in section 1.6, in this research the framework of Sargent (2008) is adopted, in which various types of validation are indentified. In order to be able to acquire a high degree of confidence in a simulation model and its results, the validation approach will be based upon comparison of a simulation model to the corresponding real system. As such, the approach designed in this report will be related to operational validation. Within this chapter the most suitable operational validation methods found in literature will be selected based on evaluation criteria.

The chapter is structured as follows. Within section 2.1 several high level steps will be identified that are relevant for comparing simulation models to real systems. In section 2.2, 2.3, 2.4 and 2.5 these high level steps will be discussed one by one in more detail. An overview will be given of available methods. With exception of the first step, the list of options will be narrowed down by comparison based on predefined evaluation criteria. Finally, in section 2.6, a discussion of the findings will be presented.

### 2.1 Defining High Level Steps

Many validation methods for simulation models have been proposed in literature (overviews can be found in: Balci, 1998; Law & Kelton, 2000; Kleijnen, 1995, 2005; Sargent, 2008; Trocine & Malone, 2001). Within this section the many operational validation methods for simulation models will be structured by identifying several high level steps. In Table 2 the classification of operational validity developed by Sargent (2008) is presented, which is based on the decision approach (subjective or objective) and whether or not the system is observable. Within Table 2 comparison means comparing the simulation model output behavior to either the system output behavior or another model output behavior using graphical displays or statistical tests and procedures. Explore model behavior implies to examine the output behavior of the simulation model using appropriate validation techniques, including a parameter sensitivity analysis. As stated before, this research is based upon an observable system, resulting in a higher degree of confidence in the validation results.

| | Observable System | Non-observable System |
|---|---|---|
| **Subjective Approach** | ▪ Comparison Using Graphical Displays<br>▪ Explore Model Behavior | ▪ Explore Model Behavior<br>▪ Comparison to Other Models |
| **Objective Approach** | ▪ Comparison Using Statistical Tests and Procedures | ▪ Comparison to Other Models Using Statistical Tests |

Table 2: Operational Validity Classification (Sargent, 2008)

Comparison by using graphical displays and comparison by using statistical tests and procedures is performed by black-box testing (also called functional testing) (Balci, 1998). Black-box testing only assesses the input-output transformation. White-box testing, used for exploring model behavior, uses the internal structure of the model to validate the output (Balci, 1998).

Although tests based upon an observable system are preferred, it should be taken into account that comparing real world observations and corresponding statistics from the model output data is very vulnerable to the inherent randomness of the observations from both the real system and the simulation model (Law & Kelton, 2000). Instead, if it is possible to collect data on both system input and output, it is recommended to compare model and system output by 'driving' the model with historical system input data, called trace-driven simulation, rather than samples from the input probability distributions. Since the system and the model experience exactly the same observations from the input random variables, it should result in a statistically more precise comparison (Balci, 1998; Law & Kelton, 2000). As a result of this method, assumptions about input probability functions should be validated separately. The requirement to validate both the output of the model and the input is also called the double validation problem (Balci, 1998).

Related to the operational validity classification and the double validation problem, several high level steps can be identified in literature; i.e., maintaining an assumptions document, validating input distributions, validating the output from the overall simulation model, and conducting a sensitivity analysis (Balci, 1998; Gass, 1983; Kleijnen, 1995; Law & Kelton, 2000) (Table 3). These steps together form the high level steps within the ex post trace-driven validation approach.

| Step | Description |
|---|---|
| 1. | Maintaining an Assumptions Document. |
| 2. | Input Validation |
| 3. | Trace-Driven Output Validation |
| 4. | Conducting a Sensitivity Analysis |

**Table 3: Relevant validation steps for simulation models in case real-world observations exist**

An assumptions document is used to make the theories and assumptions underlying the conceptual model explicit. A sensitivity analysis consists of changing the values of input and internal parameters of a model to determine the effect upon the model's behavior or output (Sargent, 2008). The sequence is determined based on both the literature as the case study. The assumption document should result from the modeling decisions made in earlier phases (see the modeling framework (Figure 1)). It is beneficial to perform input validation before trace-driven output validation because a notion of the real system input is required to determine which methods may be applied for output validation. This notion naturally results from conducting the input validation. Contrary to literature, conducting a sensitivity analysis is decided to be the last step. In contrast to black-box testing it can generate insights into what causes the differences observed at the output validation (if there are any). The adaption from black-box testing to white-box testing and the related increase in level of detail is the reason to perform this step lastly. When possible causes are not of interest this step may be omitted. Note that the absence of differences in an output analysis does not necessarily imply that no differences will be found in the more detailed sensitivity analysis. It merely indicates that the combination of these possible differences does not have a significant impact on the assessed output.

Schematically step 2, 3 and 4 can be drawn as in Figure 2.

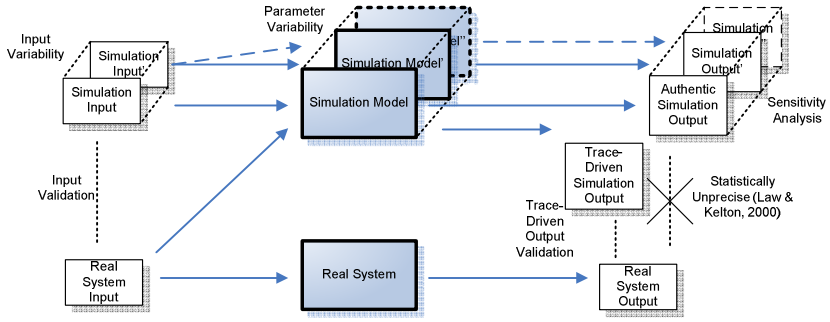



**Figure 2: Measurement of differences between a simulation model and the corresponding real system**

It should be noted that the situation of Vanderlande Industries is not exactly identical to the one described in literature. Contrary to the situation described in literature, the real system is typically built after the simulation model has been finalized. Consequently, (small) lay-out changes may occur. It is debatable whether this should be included in model inaccuracy. From the viewpoint of simulation engineers the answer is probably no; the omission of last minute lay-out changes cannot be seen as a deficiency of the model because, in terms of the discussed modeling framework (Figure 1), they were not part of the problem entity at the time. On the other hand, from the viewpoint of a customer, the answer is probably yes; the performance indicators presented based on the simulation are expected to be observed, but may not be achieved in practice. Within this research lay-out changes are included in model inaccuracy. If this is undesirable identified modifications to the problem entity should be similarly applied to the simulation model before performing the operational validation approach.

## 2.2 Maintaining an Assumptions Document

It is well-advised to maintain an assumptions document to provide information on the input values and the assumptions made in the simulation model, since "the model's assumptions and input values determine whether the model is valid, and will remain valid when the real system and its environment will change" (Kleijnen, 1995, p. 158). The document will give an indication where differences with reality can be expected. Furthermore, it can be used to evaluate the degree of similarity between simulation models; a pattern may be observed between certain assumptions and their impact on model accuracy. According to Law & Kelton (2000) relevant information can be: an overview section that discusses overall project goals, detailed descriptions of each subsystem and how the subsystems interact, what simplifying assumptions were made and why, as well as summaries of data such as the sample mean, and sources of important or controversial information.

## 2.3 Input Validation

Random input variables of the simulation model must be examined to evaluate how well they represent the true underlying distribution of the real system input data. Random input distributions will be substituted by real system input data in the output validation step. Nevertheless, these may be an important cause of perceived invalidity; for instance, observing different performance indicators in practice may be due to differences in input variables. No input variable will be exactly correct; the aim is

to validate that an applied distribution is accurate enough for the intended purpose of the model (Law & Kelton, 2000).

The first subsection identifies evaluation criteria that serve as a basis for the comparison of related techniques, as well as practical requirements to consider. Section 2.3.2 describes how to assess the general assumptions of the statistical tests that will be discussed. Section 2.3.3 depicts techniques to test whether various system samples may be combined or not, in order to make implications of the comparison as general as possible. Section 2.3.4 and section 2.3.5 specify two different required options for comparing system and simulation input (see the practical requirements for more information). In section 2.3.6 graphical techniques are presented that are applicable in both situations. The last section, section 2.3.7, presents an overview and evaluates the large amount of possible methods.

### 2.3.1 Evaluation Criteria & Practical Requirements

In order to be able to determine which input validation methods should be applied evaluation criteria are required. The criteria identified for input validation methods are: generality, objectivity, the amount of data required to obtain meaningful results, and ease of use. Furthermore, in case of hypothesis tests, statistical power should be addressed (the probability that the test will reject the null hypothesis when the alternative hypothesis is true (Montgomery & Runger, 2002)). Generality is related to the number of assumptions made and its implications. Objectivity concerns whether formal methods or subjective interpretation is being used to draw conclusions. These criteria are based upon observed differences between methods, as well as characteristics that were marked in literature as important. An overview of the criteria can be observed in Table 4.

| | Criteria |
|---|---|
| 1. | Generality |
| 2. | Power |
| 3. | Objectivity |
| 4. | Data |
| 5. | Effort |

Table 4: An overview of the evaluation criteria for input validation techniques

In addition to the evaluation criteria there are some practical requirements to consider. The most trivial requirement is that data should be available to some extent; without it the methods discussed in this chapter are not applicable. This requirement holds for every input validation attempt. Note that data validation is not considered in this approach. Within this setting it is assumed that data is correctly measured and validated prior to this approach. In practice some data evaluation may be required, such as the assessment of outliers.

Furthermore, some practical requirements specific to Vanderlande should be considered. As described in the chapter introduction, probability distributions are generally being used to generate the random input variables of a simulation. For relatively small projects this is also true for Vanderlande Industries. However, for large projects the arrival process is often developed by a client itself, using his expertise to generate an input file containing exact customer arrival times, called a load file. Therefore, in order to

validate the random input variables, comparisons of a probability distribution and a system sample, as well as comparison of the samples of a load file and of the system should be taken into account.

Lastly, it should be taken into consideration that a model is generally developed for simulating some predefined behavior. For instance, for baggage handling systems this behavior can deviate significantly from the average state of the real system; simulation studies are typically performed for evaluating design performance with the highest amount of arrivals, i.e. peak loads. Therefore only system data should be used for comparison that reflects the intended purpose of the simulation model.

### 2.3.2 Data Evaluation: Data Stationarity & Sample Independence

A common assumption in many time series techniques is that sample data is stationary. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time (NIST/SEMATECH, 2010). Said otherwise, the techniques assume that the observations are random drawings, from a fixed distribution with a common location and a common scale. A location parameter shifts a probability density function left or right on the horizontal axis (change in mean). An increase in scale parameter stretches a probability density function on the horizontal axis (change in variance). As a reference, the standard normal distribution has a location parameter equal to zero and a scale parameter equal to one (NIST/SEMATECH, 2010). It should be noted that this definition of stationarity is not sufficient for the statistical techniques that will be discussed in this chapter; not only should the autocorrelation structure not change over time, it should also be insignificant (observations are independent). The techniques may not be valid if these assumptions are not satisfied. Consequently only the graphical techniques of section 2.3.6 would be applicable (Law & Kelton, 2000).

Various techniques based on linear regression have been proposed in literature for assessing stationarity (appendix A.1.1). Most of these methods require the observations (or residuals) to be normally distributed. This is typically not to be expected for an arrival process. The non-parametric alternatives based on linear regression are difficult to apply, because they are not commonly supported by statistical software packages. Alternatively, stationarity of mean and variance can be evaluated with a run sequence plot.

Additional techniques should be used to assess data independence. Due to the presence of many ties in the data, the Von Neumann's ratio test and run tests cannot be applied. Instead the Box-Ljung test or the Pankrantz criterion may be performed. Related to graphical techniques for the evaluation of autocorrelation, the autocorrelation plot is preferred over the lag plot because it is able to assess various lags in a single graph. The Pankrantz criterion is based on the autocorrelation plot. A more in-depth discussion of relevant techniques can be found in appendix A.1.1.

### 2.3.3 Sample Selection

In many cases data will be available of more than one period. Comparison of simulation data with system data of as many periods as possible is preferred, because this will increase the generality of the findings. Two options exist for incorporating multiple samples in the comparison: simulation input can be compared with each system sample separately, or simulation input can be compared to a group of

similar system samples. In principle this last option is preferred. A larger sample size will decrease both the probability of a type I error (risk of rejecting a true hypothesis) and the probability of a type II error (risk of accepting a false null hypothesis) (Montgomery & Runger, 2002). Therefore, methods are required in order to be able to determine whether or not different samples may be combined.

Techniques based on ANOVA (analysis of variance) are only limited applicable due to normality assumptions. Of the nonparametric test, the Friedman test is preferred if data samples represent blocks, such as a day of the week. A disadvantage of the Friedman test is that the method requires different samples to be of exactly the same size (Statpoint Technologies, 2009). When this is not the case the Kruskal-Wallis test is the most appropriate. Mood's median test lacks power in comparison to the other two tests, but it is more general. Related to comparing variances of samples, Bartlett's test should be applied if observations are normally distributed. Levene's test based on the median is recommended as the choice that provides good robustness against many types of non-normal data while retaining good power. The variant based on the trimmed mean performs best when the underlying data follows a heavy-tailed distribution. Using the mean provided the best power for symmetric, moderate-tailed, distributions. For a profound assessment of these methods one is referred to appendix A.1.2.

### 2.3.4 Goodness-of-Fit Tests

"A goodness-of-fit test is a statistical hypothesis test that is used to assess formally whether the observations $X_1$, $X_2$, …, $X_n$ are an independent sample from a particular distribution with distribution function $\hat{F}$" (Law & Kelton, 2000, p. 356). Analogous, the null-hypothesis can be stated as: $H_0$: The $X_i$'s are independent and identically distributed random variables with distribution function $\hat{F}$. It should be noted that a failure to reject $H_0$ does not necessarily imply that $H_0$ should be accepted (Law & Kelton, 2000). Furthermore, for small to moderate sample sizes, these tests are not very sensitive to subtle disagreements between the data and the fitted distribution. However, if the sample size is very large, these methods will almost always reject the hypothesis that the data fits the defined distribution since this is virtually never exactly true (Law & Kelton, 2000). This effect is not desired since it is usually sufficient to have a distribution that is nearly correct (Law & Kelton, 2000).

The chi-square test, Kolmogorov-Smirnov test, and the Anderson-Darling test are goodness-of-fit tests that can be used to assess whether a system sample stems from the proposed distribution. The chi-square test is the most general test. It also has the least power. This is something that can be observed in general; tests making less assumptions, e.g. nonparametric tests, can be more widely applied, but relate to less statistical power. The Anderson-Darling test is applicable similar to the Kolmogorov-Smirnov test, though it has slightly more power. Furthermore, the Anderson-Darling test focuses mainly on differences in distribution tails, while the Kolmogorov-Smirnov test is more sensitive to differences near the center of the distribution (NIST/SEMATECH, 2010). The chi-square test requires the most effort; a suitable interval width has to be determined. Since this is generally a subjective step, the test is also less objective. Additional information about this topic can be observed in appendix A.1.3.

### 2.3.5 Comparing Two Samples

When a load file (an estimated sample defined by the client) is used in a simulation model instead of an input distribution, input validation will be based on comparing two samples.

Several two independent samples tests, e.g. Student's t-test, F-test, Wald-Wolfowitz runs test, two-sample Kolmogorov-Smirnov test, and Mann-Whitney U test, can be used to evaluate the degree of similarity between two samples. Both the t-test and the F-test assume normality of samples. This assumption may not hold; arrival rates and service times are typically characterized as an exponential process. The t-test turned out to be very robust against this assumption, while the F-test is not. The Wald-Wolfowitz test and the two-sample chi square test are the most general statistics, though at the cost of statistical power. Therefore the two-sample Kolmogorov-Smirnov and Mann-Whitney U test may be preferred alternatives. The difference between these tests lies in their focus. While the two-sample Kolmogorov-Smirnov test focuses on the largest difference, the Mann-Whitney U test focuses on differences in median.

The F-test distinguishes itself from the other tests by focusing on equality of variances instead of on the mean/median. Alternatives to the F-test that do not rely on the normality assumption have already been discussed in section 2.3.3. These tests are suitable for two or more samples and therefore are also applicable in this situation. A more elaborate debate is presented in appendix A.1.4.

### 2.3.6 Graphical Techniques

Several graphical procedures can be used for comparing fitted distributions with the true underlying distributions, as well as for the comparison of two data samples. Graphical displays of sample data are very powerful and useful ways to visually examine the data and search for existing differences (Montgomery & Runger, 2002; Balci, 1998).

For continuous data, a density/histogram overplot can be made by plotting the probability density function of the applied distribution over the histogram of the data. Alternatively, a frequency comparison can be used, which compares the intervals of two histograms. These can be based on one data sample and a distribution, as well as on two data samples. A frequency comparison can also be used in case of a discrete distribution (Law & Kelton, 2000). A disadvantage is that histograms are generally not really reliable indicators of the distribution form, unless the sample size is sufficiently large (Montgomery & Runger, 2002).

Instead of basing the graphical comparison on probability density functions, it is also possible to use the cumulative distribution functions. Such a technique is the distribution function difference plot (Law & Kelton, 2000).

Another graphical technique based on the cumulative distribution function is the probability plot. Probability plotting is more reliable than the histogram for small to moderate sample sizes (Montgomery & Runger, 2002). Several kinds of probability plots exist. The Q-Q plot (where Q stands for quantile) compares the probability distribution by plotting their quantiles against each other (Gibbons &

Chakraborti, 2003). Q-Q plots amplify differences that exist between the tails of the compared distribution functions (Law & Kelton, 2000). A different probability plot, the P-P plot (where P stands for probability or percent), on the other hand, amplifies the differences between the middle parts of the distribution functions (Law & Kelton, 2000).

### 2.3.7 Overview

In section 2.3 it has been shown that validation of random input variables should be possible by comparison of two samples as well as by comparing a sample to a distribution. For both comparisons graphical techniques and hypothesis tests can be performed. Graphical techniques can be used to indicate differences without making assumptions. Hypothesis tests are more formal methods for addressing whether random input variables are consistent with system input data. However, these require observations to be stationary and independent of each other. Several techniques have been proposed in order to assess these requirements.

The results are summarized in Table 5. Responses are rated according to a five-level Likert scale, such that a high rating is always associated with a positive result. Note that ratings should be interpreted as an ordinal scale; numbers indicate the relative position of items, but not the magnitude of difference (Cooper & Schindler, 2003). Furthermore, power is difficult to interpret in this setting: a test with less power might be preferred because some inaccuracy is acceptable (a difference always exists). The techniques that appeared to be most appropriate, based on the discussions in section 2.3, are underlined.

| | | | |
|---|---|---|---|
| Generality: | 1 | = | The technique is very limited applicable |
| | 2 | = | Very strong assumptions are made |
| | 3 | = | Considerable assumptions, though robust |
| | 4 | = | Minor assumptions, e.g. only data independence required |
| | 5 | = | Virtually no limiting assumptions used |
| Power: | 1 to 5 | with 1 very limited statistical power and 5 very high statistical power | |
| Objectivity | 1 | = | Completely subjective, no guidelines included |
| | 2 | = | Subjective, guidelines included |
| | 3 | = | Formal test, though involves subjective steps |
| | 4 | = | Objective, only sampling bias |
| | 5 | = | Completely objective |
| Data: | 1 to 5 | with 1 very large data set required and 5 very limited data required | |
| Effort: | 1 to 5 | with 1 very high amount of effort required and 5 very little effort required | |

| Category | Method | Focus | Generality | Power | Objectivity | Data | Effort |
|---|---|---|---|---|---|---|---|
| Stationarity and data independence | ▪ Linear regression based methods | Residuals | 2 | 5 | 4 | 2 | 4 |
| | ▪ Non-parametric regression based methods | Residuals | 4 | 4 | 4 | 3 | 2 |
| | ▪ Von Neumann's ratio test | Ranks | 3 | 4 | 4 | 3 | 4 |
| | ▪ Run test | Runs | 3 | 2 | 4 | 3 | 4 |
| | ▪ Pankrantz criterion | Lags | 4 | 4 | 3 | 3 | 4 |
| | ▪ Box-Ljung test | Lags | 4 | 4 | 4 | 3 | 4 |
| Comparison of simulation samples | ▪ ANOVA | Mean | 2 | 5 | 4 | 2 | 4 |
| | ▪ Friedman Test | Median | 3 | 4 | 4 | 3 | 4 |
| | ▪ Kruskal-Wallis Test | Median | 4 | 3 | 4 | 3 | 4 |
| | ▪ Mood's median Test | Median | 5 | 2 | 4 | 3 | 4 |
| | ▪ Bartlett's test | Variance | 2 | 5 | 4 | 2 | 4 |
| | ▪ Levene's test | Variance | 4 | 4 | 4 | 3 | 4 |
| Comparison of samples to distributions | ▪ Chi-square Test | Intervals | 5 | 3 | 3 | 3 | 3 |
| | ▪ Kolmogorov-Smirnov Test | Largest Distance | 4 | 4 | 4 | 3 | 4 |
| | ▪ Anderson-Darling Test | Distribution Tails | 4 | 5 | 4 | 3 | 4 |
| Comparison of a simulation input sample to a system input sample | ▪ T-Test | Mean Differences | 3 | 5 | 4 | 2 | 4 |
| | ▪ F-Test | Variance Differences | 2 | 5 | 4 | 2 | 4 |
| | ▪ Wald-Wolfowitz runs test | Runs | 5 | 2 | 4 | 3 | 4 |
| | ▪ Two-Sample Chi-square test | Intervals | 5 | 3 | 3 | 3 | 3 |
| | ▪ Two-Sample Kolmogorov-Smirnov Test | Largest Distance | 4 | 4 | 4 | 3 | 4 |
| | ▪ Mann-Whitney U test | Medians | 4 | 4 | 4 | 3 | 4 |
| General Comparisons | ▪ Graphical Techniques | - | *5* | - | *2* | *4* | *4* |

**Table 5: Comparison of input validation techniques**

It can be concluded that for both the comparison of two samples and comparing a sample to a distribution, several alternatives can be used. Alternative methods typically focus on other differences (e.g. medians, tails, and largest differences). The application of several methods can be complementary and additional insights may be reached with relatively little extra effort.

## *2.4 Trace-Driven Output Validation*

Statistical procedures can be used to compare system output data to simulation output data, which is generated based on historical system input data (e.g. actual observed interarrival times and service times). Balci (1998) discusses many statistical techniques for comparing real system output with simulation output based on the same input, such as t-test, Mann-Whitney-Wilcoxon test, two-sample chi-square test, two-sample Kolmogorov-Smirnov test, and Hotelling's $T^2$ test (the multivariate counterpart of the t-test (NIST/SEMATECH, 2010)). Though with respect to these methods the "same" implies that the model input data are coming independently from the same populations or stochastic process of the system input data, in which case the model and system output data can be expected to be independent and identically distributed, which is required for the various tests. However, since the comparison intended in this research is trace-driven, the "same" indicates that the model input data are exactly the same as the system input data (Balci, 1998). Hence the model and system output data are expected to be dependent and identical, and thus the assumptions of the tests cannot be satisfied (or only with great difficulty) (Sargent, 1996). Secondly, many statistical tests require the data to have a

normal distribution, which is usually not the case for data collected from a system or simulation model (Sargent, 1996). In addition, almost all real-world systems and simulations are nonstationary (the distributions of the successive observations change over time) and autocorrelated (the observations in the process are correlated with each other) (Law & Kelton, 2000; Sargent, 1996). Furthermore, even when all assumptions do apply, in many instances the tests cannot be used because there is not enough data available from the system in order to obtain meaningful results from the statistical tests (Sargent, 1996).

Law & Kelton (2000), Kleijnen (1995) and Sargent (2008) give overviews of applicable methods, which will be discussed in section 2.4.2, section 2.4.3, section 2.4.4 and section 2.4.5. First evaluation criteria are determined that serve as a basis for the comparison of related techniques, as well as practical requirements to consider. The last section, section 2.4.6, summarizes and compares the various possible methods.

### 2.4.1 Evaluation Criteria & Practical Requirements

For the evaluation of trace-driven output validation techniques similar criteria can be identified as for input validation, namely generality, objectivity, the amount of data required to obtain meaningful results, ease of use, and, in case of hypothesis tests, statistical power. In addition, the subject of comparison can be taken into account. Besides the usual mean and variance, correlation of output is interesting because identical input is being used. An overview of the criteria can be observed in Table 6.

| | Criteria |
|---|---|
| 1. | Subject of Comparison |
| 1.1 | Mean |
| 1.2 | Variance |
| 1.3 | Correlation |
| 2. | Generality |
| 3. | Power |
| 4. | Objectivity |
| 5. | Data |
| 6. | Effort |

Table 6: An overview of the evaluation criteria for trace-driven output validation techniques

The same trivial practical requirement is relevant in this setting as for input validation; system output data should be available to some extent. Again it is assumed that data is correctly measured and validated prior to the conceptual validation. Nevertheless, some data evaluation is required, specifically related to outliers that may be caused by downtime for instance.

### 2.4.2 The Correlated Inspection Approach

The first method, called the correlated inspection approach, prescribes to “compute one or more statistics from the real-world observations and corresponding statistics from the model output data, and then to compare the two sets of statistics without the use of a formal statistical procedure” (Law & Kelton, 2000, p. 283). Examples are the sample mean, the sample variance, the sample correlation function and graphical plots (Law & Kelton, 2000). Sargent (1996) and Kleijnen (1995) discuss several

graphical methods, namely histograms, box plots and behavior graphs, which can be used for operational validity. These do not require the data to be independent, have no distributional requirements on the data, and can be used with a limited number of observations (Sargent, 1996).

Though the correlated inspection approach does not use a formal statistical procedure to compare real-world and simulation statistics, it may provide valuable insights into the adequacy of a simulation model and it will often be the only feasible statistical approach due to limitations on available data (Law & Kelton, 2000). Due to the lack of a formal, objective procedure to compare the two sets of data, determining whether the model has sufficient accuracy for its intended purpose should be done subjectively. This comparison can be made by the model development team and/or by experts using face validity or Turing tests (Kleijnen, 1995; Sargent, 1996). A Turing test is performed by asking people knowledgeable about the system to examine and identify one or more sets of system data as well as one or more sets of model data without knowing which sets are which (Law & Kelton, 2000). Related to this, a model is said to have face validity when simulation results are consistent with perceived system behavior (Law & Kelton, 2000). Additional insights about this method can be obtained in appendix A.2.1.

### 2.4.3 Confidence-Interval Approach

When it is possible to collect a potentially large amount of data from both the model and the system it is possible to create confidence-intervals based on output differences. This is a more reliable approach for comparing a model with the corresponding real system (Balci, 1998; Law & Kelton, 2000). The combination of confidence intervals of various output differences is also called the model's range of accuracy (Balci, 1998).

Because the model and system output are dependent (the same input values have been used), the paired-t approach should be used for creating confidence intervals for the differences in responses. This method pairs dependent observations, and therefore requires the amount of observations of system output and model output to be equal. Furthermore, the paired-t method assumes the response differences to be independent and identical (IID) random variables, and normally distributed (Law & Kelton, 2000). It is important to note that the responses should be random variables over entire independent replications (e.g. a single day in a terminating system). As a result the data is IID as required (Law, 2007; Kleijnen, 1995). The method is quite robust for deviations of the normality assumption; the central limit theorem applies (also when autocorrelation exists) (Kleijnen, 1995; Law & Kelton, 2000), which means that the coverage probability will be near $1 - \alpha$ for a large number of observations (with $\alpha$ being the probability of a type I error). Therefore the test may still be applied in case of non-normality (Kleijnen, 1995). In contrast to the classical two-sample-t approach, Var(model) = Var(system) is no prerequisite (Law & Kelton, 2000). For more details about this approach one is referred to appendix A.2.2.

### 2.4.4 Regression Based Approach

In Kleijnen (1999) and Kleijnen et al. (1998) two validation approaches for trace-driven simulations are discussed, which are based on a standard regression analysis. As for the confidence interval approach,

both methods assume the outputs of the real system and the simulated systems to be identically and independently, as well as normally distributed.

It is important to note that these regression based methods are hypothesis tests. In Law (2008) it is questioned whether hypothesis tests, concerning possible differences between real system output and simulation output, are the appropriate statistical approach, since a simulation model is a simplification, and consequently merely an approximation, of the real system. Therefore a null hypothesis that the system and model output stem from the same distribution, is expected to be false. As a consequence it is more useful to ask whether or not the difference between the model and system output are significant enough to affect any conclusions derived from the model (Kleijnen, 2005; Law, 2008). A hypothesis test does not give additional insights into the magnitude of an observed difference. The various regression based methods are presented in more detail in appendix A.2.3.

### 2.4.5 Time-Series Approaches

A downside of the three methods described so far is that they provide little information about the autocorrelation structures of the two output processes (Law & Kelton, 2000). When there is a strong suspicion that autocorrelation is of major importance, time series might be more useful. “A time series is a finite realization of a stochastic process” (Law & Kelton, 2000, p. 289). Time-series approaches only require one set of each type of output and can be used to formally compare the autocorrelation functions of two samples (Kleijnen, 1995; Law & Kelton, 2000).

Difficulties with time series approaches are that they require the output processes to be covariance-stationary (generally not true in practice), a high level of mathematical sophistication, and long time series (Kleijnen, 1995; Law & Kelton, 2000; Van Horn, 1971). Furthermore, it may be difficult to relate the generated confidence interval to the validity of the simulation model (Law & Kelton, 2000; Van Horn, 1971). Lastly, some time series approaches are formulated as hypothesis tests, of which the usability has been criticized in the previous section. Additional information about time-series approach can be acquired in appendix A.2.4.

### 2.4.6 Overview

When comparing trace-driven simulation output with system output, there are three important parameters to consider, i.e., sample mean, variability, and correlation (Kleijnen et al., 1998). These issues can be addressed with the correlated inspection approach, the confidence-interval approach (only for differences in mean), regression based approach or time-series approach.

The results are summarized in Table 7. Responses are rated according to a five-level Likert scale, of which the ratings are similar to the ratings in section 2.3.7. A high rating is always associated with a positive result and ratings should be interpreted as an ordinal scale; numbers indicate the relative position of items, but not the magnitude of difference (Cooper & Schindler, 2003). As in the previous overview section, the technique that appeared to be the most suitable, based on the previous discussions, are underlined.

| Method | Mean | Variance | Correlation | Generality | Power | Objectivity | Data | Effort |
|---|---|---|---|---|---|---|---|---|
| Correlated Inspection Approach | v | v | v | 5 | - | 2 | 4 | 5 |
| Confidence Interval Approach | v | v | - | 3 | 4 | 4 | 2 | 4 |
| Regression Based Approach | v | v | Between Var. | 3 | 3 | 4 | 2 | 3 |
| Regression Based on Bootstrapping | v | v | Between Var. | 4 | 3 | 4 | 3 | 2 |
| Time-Series Approach | v | v | v | 3 | 3 | 2 | 1 | 1 |

Table 7: Comparison of trace-driven output validation techniques

In conclusion, the correlated inspection approach is practically always applicable, from which important insights can be gained, while requiring relatively little effort. In case sufficient data is available the approach can be extended with confidence intervals. Regression based methods only provide limited information in case a difference is found. The application of time series is only recommended when comparison of autocorrelation is of specific interest.

## 2.5 Conducting a Sensitivity Analysis

A sensitivity analysis is a technique for determining which model input parameters have a significant impact on the desired measures of performance, and consequently need to be modeled carefully (Law & Kelton, 2000). It can enhance model validity by assuring that those values are specified with sufficient accuracy (Balci, 1998). Even when there is abundant data on the input and output of the simulated system, this information is very useful (Kleijnen, 1995). In contrast to trace-driven output validation, which only assesses the input-output transformation, a sensitivity analysis can reveal important information about possible causes of differences. A sensitivity analysis is performed by systematically changing the values of model input variables and parameters over some range of interest and observing the effect upon model behavior (Balci, 1998). Unexpected effects may reveal invalidity. Examples of model input variables that could be investigated are: the value of a parameter, the choice of a distribution, the entity moving through the simulation system (e.g. a single item or a batch), and the level of detail for a subsystem (Law & Kelton, 2000; Law, 2008).

This section is structured as follows. The first subsection identifies evaluation criteria that serve as a basis for the comparison of related techniques, as well as practical requirements to consider. In section 2.5.2 the main method will be discussed. Available techniques related to the proposed main method will be presented in section 2.5.3, 2.5.4 and 2.5.5. Validation of these techniques will be addressed thereafter. In section 2.5.7 methods will be introduced that can reduce the variance experienced during a sensitivity analysis. The last section, section 2.5.8, presents an overview and comparison of the methods discussed in this section.

### 2.5.1 Evaluation Criteria & Practical Requirements

When selecting a procedure for conducting a sensitivity analysis there are four main criteria to consider, namely efficiency, effectiveness, robustness, and ease of use (Trocine & Malone, 2001). Efficiency relates to the amount of runs required for screening the factors. Effectiveness is whether the metric yields accurate underlying coefficients of the effects. Note that this is only measurable in simulated cases with known coefficients (i.e. the true factor effect values are known). The degree of confounding is taken into account as part of effectiveness; the confounded effects cannot be estimated accurately. Robustness

involves the conditions which may be required for a method to be applicable (Trocine & Malone, 2001). The last criterion, ease of use, is related to the time and effort required for performing an experimental design. An overview of the criteria can be observed in Table 8.

| | Criteria |
|---|---|
| 1. | Efficiency |
| 2. | Effectiveness |
| 3. | Robustness |
| 4. | Ease of use |

**Table 8: An overview of the evaluation criteria for sensitivity analysis techniques**

In theory ease of use is not a necessity, but rather a positive incidental circumstance (Trocine & Malone, 2001). For practical considerations, however, it is of high importance in order to insure a proper adoption of the proposed approach. In principle every design strategy can be performed by manually conducting all runs. Though this requires a very high amount of effort and therefore is not recommended. Thus, options for conducting simulation runs are limited to the capabilities of the simulation package used; in case of Vanderlande Industries AutoMod.

### 2.5.2 Design of Experiments

The classical approach is to check for output changes when one factor is varied at the time, while others are set to some arbitrary value. This approach is called the one-factor-at-a-time approach (Law, 2008). However, in case two or more factors exist, applying this method may not be correct because it neglects interactions among factors (Law & Kelton, 2000; Law, 2008). Furthermore, other methods are more efficient and accurate (lower variance) than the one-factor-at-a-time approach (Kleijnen, 1992). These methods vary multiple factors at a time, and involve making simulation runs based on particular configurations, so that the factor effects can be estimated with the least amount of simulation. Designing such a configuration is called Design Of Experiments (DOE) (Kleijnen et al., 2004a). In DOE terminology, model input parameters are called factors, and output measures are called responses (Law & Kelton, 1991). The simulation model is run for the set of factor combinations and the resulting input-output data are analyzed to estimate factor effects (Kleijnen, 1997). Appendix A.3.1 addresses type of models that can be used for this estimation. It is shown that a linear model can be used to estimate the factor effects within a sensitivity analysis. Consequently, methods can be may be applied that are based upon linear regression models. In the subsequent sections these techniques will be discussed.

### 2.5.3 A Full Factorial Design

For experiments involving the study of the effects of two or more factors, factorial designs are most efficient (Montgomery, 1991). A factorial design is the investigation of all possible combinations of the levels of the factors in each complete trial or replication of the experiment (Box et al., 2005; Montgomery, 1991). The factorial design associated with a first-order polynomial is an experiment with two levels for all k factors. It is possible to evaluate the effect on several responses in one experimental run. In contrast to the one-factor-at-a-time approach, the factorial design is capable of taking interactions among factors into account (Law & Kelton, 2000).

A downside of the full factorial design is that the amount of runs required tends to become large, when testing an increasing amount of factors. A relatively large part of these runs is required for determining the many degrees of freedom that are associated with higher interactions, which are often negligible (Montgomery, 1991). For example a complete $2^6$ design requires 64 runs, of which 6 of the 63 degrees of freedom are related to main effects, 15 degrees of freedom correspond to two-factor interactions, and the remaining 42 degrees of freedom are associated with three-factor and higher interactions (Montgomery, 1991). For a more in-depth discussion of the full factorial design one is referred to appendix A.3.2.

### 2.5.4 Fractional Factorial Design

In most experimental designs the sparsity of effects principle applies; the system is dominated by the main effects and low-order interactions. The three-factor and higher order interactions are usually negligible (Montgomery & Runger, 2002). In this case information about the main effects and low-order interactions may be obtained by running only a fraction of the complete factorial experiment, which is called a fractional factorial design (Montgomery, 1991). This is realized by using the same levels (both low/high at the same time) for certain factors (preferably higher interactions). As a result one cannot differentiate between these effects. This property is called aliasing (Montgomery, 1991), or confounding (Law & Kelton, 2000).

Several important fractional factorial designs have been classified (Box et al., 2005; Law & Kelton, 2000; Montgomery, 1991). Resolution III designs (denoted as $2_{III}^{k-p}$), which are defined as "designs in which no main effects are aliased with any other main effect, but main effects are aliased with two-factor interactions and two-factor interactions may be aliased with each other" (Montgomery, 1991, p. 339). Resolution IV designs (denoted as$2_{IV}^{k-p}$) are "designs in which no main effect is aliased with any other main effect or with any two-factor interaction, but two-factor interactions are aliased with other" (Montgomery, 1991, p. 339). Resolution V designs (denoted as $2_{V}^{k-p}$) can be defined as "designs in which no main effect or two-factor interaction is aliased with any other main effect or two-factor interaction, but two-factor interactions are aliased with three-factor interactions" (Montgomery, 1991, p. 339). "In simulation there will often be at least two-way interactions of interest" (Law & Kelton, 2000, p. 639). Therefore, resolution IV designs may be inadequate and it is strongly recommended to use resolution V designs in simulation (Law & Kelton, 2000). Additional insights about fractional factorial designs can be acquired in appendix A.3.3.

### 2.5.5 Factor-Screening Strategies

Screening is defined as "the search for the most important factors among a large set of factors in an experiment" (Kleijnen et al., 2006, p. 287). "The purpose of factor screening is to eliminate negligible factors in favor of concentrating experimental efforts on those factors that are important" (Trocine & Malone, 2001, p. 169). This is possible because, equivalent to the Pareto rule, only a few factors are responsible for most of the effect in a response (Kleijnen et al., 2006; Box et al., 2005; Trocine & Malone, 2001). Many different screening strategies can be identified. Examples are Plackett-Burman Designs,

supersaturated designs, and group screening designs. These are discussed in detail in appendix A.3.4, A.3.5 and A.3.6.

### 2.5.6 Validation of the Sensitivity Analysis

Determining the validity of the sensitivity analysis can be done by running new scenarios and comparing simulation output with sensitivity analysis prediction by calculating the correlation (Vonk Noordegraaf, 2002). An alternative procedure, which requires no new simulation runs, is cross-validation (Kleijnen, 1995). Cross-validation eliminates scenarios one by one and re-estimates the regression model. Subsequently, the resulting factor effects are used to predict the simulation realization of the deleted scenario. These predictions can be compared with the corresponding simulation responses using the Pearson linear correlation coefficient or comparing the responses through a scatter plot (Van Groenendaal & Kleijnen, 1997). For the last option it can be decided whether the factor estimates are acceptable by eyeballing the plot (the points will lie upon an approximately linear line with an intercept of 0 and a slope of 1) (Kleijnen, 2005).

### 2.5.7 Variance Reduction Techniques

The logistic systems simulated by Vanderlande Industries are so-called Discrete-Event Dynamic Systems; i.e. the simulation is inherently stochastic (Kleijnen, 2008). Law & Kelton (2000), as well as Farrington & Swain (1993), state that random inputs will produce random outputs. As a result more replications of the experiment are required in order to acquire acceptable confidence interval widths. Consequently the amount of computer time needed for the experiment will increase drastically. Variance reduction techniques can be used to keep the number of replications required to a minimum, while preserving statistical adequacy (Kleijnen, 2008). Variance reduction techniques aim to reduce the variance of an output random variable without disturbing its expectation, and consequently obtain greater precision (e.g. smaller confidence intervals) (Law & Kelton, 2000). The most useful technique is common random numbers (Kleijnen, 2008; Law & Kelton, 2000). A more extensive discussion of common random numbers can be found in appendix A.3.7.

### 2.5.8 Overview

In section 2.5 various techniques for performing a sensitivity analysis based on a linear model have been identified. The results are summarized in Table 9. Responses are rated according to a five-level Likert scale, such that a high rating is always associated with a positive result. Effectiveness is rated by taking practical significance into account. Note that ratings should be interpreted as an ordinal scale; numbers indicate the relative position of items, but not the magnitude of difference (Cooper & Schindler, 2003). As for the other overview sections, the technique that appears to be most appropriate is underlined.

| Method | Efficiency | Effectiveness | Robustness | Ease of Use |
|---|---|---|---|---|
| Full Factorial Design | 1 | 5 | 5 | 5 |
| Fractional Factorial Design | 2 | 4 | 4 | 4 |
| Plackett-Burman Designs | 3 | 3 | 2 | 1 |
| Supersaturated Designs | 4 | 1 | 1 | 1 |
| Two-Stage Group Screening | 4 | 3 | 3 | 3 |
| Sequential Bifurcation | 5 | 2 | 3 | 2 |
| Iterated Fractional Factorial Design | 2 | 2 | 3 | 1 |
| Controlled Screening | 3 | 4 | 3 | 1 |

Table 9: Comparison of sensitivity analysis techniques

In conclusion, the appropriate approach depends on how the different criteria are weighted. In order to achieve practically meaningful results, without requiring too many runs, the two-stage group screening method is preferred. Possibly this technique may be extended with CRN, dependent of the additional efforts required.

## 2.6 Discussion

In this chapter operational validation methods found in literature have been evaluated based on predefined criteria. It has been shown that four high level steps are important for comparing a simulation model to the corresponding real system, namely maintaining an assumptions document, validating input distributions, validating trace-driven output, and conducting a sensitivity analysis. However, some considerations remain concerning the selection of specific methods related to the last three steps.

Selections within this chapter are partly based on assumptions about simulation models, which literature claims to be generally true. For example, it is assumed that interarrival times are typically best characterized as an exponential process and that simulation output is nonstationary and autocorrelated. It should be verified that these assumptions hold for simulation models of Vanderlande Industries, before the selections in this chapter can be used within the operational validation approach.

Evaluating which assumptions hold for system and simulation data is important because, in general, tests making less assumptions, e.g. nonparametric tests, can be more widely applied, but relate to less statistical power. However, although higher statistical power is normally to be preferred, within this setting it may not be; a test with less power might actually be favored because some inaccuracy between a simulation model and the real system is acceptable (a difference always exists).

The applicability of the methods and the legitimacy of selections made will be evaluated within the subsequent chapter.

# 3. Practical Findings

For the development of the operational validation approach for Vanderlande Industries, the methods which are best suited for application within the context of Vanderlande Industries should be selected to form a practical operational validation approach. In this chapter the design proposed in the previous chapter will be evaluated by means of the baggage handling system selected in the case study. Practicalities will be discussed that became apparent when performing the methods proposed in the previous chapter, as part of the case study. Practicalities can imply general positive or negative findings related to certain proposed methods and their assumptions, or additional minor steps that turned out to be required, but not yet have been discussed.

No extraordinary findings have been observed while creating an assumptions document. Therefore, this step will not be incorporated in this chapter. Consequently, section 3.1 will address input validation. In section 3.2 trace-driven output validation will be evaluated and in section 3.3 practicalities related to the sensitivity analysis will be discussed. The general findings will be discussed in section 3.4.

## 3.1 Input Validation

In order to validate input distributions for the arrivals of baggage items, proper samples should be selected. Because simulation models are generally developed for evaluating system performance under peak loads, input distributions are defined accordingly. Consequently, for validating input distributions data should be used that represents peak loads. It should be noted that these peak loads are related to the total system arrival process. When evaluating peak loads per input line, a total system capacity may be required that greatly exceeds real total peak loads. Once the system peak loads have been determined, they should be separated into samples of individual input lane arrival processes.

For the purpose of getting an idea of the arrival process, the total system arrivals should be visualized first. Once the total system peak loads are identified, the related input distribution samples can be acquired. In order to be able to evaluate the trend of the arrival process properly, a moving average will be used of the arrival rate. A moving average, also called a rolling average, rolling mean, or running average, is a technique to smooth out short-term fluctuations and highlight longer-term trends or cycles (NIST/SEMATECH, 2010). Although the variable evaluated is the interarrival time of baggage items, it is not used as the dependent variable in the moving average plot. Instead the variable is converted into the capacity level that it requires. This way the moving average plot can be easily related to system capacities, in order to assess whether the highest observed capacity requirements can be considered a “real” peak load. For similar reasons the x-axis indicates the time of the last arrival of the moving average subset. Timing is more meaningful than the number of the last arrival. Furthermore, this makes the moving average plots of different days directly comparable. As a downside, changing the x-axis results in the moving averages not being equally distributed over the axis. As it turned out, system capacity requirements were highly dynamic. Therefore, a relatively large moving average length was selected (100 bags). Also, various alternative moving average lengths have been evaluated in order to identify the relation of a sample size to a peak load (up to 2000).

Unlike the moving average technique described, simulation engineers at Vanderlande Industries often base capacity requirements on specific time intervals. The benefit of a constant time interval is that it makes a plot easy to interpret; an output can be directly related to the time it occurs. Furthermore, measuring points are distributed equally over the total amount of time. Though it should be noted that using a constant time interval may lower the peak results; it is unlikely that a peak load exactly matches a time interval. This downside can be partly overcome by creating a hybrid; a constant time interval can be used which is not shifted completely, but with very small steps (as with the moving average, where it is shifted by one bag at the time). This hybrid can be observed on the right hand side of Figure 3, where the time interval is 38 minutes (which is on average equal to 1250 bags at a peak load) and the interval is shifted per minute. However, it is more difficult to relate the peak load to a sample size, in order to evaluate whether it is sufficiently large.

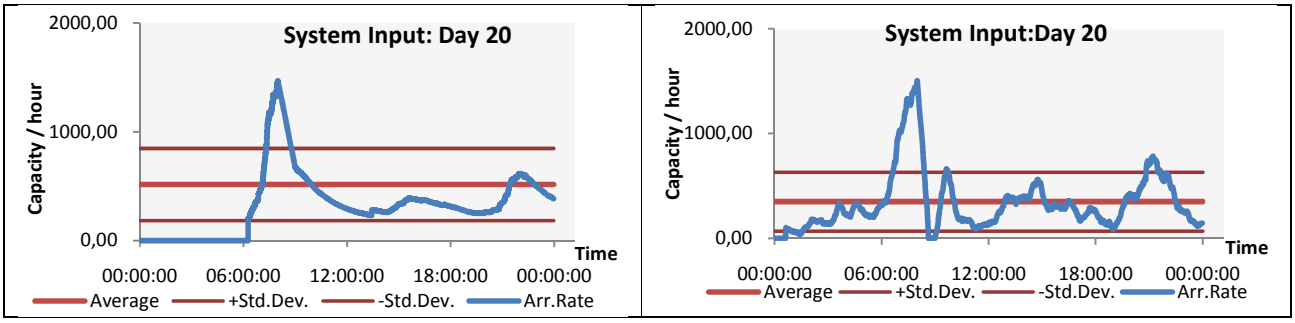



Figure 3: Comparison of case-based moving average to time-based moving averages (1250 cases ~ 38 minutes)

Now that peak loads can be identified, the next issue is which bags to include into a sample and which not. In order to increase statistical power the sample size should be as large as possible. On the other hand, from Figure 3 it can be observed that the arrival process is nonstationary. Therefore, a too large sample will result in non-identical and dependent observations, which violates statistical tests' assumptions. Another disadvantage of an increasing sample size in combination with a nonstationary process is that it will decrease the average capacity requirements of the sample, while a high system capacity is pursued. This is because data with a higher average interarrival time is added to the sample. Plotting the highest observed systems capacity against the used sample size can give insights into this effect. Such a graph is called a sample response plot in this report.

Within the case study it turned out that the interarrival times of bags did not form a random sample of a normal distribution. The process is better described by an exponential or a lognormal distribution. This implies that various tests relying on the normality assumption are not applicable, such as the standard linear regression based methods for stationarity and independence, ANOVA, Bartlett's test, and the F-test. Furthermore, the non-parametric Von Neumann's ratio test and run tests were not applicable due to the high amount of ties in the data. Arrival times were measured per second. With over 60 000 arrivals of bags and an interarrival time mainly ranging from 2 to 15 seconds it can be imagined that many interarrival times had the same value. Consequently the assumption is violated.

Determining which system samples were sufficiently similar in order to assume that they originated from the same distribution was very well possible with the proposed techniques. The techniques described to assess goodness-of-fit on the other hand, appeared to be too sensitive for modeling purposes.

Differences were found between the data and distributions, even for the best fitted distributions. However, it is possible to apply goodness-of-fit tests in order to make an objective comparison between alternative distributions possible. For instance the best fitting distribution will result in an upper limit of the accuracy of an input distribution.

Several statistical packages have been evaluated for applying the various tests, namely Statgraphics, PASW SPSS, Palisade @Risk/Stattools, and Expertfit. As it turned out, graphical techniques were most easily applied within Palisade @Risk/Stattools and Expertfit. Palisade @Risk/Stattools was also very suitable for performing the proposed tests related to comparison of simulation and system input. Furthermore, data independence could be investigated with aid of the correlation plot and Pankrantz criterion. The alternative, the Box-Ljung test, is only applicable within PASW SPSS. For conducting statistical tests for the comparison of two or more simulation samples Statgraphics or PASW SPSS is required.

## *3.2 Trace-Driven Output Validation*

In order to perform trace-driven output validation the simulation model should be adapted to acquire interarrival times from an external data source. This data source can be a text or an Excel file, derived from the real system's logging.

The sample mean, variance and sample correlation, as well as Box plots and frequency diagrams/histograms have been proposed for comparing the output of the simulation model and the real system. However, some statistics may be added. In real system data very high cycle times can be observed, for instance due to errors when registering that a bag leaves the system. These values highly influence the sample mean. As identified earlier, if it is suspected that outliers are present or the data is skewed, samples are more appropriately compared by their median (Green & Salkind, 2004). Additionally, the trimmed mean can be used, which only takes into account the trimmed probability density function. As an example, the 5% trimmed mean is the average value of the middle 95% of the probability density function.

As it turned out, the sample correlation function is not directly applicable. Individual observations cannot be related to each other for both samples because of a significant number of bags no cycle times are available in the real system (bags go out of tracking). Furthermore, due to stochastic behavior, bags might belong to different classes in the simulation model (normal baggage, early baggage, unsafe baggage). In order to overcome this problem, the correlation can be based upon an average output per period. For assessing the correlation Pearson's coefficient is generally preferred. However, when outliers are present or the data is skewed the alternative, Spearman's coefficient, should be used. It computes coefficients based on the ranks of the data rather than on the data values themselves (Statpoint Technologies, 2009).

A difficulty with histograms is that no definite technique exists for determining interval widths. However, some guidelines can be given. It is generally recommended to choose the smallest interval width that results in a "smooth" diagram. Too small intervals will lead to a "ragged" shape because the variances will be large. Too large intervals, on the other hand, will result in a "block-like" shape since the data has

been overaggregated (Law & Kelton, 2000). Some additional rules of thumb have been suggested, but it should be noted that they will not result in an optimal outcome for every situation. The best known rule is probably Sturges's rule, which says that the amount of intervals should be equal to $1 + \log_2 n$, with n equal to the amount of observations (Engel, 1997). A different rule states that the interval width should be equal to 0,3 times the standard deviation of the data (NIST/SEMATECH, 2010). However, when performing the case study these rules appeared to result in too large interval widths. Therefore, the general guidelines are preferred. Furthermore, it should be noted that relative frequency diagrams are preferred over regular variants. This is due to the differences between the numbers of cycle times that will possibly be obtained. Differences are caused not only by disparities in parameter values, but also by bags going out of tracking. These bags cannot be traced, and consequently no cycle times can be obtained from this subgroup.

Behavior graphs are also very valuable. In section 3.1 it has been shown that the arrival times of bags are not stationary; the average arrival rate changes over time. Changes in the arrival rate relate to alterations in the utilization of the system. Certain assumptions, for instance about merging behavior, are affected by the utilization of the system. They may suffice when the utilization is low, but become more and more problematic when the utilization increases. For the example of merging behavior: a possible assumption is that merging two flows does not lead to any interruption. With a low utilization the probability of being blocked at a merge is little and the assumption suffices. However, in case of a high utilization blocking becomes more likely, causing a significant interruption. Generally, an increase in utilization is expected to increase the difference between simulation and the real system output. As a result it is not the overall average difference that is of main interest, but the average difference per range of utilization. The relatively high arrival rates occur only on a very limited basis. Consequently, the techniques for output validation that require a high amount of data may be problematic. Specifically the regression based methods are vulnerable for nonstationary differences.

The optimal outcome would be an indication of a causal relation between utilization and the average difference between simulation and a real system, such that the relation can be used to estimate the difference for utilization rates not observed. Behavior graphs can be used to gain insights into this relation. When applying behavior graphs it becomes apparent that the results are too variable to see a general trend. Therefore, the output should be smoothened out, as was the case for the arrival process. An additional issue is that system and simulation output require an identical independent variable in order to make an accurate comparison possible. The most evident independent variable is the time of occurrence. However, the aim is to investigate whether a causal relation can be found between utilization and the average difference. Because the utilization of the system is not directly known, and is not necessarily identical to the utilization of the simulation model, the arrival rate can be used as an alternative. A difficulty with this independent variable is, however, that the expected effect on the output difference is delayed (bags will not affect cycle times until they are being processed). Combining the knowledge that results should be smoothened and that possible effects are delayed leads to usage of the average result per period. Furthermore, high order polynomial regression lines can be added to these behavior graphs to underline the trends.

The methods recommended within this section are available in all regular statistical packages.

## 3.3 Conducting a Sensitivity Analysis

In order to be able to conduct a sensitivity analysis the simulation model should be adapted such that AutoStat is able to alternate factor values. This implies that all factors should be modeled as variables within AutoMod. Hereby it should be noted that, though AutoStat adapts decimal placeholders to regional settings, AutoMod does not. Thus whenever regional settings are set on Dutch, AutoStat sends messages to AutoMod possibly containing values based on a comma, while AutoMod expects decimal values based on a dot. Consequently, the analysis will not run and an error is returned.

The amount of replications required for a sensitivity analysis can be greatly reduced by applying CRN (section 2.5.7). Although this method can generally be applied within AutoStat, it is not available for DOE (after consulting AutoMod developers it was assured it will be available within the upcoming version of AutoStat). However, it can be applied manually. In an effort to force AutoStat to perform CRN, distribution samples have been created within Excel. This has been done by using the built in random number generator (resulting in a random number between 0 and 1), and converting it to a distribution seed by applying the cumulative distribution function (appendix B.5.1). Identical random numbers have been used for the alternative distributions. The simulation model should be adapted to use the generated data as input. Since the model is already configured to acquire interarrival times from an external source for trace-driven output validation, CRN can be forced with relatively little effort. Therefore this procedure is recommended.

No additional software package is required for performing a sensitivity analysis; all recommended actions can be performed with AutoStat. However, AutoStat is only capable of handling up to 11 factors in a design of experiments. It is possible to create combinational factors. Consequently, screening methods have to be applied. Probably the most self-evident factors to combine are the velocities and window lengths of the various section types. It reduces the amount of factors greatly, while the direction of the effects is expected to be identical for all parameters within the groups.

The design of experiments requires two levels per factor. One level, the base level, has already been specified while the simulation model was configured (appendix B.1.1). Concerning the alternative level, two options are available: values can be used based on system logging, in which parameter values can be identified, or somewhat more extreme values may be used based on expert opinions for example. In the case study the first option is chosen. The sensitivity analysis is performed in order to give more insight in the effects of observed differences between the simulation model and the real system. However, even when basing alternative values on data exact values are not straightforward. A balance should be sought between the difference between the factor levels (larger differences are expected to result in more significant effects) and the likelihood that the values are plausible within the real system. Therefore, the system data has been evaluated per day when possible; the most deviating value has been selected. It should be noted that not all factors could be identified in the system logging, e.g. conveyor velocities. For these exceptions the alternative value has been determined by evaluating expert opinions.

Note that setting the base level to values in the simulation model and the alternative level to values mainly based upon data can imply that some values are descending from the low to the high level, while

others are ascending. Normally factor values are put in an ascending order, in order to facilitate interpretation. Though, this is less practical in this situation. The experiment results in estimated factor effects for changing the factors from the base level to the alternative level. Thus, effects can now be interpreted directly as the result of a modification to the standard simulation model.

However, this method has consequences for applying the group screening method. The method prescribes to put factors in a group such that effect directions are aligned. As it will turn out, though the combined velocities as well as the combined window lengths are likely to fulfill this requirement, EBS control is not. Therefore cancellation of effects is possible or even probable. However, it can be justified that this is not a real threat. The various factors concerning EBS control are linked to each other in practice. In fact some cannot be changed separately; the time a bag is delayed by the EBS is virtually only dependent of the amount of time the bag arrived too early. Consequently, reducing the time to flush by 50 % implies that the amount of flush-backs should approximately double. The effect of changing EBS control to a more real life example is of main interest. Therefore possible cancellation of effects is accepted.

Evaluation of factor significance can be done by creating confidence intervals of factor effects. A 95 % confidence interval is equivalent to a hypothesis test with a type I error of 5 %. Applying confidence intervals is an option within AutoStat.

With respect to validation of the sensitivity analysis, in AutoStat it is not possible to eliminate individual scenarios and recalculate the factor effects. Therefore, cross-validation cannot be performed. Instead new scenarios should be run and simulation output should be compared to metamodel prediction by calculating the correlation, or eyeballing a scatter plot (Kleijnen, 2005).

## *3.4 Discussion*

Within this chapter the applicability of the methods and the legitimacy of selections made in the previous chapter have been evaluated. Also, additional actions and practicalities have been addressed.

As it turned out, the assumptions related to real system and simulation model data that were used in literature for evaluating suggested methods, do hold for BHS models of Vanderlande Industries. Although this implied that less powerful nonparametric tests should be used, statistical methods for comparing system and simulation input still appeared to be too sensitive. However, the statistics did appear valuable for making an objective comparison between the fit of alternative distributions possible.

With respect to output validation, confidence intervals are preferred, but do result in additional requirements related to data characteristics. Although the requirement of a high amount of data per input range was not met within the case study, it might be available in other future projects. Consequently, it can still be noted as an option within the operational validation approach.

The possibilities related to the sensitivity analysis depended heavily of the available functions within AutoStat. However, although somewhat affecting the selection of methods, it did not change the recommendations based on literature greatly.

# 4. The Operational Validation Approach

In this chapter the operational validation approach that has been designed based on theoretical and practical findings will be presented. The various steps that should, or may, be performed will be explained and the accompanying selected techniques will be addressed. Hence, this chapter addresses the main research question. For more in depth knowledge about a specific selection the reader will be redirected to the appropriate approach development section. Illustrations of the recommended techniques can be observed in the case study (Chapter 5).

This chapter is structured as follows. In section 4.1 a schematic overview of the approach is presented. From this overview it can be derived that the operational validation approach has been divided into several high level steps, each related to a different aspect of the validation problem. The first step, maintaining an assumptions document, is addressed in section 4.2. Section 4.3 discusses the second step, called input validation. Trace-driven output validation, the third step, is handled in section 4.4. Finally, the last step, conducting a sensitivity analysis, is discussed in section 4.5.

## 4.1 Approach Overview

A schematic overview of the approach can be observed in Figure 4. Attached to the various steps are the different techniques that could be applied. The reader is being referred to specific sections for more in depth discussions about specific methods.

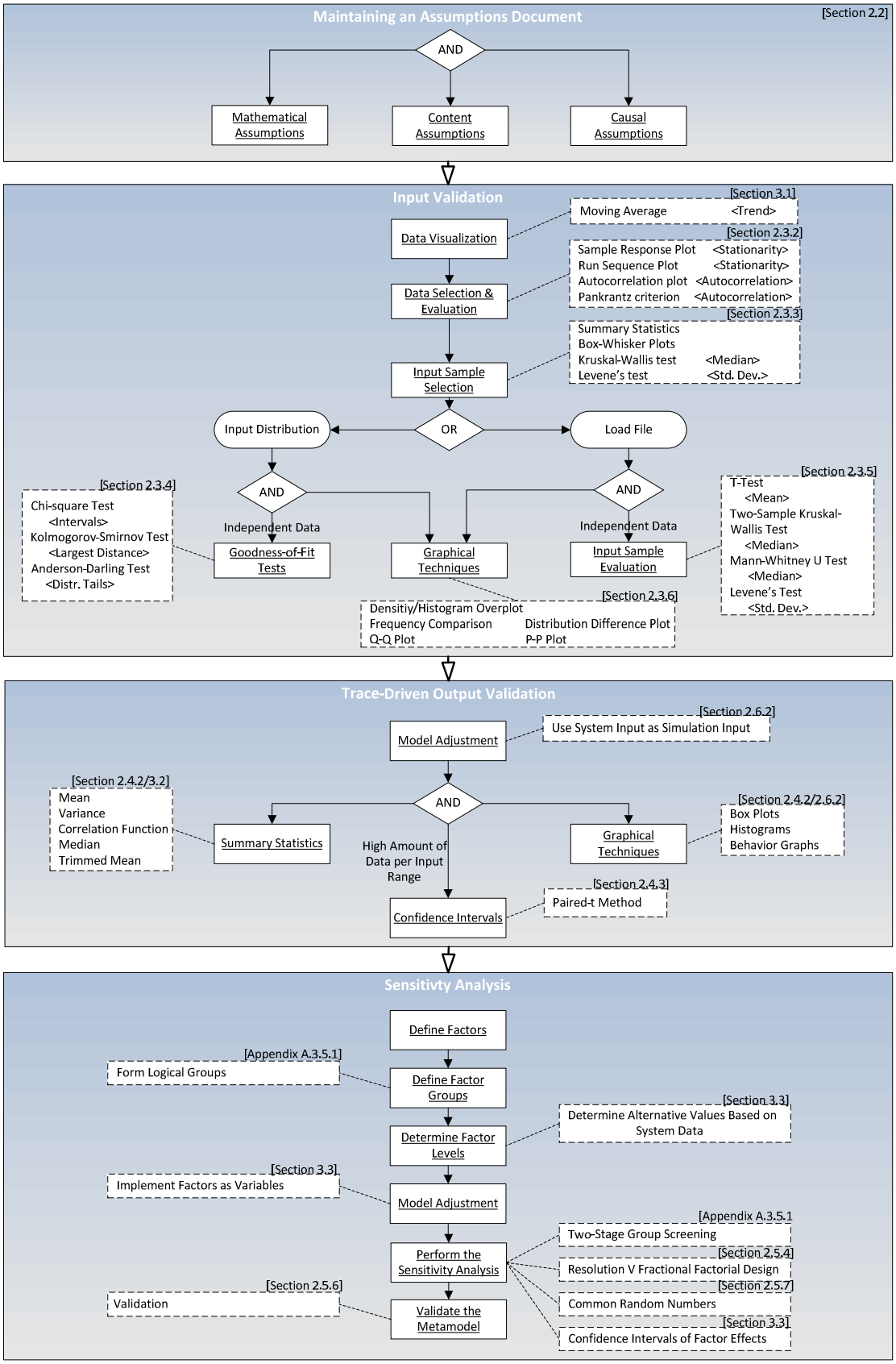



**Figure 4: A schematic overview of the operational validation approach**

## 4.2 Maintaining an Assumptions Document

The first step to perform is to maintain an assumptions document. The document should provide information about the input values, the assumptions made and the level of detail contained in the simulation model. This will give an indication where differences with reality can be expected. Furthermore, it can be used to evaluate the degree of similarity between simulation models; a pattern may be observed between certain assumptions and their impact on model accuracy. The types of assumptions identified are: “mathematical assumptions including the model form and continuity of the relationships, content assumptions dealing with the scope and definition of model terms and variables, and causal assumptions concerning assumed or hypothesized relationships between terms and variables” (Gass, 1983, p. 612). Examples of assumptions used in the case study can be used to illustrate the different types. A used mathematical assumption is that the underlying model used to estimate factor effects in the sensitivity analysis is linear. A content assumption is that the velocity of a standard belt floorveyor is 1 m/s. An example of a causal assumption is that an increase in utilization generally leads to an increase in the difference between simulation output and system output.

## 4.3 Input Validation

The second step is input validation. Validating the input used in the simulation model separately is required because it will be substituted by real system input data in the output validation step. Furthermore, it is also valuable for the trace-driven output validation; an idea is developed about the behavior of the real system input. This is required in order to be able to evaluate the options available for output validation.

### Visualization

Related to this, the first action to perform is to visualize the arrival process. This is especially relevant for baggage handling systems, because its arrival process is typically highly nonstationary (the mean, variance and autocorrelation function of the interarrival times shifts over time). Because individual observations are also highly dynamic a moving average should be used. A moving average is a technique to smooth out short-term fluctuations and highlight longer-term trends or cycles. Given a series of numbers and a fixed subset size, the average of the first subset is calculated. The fixed subset is moved forward one number and the new average is calculated. This process is repeated over the entire data series. The line connecting all the acquired averages is the moving average (NIST/SEMATECH, 2010). Although the variable evaluated is the interarrival time of baggage items, it is not recommended as the dependent variable in the moving average plot. Instead the variable should be converted into the capacity level that it requires. This way the moving average plot can be easily related to system capacities, in order to assess whether the highest observed capacity requirements can be considered a “real” peak load. In order to make the arrival process of different days more comparable, the x-axis may be converted to a time based scale. A disadvantage is that it results in the moving averages not being equally distributed over the axis. An alternative moving average based on time intervals is discussed in section 2.6.1.

### Data Selection and Evaluation

In case a peak load should be identified, a sample response plot is proposed in order to observe output changes over a range of samples sizes. For baggage handling systems it is useful to plot the highest observed system capacity against the used sample size. When increasing the sample size, a high negative slope indicates that the added data differs considerably from the current sample. Contrarily, a gradual decline implies that the added data differs little from the current sample. In this manner it is possible to assess quickly which sample sizes may be viable options for representing the peak load.

Viable data sets should be investigated for stationarity. Nonstationary data cannot be represented by a single distribution in the simulation model. Furthermore, no statistical tests can be applied. Stationarity can be determined with a run sequence plot (NIST/SEMATECH, 2010). Run sequence plots are an easy way to summarize a data set. The graph is formed by plotting the response variable on the vertical axis, and the observation index to the horizontal axis. Shifts in location and scale are typically quite evident. Furthermore, outliers can easily be detected using a run sequence plot. For a constant location and scale the response should appear constant +/- a random error (NIST/SEMATECH, 2010). Although being very useful for changes in mean and variance, the run sequence plot is not suitable for addressing autocorrelation.

Instead the autocorrelation plot is recommended. The autocorrelation plot is a graph of the sample autocorrelations for data values at varying time lags. A lag is a fixed time displacement: a plot of lag 1 is a plot of the values $Y_i$ versus $Y_{i-1}$ (IST/SEMATECH, 2010). It is important to note, that the indicated sample correlation will not be 0 even when the data is independent. Only if the data differs significantly from 0, strong evidence of correlation exists. Another indication is a specific pattern in the correlation plot, such as a linear trend (Law & Kelton, 2000). In order to assess more formally whether the autocorrelation is too high, the graph can be extended with the Pankrantz criterion. It states that the autocorrelation divided by its standard error must be less than 1,25 for the first three lags and less than 1,60 for subsequent lags, in order to conclude that the series is not significantly autocorrelated.

### Input Sample Selection

Now that various suitable data sets have been acquired, it should be decided how to compare the samples to the simulation input. Comparison of simulation data with system data of as many periods as possible is preferred, because this will increase the generality of the findings. Two options exist for incorporating multiple samples in the comparison: simulation input can be compared with each system sample separately, or simulation input can be compared to a group of similar system samples. In principle this last option is preferred. A larger sample size will decrease both the probability of a type I error (risk of rejecting a true hypothesis) and the probability of a type II error (risk of accepting a false null hypothesis) (Montgomery & Runger, 2002). When, during data selection and evaluation, it became apparent that data is significantly autocorrelated, simulation input should be compared graphically only, with each system sample separately.

In order to test whether various samples originate from the same distribution based on their median the Kruskal-Wallis test should be applied. The median is used instead of the mean to be able to cope with non-normality. This is especially appropriate if it is suspected that outliers may be present or the data is

skewed. In order to test whether various samples originate from the same distribution based on their standard deviations Levene's test is recommended. The test can be based on various metrics. Using the trimmed mean performs best when the underlying data follows a heavy-tailed distribution and the median performs best when the underlying data follows a skewed distribution. Using the mean provided the best power for symmetric, moderate-tailed, distributions (NIST/SEMATECH, 2010).

### Comparison of System Input and Simulation Input

The next action is to actually perform the comparison of system input data and simulation input. The techniques to apply depend on whether a distribution or a load file is used as simulation input. Two types of techniques can be used: statistical tests and graphical methods. While the statistical technique to apply depends on the type of simulation input, similar graphical techniques can be applied for both input options. Furthermore, graphical methods make no assumptions about data independence and therefore can always be applied.

A goodness-of-fit test is a statistical hypothesis test that is used to assess formally whether observations are an independent sample from a particular distribution (Law & Kelton, 2000). Several goodness-of-fit tests are recommended. The Kolmogorov-Smirnov test and Anderson-Darling test provide good statistical power. They compare the empirical distribution function of the sample with the cumulative distribution function of the hypothesized reference distribution. Both are suggested because they focus upon a different area; the Anderson-Darling test is designed to specifically detect discrepancies in the tails, while the Kolmogorov-Smirnov is more sensitive to differences near the center of the distribution. Applying both methods will require virtually no additional effort. The chi-square test is less appropriate, but, contrary to the other methods, it is not limited to specific distributions (i.e. normal, exponential, weibull, lognormal, and log-logistic distributions). Although the statistical techniques presented are inclined to be too sensitive for modeling purposes, they can facilitate an objective comparison between alternative distributions. This is useful because the best fitting distribution will for instance result in an upper limit of the accuracy of an input distribution.

Related to comparing a load file to a system sample, the most powerful test is the Student's t-test for equality of means. Although assuming that samples are normally distributed, Rasch et al. (2007) state that the t-test is robust against this assumption to such an extent that it can be recommended in nearly all applications. However, again more tests can be recommended due to a difference in focus, while requiring little effort. The two-sample Kolmogorov-Smirnov test is an adaption from the one-sample test discussed earlier. The Mann-Whitney U test is an alternative to the t-test without its limiting assumption of normality (Cooper & Schindler, 2003). It is a rank test that evaluates whether the medians on a test variable differ significantly between two samples (Green & Salkind, 2004). Related to equality of variances, Levene's test is once more recommended.

Several graphical procedures can be used for comparing fitted distributions with the true underlying distributions, as well as for the comparison of two data samples. In general a close match of the data and the distribution, or both data samples, implies a better fit. For continuous data, a density/histogram overplot can be made by plotting the probability density function of the applied distribution over the histogram of the data. Alternatively, a frequency comparison can be used, which compares the intervals

of two histograms. These can be based on one data sample and a distribution, as well as on two data samples. Furthermore, graphical comparisons can be applied that are based upon the cumulative distribution functions. Such a technique is the distribution function difference plot (Law & Kelton, 2000). A perfect fit will result in a horizontal line at height 0; the greater the vertical deviation, the worse the fit (Law & Kelton, 2000). However, most statistical packages simply plot both cumulative density functions in one graph, from which the difference can be derived by eyeing the plot. Another graphical technique based on the cumulative distribution function is the probability plot. Several kinds of probability plots exist. The Q-Q plot (where Q stands for quantile) compares the probability distribution by plotting their quantiles against each other (Gibbons & Chakraborti, 2003). The distributions being compared are similar if the point in the Q-Q plot will approximately linear with an intercept of 0 and a slope of 1 (Gibbons & Chakraborti, 2003). Q-Q plots amplify differences that exist between the tails of the compared distribution functions (Law & Kelton, 2000). A different probability plot, the P-P plot (where P stands for probability or percent), on the other hand, amplifies the differences between the middle parts of the distribution functions (Law & Kelton, 2000). This difference can be observed in Figure 5. The Q-Q plot graphs the different values $x^S_q$ and $X^M_q$ for different quantiles q. The P-P plot graphs the model probability $\hat{F}(p_i)$ versus the sample probability $\tilde{F}_n(p_{(i)})$ for different values of $p_i$ with i = 1, 2, .., n. Note that even if the correct distribution has been used, or both input samples are identical, there will be departures from linearity for small to moderate sample sizes (Law & Kelton, 2000).

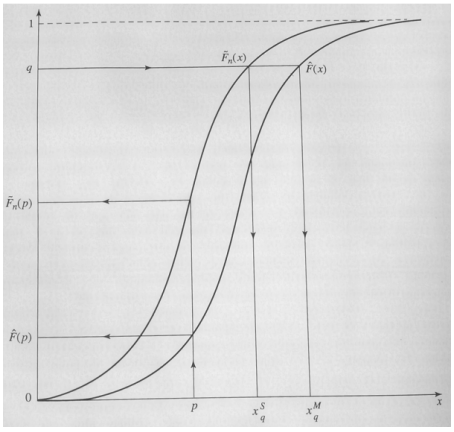



**Figure 5: Definitions of Q-Q and P-P plots (Law & Kelton, 2000)**

## *4.4 Trace-Driven Output Validation*

The third high level step is to validate trace-driven simulation output. This can be considered the main step of the validation approach; based on these results it is concluded whether the main model is valid or not. This is done by assessing the input-output transformation.

### Model Adjustment

In order to perform the validation methods that will be described, the model should be adjusted such that it can use system input data as input for the simulation. Furthermore, an input file should be created based on exact arrival times of bags or products. With these adaptations trace-driven output can be generated.

### Output Comparison

For comparing trace-driven simulation output with system output the correlated inspection approach should be used. The correlated inspection approach does not require the data to be independent, has no distributional requirements on the data, and can be used with a limited number of observations (Sargent, 1996). It prescribes to compare the sample mean, the sample variance or standard deviation, the sample correlation function and to apply several graphical plots, namely histograms, box plots and behavior graphs (Sargent, 1996; Kleijnen, 1995). Additionally, the median and the trimmed mean are recommended, which are especially relevant in case outliers are present.

For the construction of histograms it is generally recommended to choose the smallest interval width that results in a "smooth" diagram. Too small intervals will lead to a "ragged" shape because the variances will be large. Too large intervals, on the other hand, will result in a "block-like" shape since the data has been overaggregated (Law & Kelton, 2000).

A detailed explanation of Box-Whisker plots can be found in Figure 6. It can be added to this information that quartiles are the points from the cumulative distribution function that divide it in four regular intervals (Montgomery & Runger, 2002). Note that the second quartile is equal to the median.

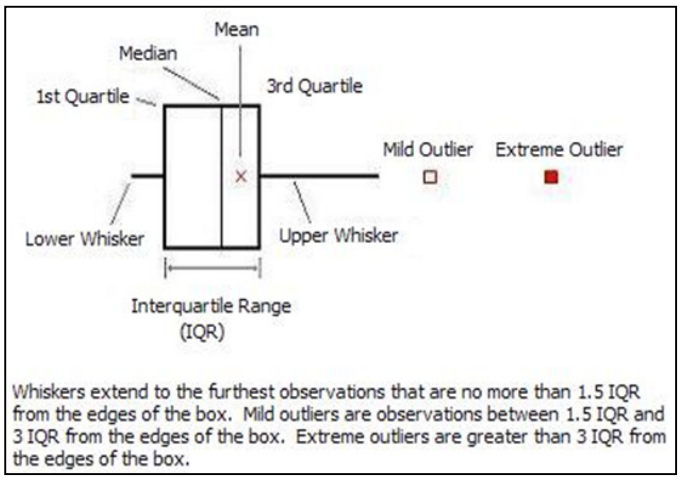



**Figure 6: An explanation of a Box-Whisker plot (Statpoint Technologies, 2009)**

For assessing the correlation Pearson's coefficient is generally preferred. However, when outliers are present or the data is skewed the alternative, Spearman's coefficient, should be used. It computes coefficients based on the ranks of the data rather than on the data values themselves (Statpoint Technologies, 2009). Correlation coefficients range between -1 and +1 and measure the strength of the association between the variables. When individual observations of both samples cannot be directly related to each other, for example due to bags going out of tracking, average output per period should be used.

In section 2.6.2 it has been shown that this also holds for behavior graphs. With behavior graphs one attempts to clarify how the system and simulation output behave in time and with respect to changes in input. Responses per period should be used because the effect that the arrival rate has on cycle time is delayed in time. Furthermore, the interval smoothens out the high frequency oscillations so that trends

can be observed. When trends are still difficult to identify, high order polynomial regression lines can be added.

For these methods, determining whether the model has sufficient accuracy for its intended purpose should be done subjectively. This comparison can be made by the model development team and/or by experts using face validity or Turing tests (Kleijnen, 1995; Sargent, 1996). A Turing test is performed by asking people knowledgeable about the system to examine and identify one or more sets of system data as well as one or more sets of model data without knowing which sets are which (Law & Kelton, 2000). Related to this, a model is said to have face validity when simulation results are consistent with perceived system behavior (Law & Kelton, 2000).

When input data is stationary over a relatively long period, confidence interval can be constructed per range of input. The paired-t method should be used due to the dependence between simulation and system output. Although the method relies on the normality assumption, it is quite robust for deviations (Kleijnen, 1995). However, it does require observations to be independent and identically distributed. Because this is not the case for individual observations, average output per period should be used for this method as well. Additionally, this is required because the paired-t method can only be applied to equal numbers of observations.

A 100(1 – α) percent confidence interval is statistically significant at level α in case the interval does not contain 0. When it does contain 0 any observed difference may be explained by sampling fluctuation (Law & Kelton, 2000). However, differences are to be expected and do not necessarily imply that the model is invalid. Therefore practical significance is defined as the magnitude of the difference being large enough to invalidate any inferences about the system that would be derived from the model (Law & Kelton, 2000). As for the inspection approach, the decision whether the difference between the model and the system is practically significant, is a subjective one, and should be decided on by the model development team or expert.

## *4.5 Sensitivity Analysis*

A more detailed comparison can be made by conducting a sensitivity analysis. It can generate insights into what causes differences that may be observed when evaluating simulation and system output. However, it may also be of interest when no practical significant differences are found; the absence of differences in an output analysis does not necessarily imply that no differences will be found in the more detailed sensitivity analysis. It merely indicates that the combination of these possible differences does not have a significant impact on the assessed output. Furthermore, a sensitivity analysis indicates the system's robustness to changes in variables.

### Define Factors

When performing a sensitivity analysis the first step is to define the factors to evaluate. Important factors to consider are factors for which differences are expected or observed in the system's logging. Correspondingly, factors related to assumptions made in the simulation model are likely candidates. Examples of factors that could be investigated are: the value of a parameter, the choice of a distribution,

the entity moving through a system (e.g. a single item or a batch), and the level of detail for a subsystem (Law, 2008).

### Define Factor Groups

In order to be able to evaluate factors relatively efficiently, and because AutoMod is only capable of taking up to 11 factors into account in one experiment, two-stage group screening should be applied. For two-stage group screening the experimenter uses experience and knowledge of the problem and the factors to arrange the factors into logical groups (Trocine & Malone, 2001). A fractional factorial design is run on the groups in order to indentify the important ones. Subsequently, a new fractional design may be run on the factors or subgroups within an important group until the important factors are identified. The method is iterative since the results of the first stage are used in the second stage. Note that interactions between factors in different groups are not measured and if they exist may confound the results of the groups (Trocine & Malone, 2001).

In order to avoid cancellation of factors and to detect as many of the effective factors as possible Ivanova et al. (1999) identified several guidelines.

- A factor with an unknown direction of effect should be placed alone in a group.
- Factors with assumed important positive effects should be placed in one group.
- Factors with assumed small effects and the same direction should be placed in a group.
- Factors with possible effects and the same direction should be placed in a group.
- Resolution V designs should be used to calculate main effects and two-factor interactions unbiased by any other main effect or two-factor interaction (section 2.5.4)

Contrary to these suggestions, in some situations cancellation of effects can be acceptable, because factors are linked in practice for instance. When such groups are selected it should be kept in mind though that, based on perceived group effects, no conclusions can be drawn for the corresponding individual factors. A more detailed discussion about this subject can be found in section 2.6.3.

### Determine Factor Levels

The design of experiments requires two levels per factor. The standard levels should be set equal to the parameter values of the simulation model. The alternative levels should be identified in the real system's logging. In case no equivalents of parameters can be identified in the logging, the alternative levels should be based upon simulation engineer's experience.

### Model Adjustment

In order to be able to conduct a sensitivity analysis the simulation model should be adapted such that AutoStat is able to alternate factor values. This implies that all factors should be modeled as variables within AutoMod. Furthermore, additional actions may be required, dependent on the version of AutoStat, when deciding to apply common random numbers (see the subsequent action). More details concerning these adjustments can be found in section 2.6.3

### Perform the Sensitivity Analysis

A resolution V fractional factorial design should be run on the factors and factor groups in order to indentify the important ones. Subsequently, a new fractional design can be run on the factors or

subgroups within an important group until the important factors are identified. Furthermore, usage of common random numbers is recommended.

Common random numbers (CRN) strives to compare alternative configurations under similar experimental conditions, so that any observed difference in performance is due to differences in the system configurations rather than to fluctuations of the experimental conditions (Law & Kelton, 2000; Farrington & Swain; 1993). CRN tries to induce a positive correlation by generating corresponding random variables across simulations from the same random numbers (Glasserman & Yao, 1992). This effect is illustrated by an example in Figure 7. Note that the idea of comparing a model and the corresponding system under the same statistical conditions is similar to the use of identical input in trace-driven output validation (Law & Kelton, 2000).

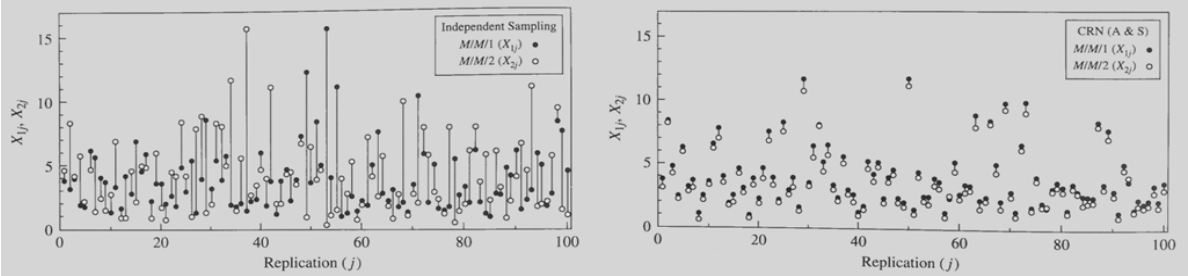



**Figure 7: An example of the effect of CRN, based on an M/M/1 and M/M/2 queue (Law & Kelton, 2000)**

In order to determine whether the observed factor effects are significant, confidence intervals should be constructed. A 95 % confidence interval is equivalent to a hypothesis test with a type I error of 5 %. Applying confidence intervals is an option within AutoStat.

### Validate the Sensitivity Analysis

In order to increase the confidence in the outcome of the sensitivity analysis, the results should be validated. Determining the validity of the estimated effects can be done by running new scenarios and comparing simulation output with predictions based on the sensitivity analysis. These predictions can be compared with the corresponding simulation responses using the Pearson linear correlation coefficient or comparing the responses through a scatter plot (Van Groenendaal & Kleijnen, 1997). For the last option it can be decided whether the estimated effects are acceptable by eyeballing the plot (the points will lie upon an approximately linear line with an intercept of 0 and a slope of 1) (Kleijnen, 2005).

# 5. Case Study

For illustration as well as evaluation of practicalities of the designed operational validation approach, a case study is performed. Practical findings that are based on this case study, which are related to the evaluation of the conceptual approach based on literature, have been discussed in chapter 3. The purpose of this chapter is to indicate what kind of results can be expected from applying the approach and to provide a reference for subjective actions and interpretations.

For performing a case study a baggage handling system was preferred. Parcel and postal solutions can generally be described as relatively simple processes. As such, a simulation study is normally not applied within these projects. For distribution simulations are becoming increasingly common. However, the main application area of simulation is still baggage handling. Other arguments for not selecting a distribution project are the enhanced complexity and wide variety of the models. Distribution solutions are very client specific; customer demands vary greatly, as well as the products to be handled by the system. Baggage handling systems, on the other hand, are far more standard; the input is similar (suitcases) with standardized dimension and weight requirements, the configurations are typical, and the design variation is constrained by international norms and regulations (Vanderlande Industries, 2009d). The combination of most simulations being performed for baggage handling systems and these systems being the most comparable, leads to insights based on the case study being relevant for the largest project base.

As baggage handling project Cairo International Airport Terminal 3 has been selected. This selection was based upon data availability, the presence of a simulation model, and the model not being too complex (as opposed to major projects such as Heathrow T5 and Schiphol Airport for instance). Due to international regulations, not all data may be shared, for instance relating to privacy issues. Furthermore, enhanced software needs to be installed at an airport, which allows for retrieving system data from location outside the airport.

This chapter is structured as follows. Within the first section a description of the baggage handling system will be given. The first high level step, maintaining an assumptions document, can be found in appendix B.2, because creating such a document is a relatively straightforward matter. The other validation steps will be addressed in section 5.2, section 5.3 and section 5.4. Lastly, a discussion of the case study results will be presented in section 5.5. The general configuration of the simulation model and standard simulation results can be found in appendix B.1.

## 5.1 Case Description

In April 2009, Terminal 3 of Cairo International Airport went “live” (Vanderlande Industries, 2009e). The departure system in Cairo consists of 10 check-in islands (110 check-in counters in total) and two transfer infeed lines. The check-in counters can handle a total of 4800 bags per hour. The baggage is identified at various locations by Automated Code Reader Stations (ACRS) that read the bag’s License Plate Code (LPC). When those stations fail to identify a bag, it is sent to a Manual Encoding Station (MES). All baggage is security checked by means of a 5-level security concept (Figure 8). In general, the 5-level

screening concept is a mixture of inspection by automatic screening devices (EDS) (level 1), common X-ray machines (level 3), manual judgment of the images by screening operators (level 2 and 4), and manual inspection of rejected bags (level 5) (Vanderlande Industries, 2009b). The time that operators have to make a decision about the security of an item coincides with the arrival of the bag at a Vertisorter. A Vertisorter redirects a bag in case it does not pass a security level; it is a sorting unit that sorts individual baggage items, arriving from one feeding conveyor, to two above each other installed take away conveyors (Vanderlande Industries, 2010) (Figure 9).

Figure 8: A security screening machine (Vanderlande Industries, 2009d)

Figure 9: A Vertisorter (Vanderlande Industries, 2010)

When baggage has passed the screening process is sent to two sorters. There the bags are automatically sorted to 33 laterals through both spiral and straight chutes. Laterals are accumulating conveyors, of which baggage is loaded onto transporting units. Chutes are slides that depend on gravity for the movement of items. The two sorters consist of Flat Triplanar Carousels with pushers. A Flat Triplanar Carousal is made of flame-retardant black slates that are mounted on aluminum carriers. The Divert Parallel Pushers are motor driven pushers installed aside of the carousal. They are started when a baggage item is aside of the pusher plate (Figure 11).

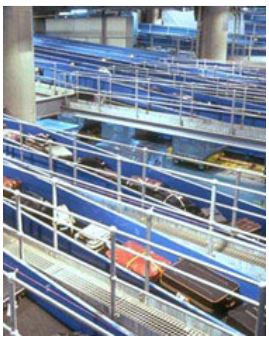


Figure 10: A lane storage system (Vanderlande Industries, 2009d)

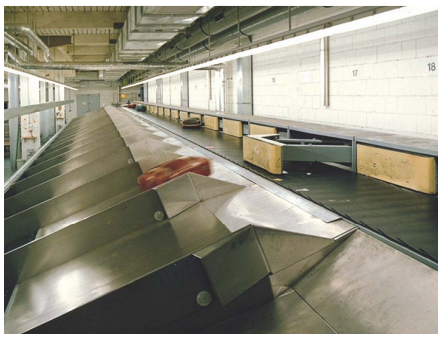


Figure 11: A flat Triplanar carousal with Divert Parallel Pushers and chutes (Vanderlande Industries, 2010)

Early baggage is temporary stored in the Early Bag Stores (EBS). Main parameters are capacity and storage/retrieval throughput. In Cairo Terminal 3 a lane storage system (Figure 10) is applied, consisting of accumulating belt conveyors. Bags can be categorized per flight, class or time-slot (Vanderlande Industries, 2009d).

Furthermore, the arrival system consists of 7 cresplanar carousels with corresponding feed lines and two out of gauge lines. The total system contains over 2,5 km of conveyors. Various back-up lines are incorporated to make all components redundant.

A complete schematic overview of the system can be observed in (Figure 12).

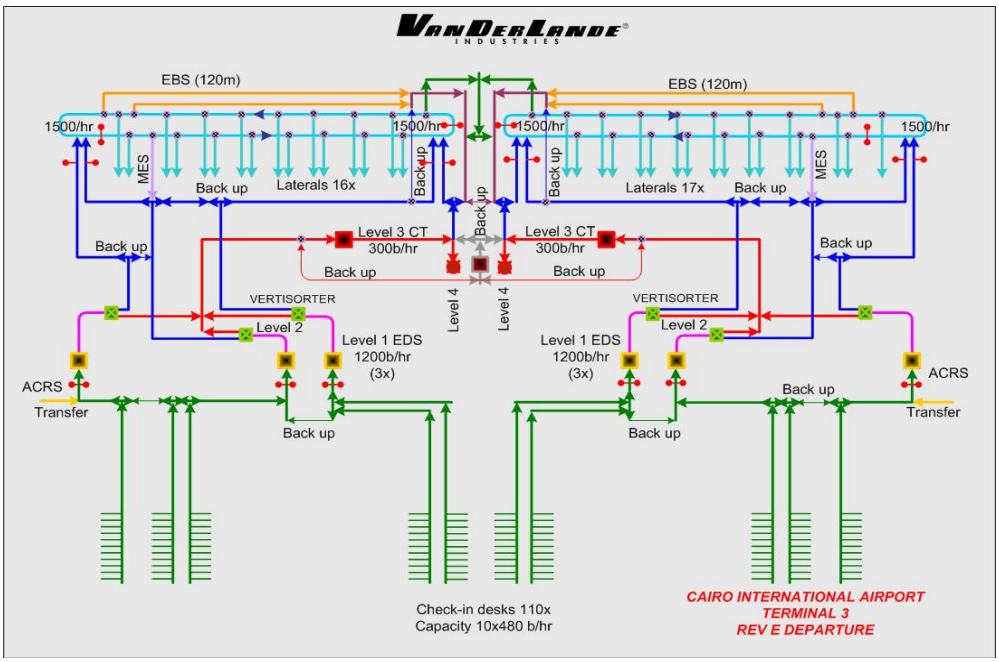



**Figure 12: An abstract drawing of the baggage handling system of Cairo International Airport Terminal 3 (Vanderlande Industries, 2009b)**

### Data availability

The BHS project is characterized by high tech features, for instance for system control. It was developed based on the London T5 software standards that were available at that time (Vanderlande Industries, 2009e). These enhanced software components made it possible to obtain the data required to apply the trace-driven validation approach. Due to airport regulations there is no data available from the check-in counters. As a consequence, the first time registrations are performed when a bag passes the first ACRS. Therefore the check-in desks and subsequent collector belts are considered out of scope of the system. Similarly, at the backend of the system, the last registration time is when a bag leaves the sorter onto a chute. As a result, also the laterals of the system are considered out of scope.

## *5.2 Validating Input Distributions*

The purpose of input validation is to evaluate how well probability distributions, which serve as random input variables, represent the true underlying distributions of the real system input data. Input distributions have to be validated separately because they have been substituted by real system data within trace-driven output validation.

### *5.2.1 Data Visualization*

For this case study system data has been made available of six successive days (20th - 25th day of the month). Graphs with a moving average of 100 bags can be observed of the 20th and the 21st in Figure 13. The x-axis denotes the point in time the last of 100 bags arrives, while the y-axis denotes the arrival

rate per hour. The figure contains the average and the standard deviation of the daily required capacity as well. The graphs of the other days are very similar and can be seen in appendix B.3.1. From the plots it can be concluded that a clear peak load can be identified every day.

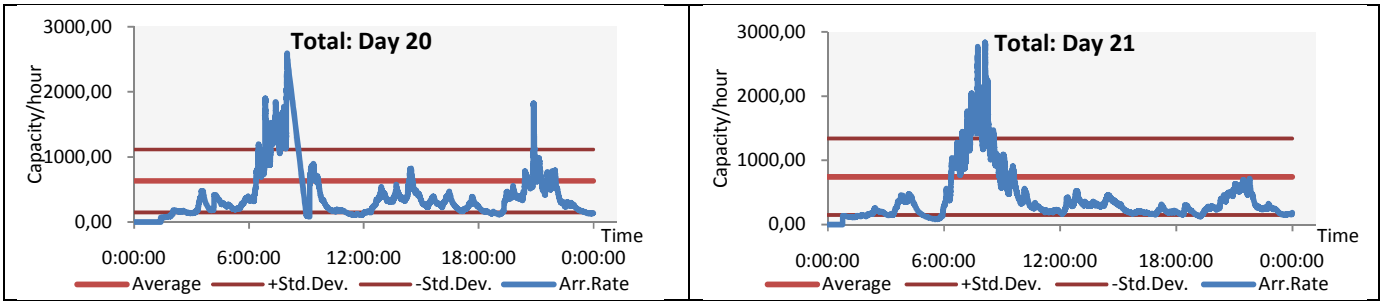



**Figure 13: System input capacity requirements, based on a moving average of 100 bags**

### *5.2.2 Data Selection & Evaluation*

Now that peak loads have been identified, the next issue is which bags to include into a sample and which not. In order to increase statistical power the sample size should be as large as possible. On the other hand, from Figure 13 it is known that the arrival process is nonstationary. Therefore, a too large sample will result in non-identical and dependent observations, which violates statistical tests' assumptions.

Another disadvantage of an increasing sample size in combination with a nonstationary process is that it will decrease the average capacity requirements of the sample, while a high system capacity is pursued. This is because data with a higher average interarrival time is added to the sample. This effect can be observed in Figure 14, where the highest observed system capacity is plotted against the used sample size. Furthermore, an indication is shown of the average time interval the sample relates to. A high negative slope indicates that the added data differs considerably from the current sample. An increase in sample size from 100 to 250 relates to a relatively large change in maximum capacity observed. Though, note that an overall sample size of 100 relates to an average ACRS sample size of 16,6 (the system contains 6 scanner stations at the starting section). Furthermore, a sample of 100 relates to approximately 2,1 minutes of input. It is assumed that the system is capable of handling input variation within such period.

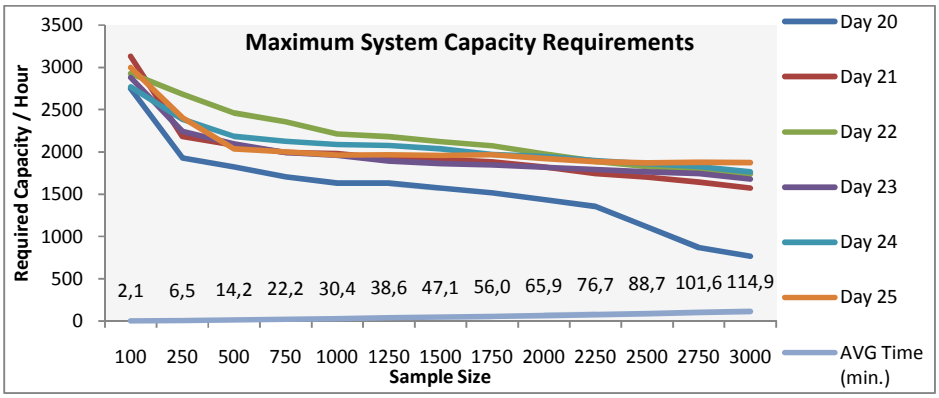



Figure 14: A sample response plot; maximum system capacity requirements based on different sample sizes

Based on Figure 14 and run sequence plots of various sample sizes, it was decided to select a sample size of 1250. In order to evaluate whether the sample is stationary the time series of interarrival times has been plotted. These run sequence plots can be found in appendix B.3.3. The arrival pattern of the sample interval is reasonably stable for the various days. An example of a run sequence plot can be found in Figure 15. In the figure the effect of a larger sample size is once more addressed. The left time series contains a sample size of 2250, while the right graph is based on a sample of 1250 bags. The sample of 1250 bags is also indicated in the left plot, within the red lines. Clearly, in the left graph the average interarrival time tends to drift upwards for the last section, simultaneously raising the standard deviation. More, related examples can be found in appendix B.3.2.

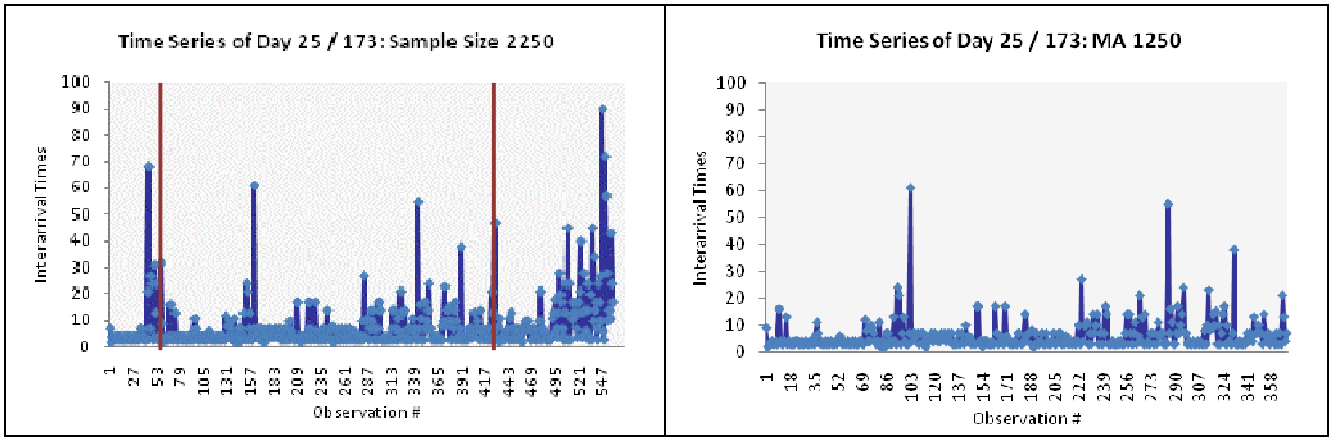



Figure 15: Time series comparison based on a sample size of 2250 and 1250

The peak load to select with a sample size of 1250 can be made more apparent by setting the moving average length to 1250. The resulting graph can be observed in section 2.6.1, in Figure 3 at the left hand side. The average capacity required remains null until the first 1250 bags have arrived. Note that the peak load matches the peak load in Figure 14 related to day 20 with a sample size of 1250.

### *Autocorrelation*

The autocorrelation of a time series can be evaluated with an autocorrelation plot (NIST/SEMATECH, 2010). The autocorrelation plot of the various days of ACRS 171 can be found in Figure 16. The values are relatively low, and appear to be having a more or less random pattern per sample, both indicating that

no significant autocorrelation is present. This is confirmed by the autocorrelation tables (appendix B.3.4, Table 22 and Table 23), which contain the exact values and indicate which measurements are significant based on the Pankrantz criterion. Autocorrelation plots and tables of the other automated code reader stations can be observed in Figure 37, Table 22 and Table 23.

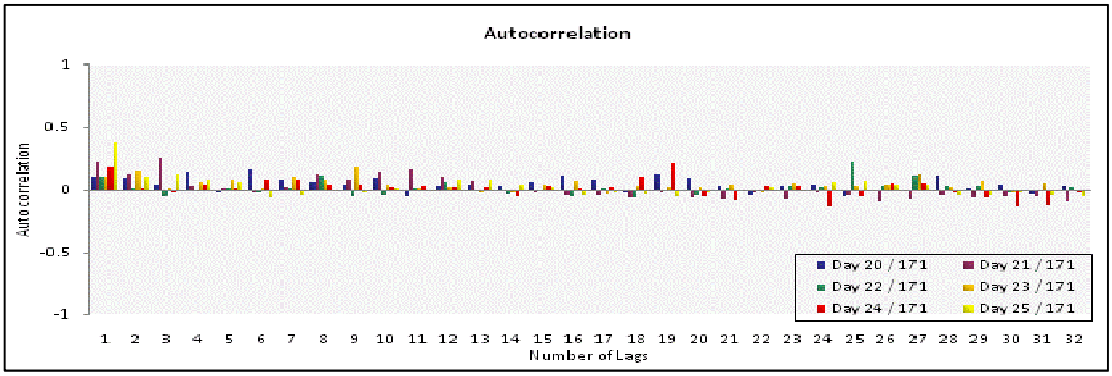



Figure 16: Autocorrelations of ACRS 171

### 5.2.3 Input Sample Selection

Now that peak loads have been identified for the total system, the samples can be split up into samples per input distribution / ACRS. The remainder of this chapter will focus upon evaluating the input distribution of ACRS 171. Evaluation of the five other stations is performed in a similar manner, with comparable results. For these results one will be redirected per section to the corresponding appendix.

For ACRS 171 summary statistics of interarrival times are presented in Table 10 (measured in seconds). The mean, variance, standard deviation, median (separating the higher from the lower half of the values), mode (the value that occurs most frequently), minimum, maximum, and amount of observations can be found per day. The median is lower than the mean, indicating that the underlying distribution is not symmetrical. Although the mode is similar for all samples, the mean values show more discrepancies. It becomes apparent that three of the samples are more closely related than the others; the means of day 21, 22 and 25 are more similar and the medians are equal. Summary statistics of the other scanning stations can be found in appendix B.3.5.

| *One Variable Summary* | Day 20 171 | Day 21 171 | Day 22 171 | Day 23 171 | Day 24 171 | Day 25 171 |
|---|---|---|---|---|---|---|
| Mean | 15,90 | 10,51 | 9,21 | 12,20 | 13,45 | 9,52 |
| Variance | 362,29 | 340,97 | 250,77 | 338,53 | 238,51 | 219,91 |
| Std. Dev. | 19,03 | 18,47 | 15,84 | 18,40 | 15,44 | 14,83 |
| Median | 10,00 | 4,00 | 4,00 | 6,00 | 7,00 | 4,00 |
| Mode | 3,00 | 3,00 | 3,00 | 3,00 | 3,00 | 3,00 |
| Minimum | 2,00 | 2,00 | 1,00 | 2,00 | 2,00 | 1,00 |
| Maximum | 127,00 | 136,00 | 150,00 | 171,00 | 125,00 | 121,00 |
| Count | 173 | 215 | 223 | 195 | 160 | 240 |

Table 10: Summary statistics of ACRS 171 interarrival times (sec), within the selected system peak of 1250 bags

More insights in the differences between populations can be acquired by using Box-Whisker plots. The Box-Whisker plot comparison of ACRS 171 is presented in Figure 17. From the Box-Whisker plots it can

be observed that the input data is right-skewed, meaning that the right tail is longer and the mass of the distribution is concentrated on the left (mean and median are placed far left). Furthermore, some differences between samples can be observed; day 20 and 24 contain on average higher interarrival times, and are more widely distributed (larger interquartile ranges). Similar Box-Whisker plots can be found in appendix B.3.5 for the others scanning stations.

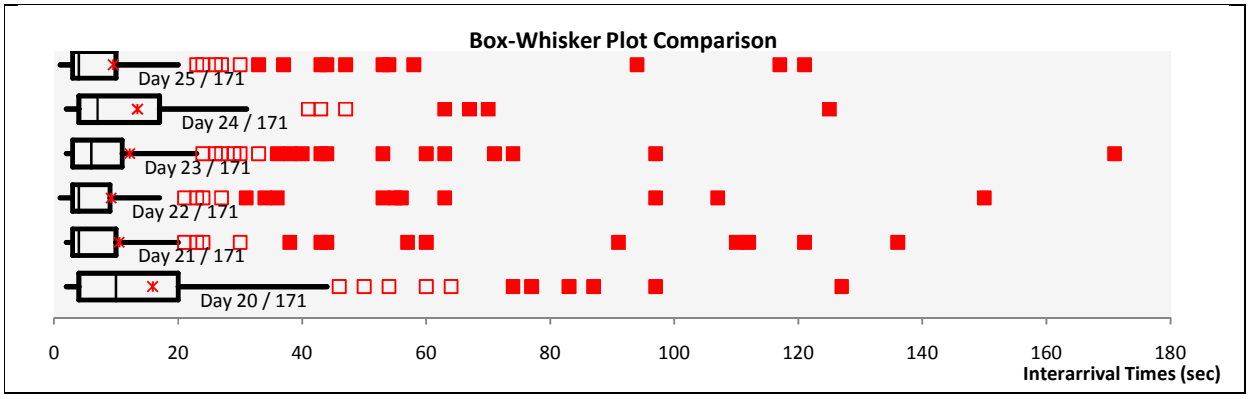



Figure 17: Box-Whisker plot comparison of ACRS 171

Although differences and similarities between samples of the various days have become clearer, this information is not sufficient to conclude that certain samples stem from the same distribution. In order to make such conclusion possible some statistical tests should be applied. As from the Box-Whisker plots it can be observed that the data is not normally distributed, nonparametric tests will be used.

### *Comparing Medians*

For comparing data samples that are not normally distributed the medians should be used rather than the means. Related to this the Kruskal Wallis test is the appropriate statistical technique. Its tests the null hypothesis that the medians within the samples are the same.

A downside is that when the hypothesis turns out to be insignificant, no information is obtained about which specific sample is particularly different. Consequently, the selection of multiple identical samples is an iterative process. If the test is not significant the most offset sample should be removed and the process should be repeated. Determining which sample is the most offset can be based upon the Box-Whisker plots. When more options are available, a combination of samples with the lowest mean and median is preferred, since high peak loads are pursued.

Test results of ACRS 171 can be observed in Table 12. Day 20, 24, and 23 have been successively removed of the total sample. A type I error of 5 % has been used, and consequently the null hypothesis will be accepted when the P-Value is larger than 0,05. The test appeared to be significant for the remaining combination of samples.

| | Kruskal Wallis test | | |
|---|---|---|---|
| Included Samples | Test Statistic | P-Value | Comments |
| Day 20 / 21 / 22 / 23 / 24 / 25 | 63,4451 | 2,354E-12 | Day 20 has the most offset median and mean |
| Day 21 / 22/ 23 / 24 / 25 | 36,0057 | 2,886E-7 | Day 24 has the most offset median and mean |
| Day 21 / 22 / 23 / 25 | 10,6715 | 0,0136 | Day 23 has the most offset median and mean |
| Day 21 / 22 / 25 | 1,4165 | 0,4925 | Accept null hypothesis |

Table 11: Comparing sample medians of ACRS 171

### *Comparing Standard Deviations*

Although various samples have been proven identical based on their median, this does not necessarily imply that they may be merged. Differences may exist based on their standard deviations. In order to verify this Levene's test should be applied. For ACRS 171 the result can be found in Table 12. Again, a type I error of 5 % has been used, which implies that the null hypothesis will be accepted when the P-Value is larger than 0,05. The hypothesis that samples of day 21, 22, and 25 are similar with respect to their standard deviation is accepted.

| Included Samples | Test Statistic | P-Value | Comments |
|---|---|---|---|
| Day 21 / 22 / 25 | 0,3077 | 0,7353 | Accept null hypothesis |

**Table 12: Comparing sample standard deviations of ACRS 171**

### *Assessing feasibility*

Previously it was discussed that, in case multiple combinations were possible, the one with the lowest mean and median was preferred. However, when most bags arrive for example the first day at ACRS 171, the second day at ACRS 172, and the third day at ACRS 173, retaining only the samples with the lowest interarrival times results in an overall system capacity requirement that in reality is never observed. Therefore, the feasibility of the combination of the selected samples per ACRS should be evaluated. The results can be observed in Table 13. The first column indicates the respective ACRS. In the second column usable sample combination are depicted, followed by the sample size and the average interarrival time. The most likely required capacity (ML) relates to the combination containing the most samples. The high capacity values are related to the combination of samples that result in the highest required capacity. When only one option is available the most likely capacity is equal to the high capacity combination. When comparing the total of the highest required capacities (2429 bags per hour) to the maximum observed system capacities in Figure 14, it can be concluded that it is still a feasible system capacity requirement.

| ACRS | Sample | Sample Size | Average | ML[Cap/h] | High[Cap/h] |
|---|---|---|---|---|---|
| 171 | 21/22/25 | 678 | 9,73 | 369,91 | 369,91 |
| 172 | 20/22/23/24/25 | 552 | 20,83 | 172,84 | 172,84 |
| 173 | 20/21/22/24 | 1180 | 7,89 | 456,39 | - |
| 173 | 23/25 | 717 | 6,40 | - | 562,26 |
| 271 | 22/23/24/25 | 950 | 9,28 | 387,81 | - |
| 271 | 21/22 | 536 | 8,18 | - | 440,04 |
| 272 | 20/24/25 | 756 | 9,41 | 382,77 | - |
| 272 | 21/23 | 580 | 7,41 | - | 485,54 |
| 273 | 22 | 140 | 14,36 | - | - |
| 273 | 24 | 238 | 9,03 | 398,67 | 398,67 |
| | | | Total | 2168,40 | 2429,28 |

**Table 13: Determining overall system capacity requirements when combining different ACRS samples**

## *5.2.4 Comparison of System and Simulation Input*

Because the total of the high capacity combinations appeared to result in a feasible system capacity requirement, these sample combinations have been used for the comparison with simulation distributions. Note that, in order to gain more insights or confidence, the most likely combinations may be compared additionally. This has not been done in this case study.

In order to facilitate deriving conclusions about the used simulation distribution, the best fitting distribution has been added to the comparison. This provides an upper limit of how close a distribution can get to the observed input data and consequently can be used as a reference. Related to ACRS 171 a lognormal distribution appeared to result in the best fit.

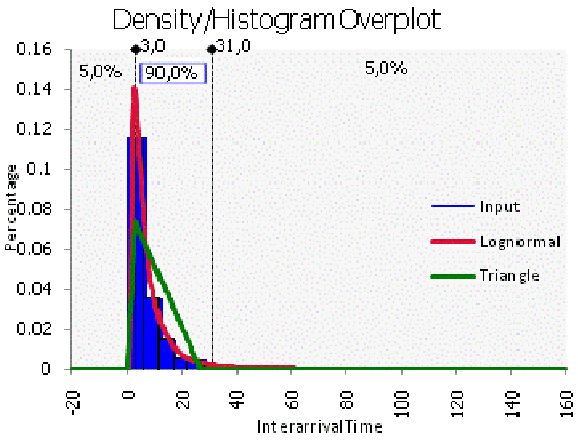
In the table in Figure 18 some summary statistics can be observed, as well as the results of the goodness-of-fit tests. In the data column the critical values of the different tests are depicted for α is 5 %. The triangle distribution scores infinity at the Anderson-Darling test because the test focuses primarily on differences in tails, which the triangle distribution lacks. Note that the Kolmogorov-Smirnov test and the Anderson-Darling test indicate that relatively large differences exist between the used and the best fitted distribution. The less suitable chi-squared test, on the other hand, even indicates the triangle distribution results in a slightly better fit.

The density/histogram overplot, frequency comparison and cumulative distribution function plot show a reasonable fit for the triangle distribution and a good fit for the lognormal distribution. The P-P plot amplifies differences between the middle parts of the distribution functions. Although the other graphical techniques showed a reasonable to good fit, this graph indicates somewhat more differences. When observing the Q-Q plot it becomes clearer why the goodness-of-fit tests' results are insignificant; too many discrepancies can be found in the tails of the distributions. However, it is expected that differences related to low interarrival times have a more significant impact on the system behavior. As such the distributions may still be suitable.

Without extensive experience with these graphs it remains difficult to conclude whether the observed differences are practically significant or not. Therefore, the effect of using the simulation input distribution or the best fitting distribution will be further evaluated within the sensitivity analysis.

Comparisons related to the other scanners can be found in appendix B.3.6.

| ACRS 171 | Data *(Cr. Value @0,05)* | Triangle Distr. | Lognormal Distr. |
|---|---|---|---|
| Minimum | 1 | 0 | 0,9227 |
| Maximum | 150 | 27 | +Infinity |
| Mean | 9,7415 | 10 | 8,4608 |
| Mode | 2 | 3 | 2,7148 |
| Median | 4 | 9 | 5,5924 |
| Std. Dev. | 16,3791 | 6,0415 | 9,5522 |
| *K-S* | *0,0519** | *0,3599* | *0,2086* |
| *A-D* | *2,4920** | *+Infinity* | *27,8989* |
| *Chi-Sq* | *36,4150** | *2182,6411* | *2190,6145* |



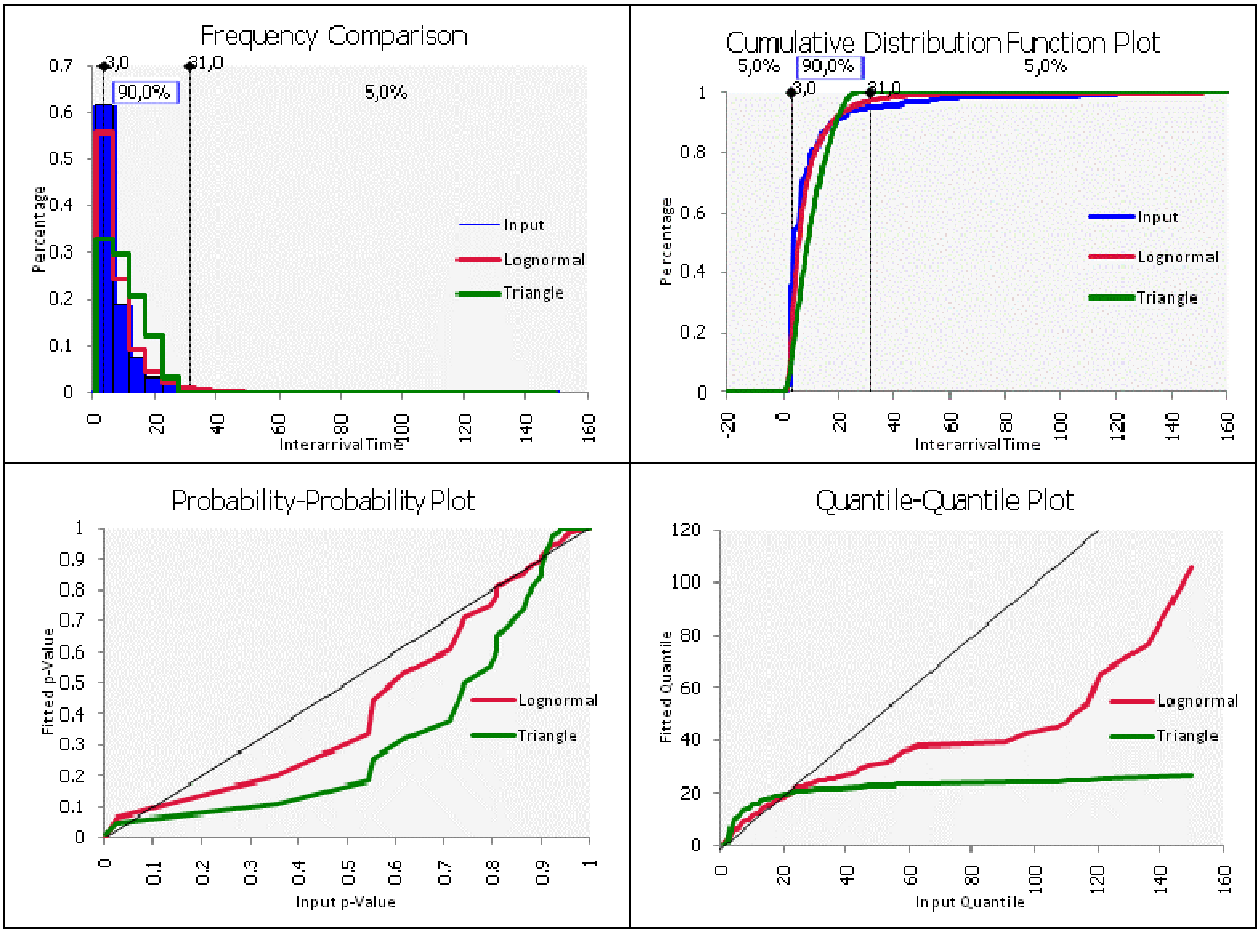



Figure 18: Comparisons for ACRS 171 – Sample: Day 21 / 22 / 25

## 5.3 Trace-Driven Output Validation

This section aims to validate the trace-driven output of the simulation model. The output evaluated is the cycle time of baggage items. Due to large differences within this response it should be divided into subgroups. These subgroups are formed by cycle time of: standard bags, which follow the standard routing (containing level 1 and level 2/3 screening, as well as manual coding), early bags, which require use of the EBS, and baggage items that leave the systems at the inspection zone. Furthermore, it appeared that a group of bags traverses both the right and the left subsystem; the assumption that back-up lines or not used is not valid. These cycle times also have been considered as a separate group, in order to keep the results of standard bags comparable. Within this section only the results of standard baggage items will be presented. Analyses of the other groups are depicted in appendix B.4.

Summary statistics of standard baggage items can be observed in Table 14. The number of cycle times counted differs between the system and simulation runs due to bags going out of tracking within the real system and bags going to the EBS or inspection based on a probability within the simulation model. Considering the real system, a standard bag requires on average 23 seconds more time to travel to a chute than within the simulation model. Though, as indicated by the 5% trimmed mean, a large part of

this difference is due to outliers. This is even more emphasized by the median, which is on average 11 seconds less for the real system than for the simulation model. These findings indicate that some differences exist for the lower half of observations and some differences exist for the higher half of observations. Finally, the standard deviation of the system cycle time is more than twice the size of the standard deviations of simulated cycle times. This is largely caused by the presence of outliers as well.

| Statistic | System | Run 1 | Run 2 | Run 3 |
|---|---|---|---|---|
| Count | 44839 | 60769 | 60823 | 60771 |
| Mean | 00:03:34 | 00:03:12 | 00:03:11 | 00:03:11 |
| 5% Trimmed mean | 00:03:21 | 00:03:10 | 00:03:10 | 00:03:10 |
| Median | 00:03:00 | 00:03:11 | 00:03:11 | 00:03:10 |
| Standard deviation | 00:02:14 | 00:01:03 | 00:01:02 | 00:01:03 |

Table 14: Summary statistics of the standard cycle times

The findings based on the summary statistics are supported by the Box-Whisker plots (Figure 19). A clear cause of differences related to the higher half of observations is the significant amount of high outliers found for the real system. As suspected these have also greatly influenced the standard deviation; no large deviations can be found between the quartiles and the spread of the whiskers.

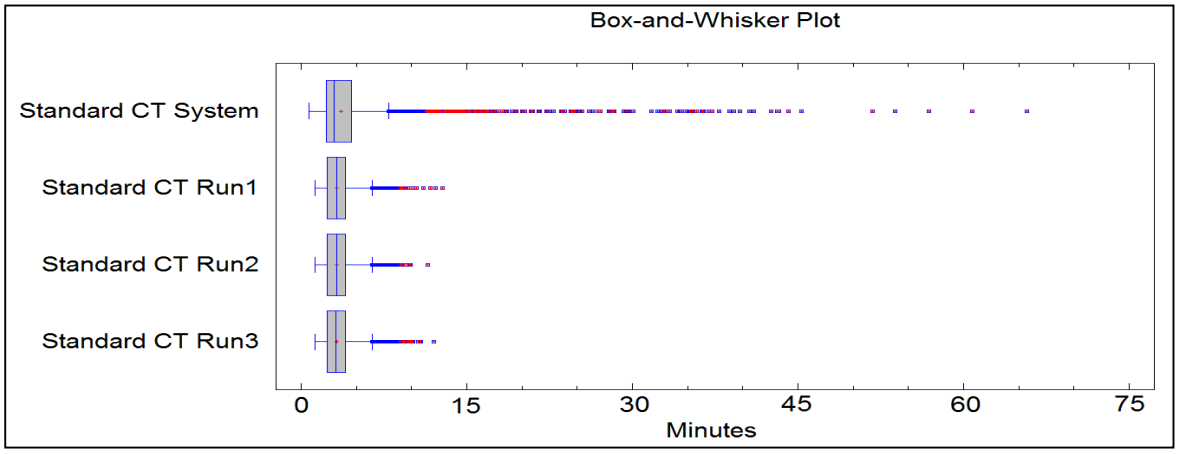



Figure 19: A box plot comparison of standard baggage cycle times

A histogram comparison of system data and run 1, concerning the cycle times of standard bags, can be observed in Figure 20. No additional histogram comparisons have been made with respect to run 2 and run 3, because the different runs are virtually identical. The histogram ranges from 0 to 1000 seconds, and does not contain a part of the outliers of the standard cycle time of the system.

Based on the figure it can be concluded that the probability density functions are reasonably similarly shaped. Within both histograms two peaks can be identified. However, concerning the real system data the first peak relates to a higher probability and a lower cycle time in comparison to the simulated data. Contrary, although again being higher, the second peak of the real system data relates to a higher cycle time. The dissimilarities of the first peak are expected to cause the differences in median. From the histogram comparison it can also be concluded that, although significantly affecting the mean, the outliers have no high impact on the general shape of probability density function; it concerns only a limited number of observations.

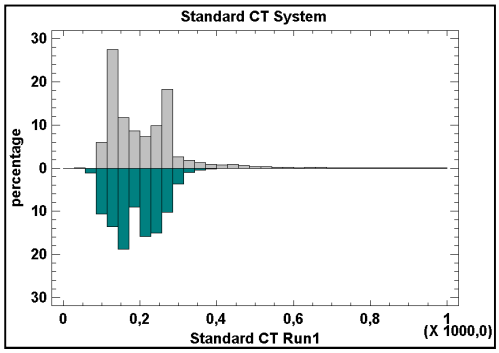



Figure 20: A histogram comparison of system data and run 1 for standard baggage cycle times (in seconds)

Behavior graphs have been created for both subsystems (left and right side of the BHS) separately. This is because the cycle times of bags traveling through a subsystems will virtually not be affected by an increase in utilization of the other subsystem. The behavior graph of the left subsystem can be found in Figure 21. The behavior graph of the right subsystem is presented in appendix B.4.4. In the figure high order polynomial regression lines are shown of average responses per half hour in order to present the general trend. These responses are the standard cycle times of the real system and the simulation model, as well as the mutual arrival rate (depicted per hour).

Some indication is present that the differences increase when the average cycle times start rising as a result of an increased arrival rate. However, in order to support this finding, data of even higher arrival rates appears to be required. Furthermore, a large difference can be observed related to the interval in which no bags arrive. The high cycle times of the average simulation run were unanticipated. However, further investigation indicated these were mainly due to a single run. Also it should be taken into account that average cycle times at the left side of the graph are based on very little bags, since no new baggage items arrive in those intervals.

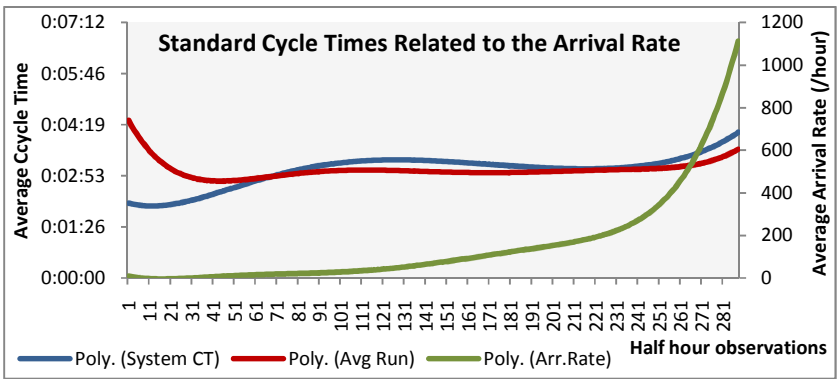



Figure 21: A behavior graph; standard cycle times related to the arrival rate, based on the left subsystem

Related to the behavior graph, in Table 15 the correlations of pairs of variables can be observed, based on average cycle times per half hour. Spearman's rank correlation was used due to the outliers present in system data. The table shows that significant positive correlations have been found between the system data and the various runs. As a reference, the correlations between the different runs are for run

1 – run 2, run 1 – run 3, and run 2 – run 3 respectively 0,5009, 0,4963, and 0,3790. From this it can be derived that the average standard cycle times of the system and simulation behave reasonably similar.

| System | Run 1 | Run 2 | Run 3 |
|---|---|---|---|
| Correlation | 0,2607 | 0,2024 | 0,4197 |
| P-Value | 0,0000 | 0,0006 | 0,0000 |

Table 15: Spearman rank correlations between the system and various runs

## 5.4 Conducting a Sensitivity Analysis

Within this section a sensitivity analysis will be performed in order to gain more insights about main causes of differences between the simulation model and the real system. A resolution V design is used, such that second order interactions are taken into account. Furthermore, group screening as well as common random numbers has been applied. The warm-up period, run length, and replications as described in appendix B.1.2 have been used. This section will mainly focus upon the results of the sensitivity analysis. Detailed information about setting up the experiment can be obtained from appendix B.5.2.

### 5.4.1 Sensitivity Analysis Results

The selected 11 factors can be observed in Table 16, which contains the factor names as applied in AutoStat and a small description.

| Factor Nr. | Factor Name | Description |
|---|---|---|
| 1 | Distribution | The distribution to use as system input |
| 2 | L1_fr | The failure rate, or reject rate of cases at screening level 1 and 2 |
| 3 | L3_fr | The failure rate, or reject rate of cases at screening level 3 and 4 |
| 4 | MC_r | The rate of bags that requires manual coding |
| 5 | EBS_r | The rate of bags that requires the EBS |
| 6 | EBS_Control | A combination of factors that together determine how the EBS is controlled |
| 7 | EBS_to_MC | Recurrence to manual coding after usage of the EBS |
| 8 | Lateral Assignment | Assignment of the location where a bag leaves the system |
| 9 | ST | Service times for manual encoding and manual inspection |
| 10 | Velocity | The velocity of system components |
| 11 | WindowLength | The window length of system components |

Table 16: Factors used within the DOE

The effects of varying the presented factors are investigated for various responses. Besides the cycle times of different groups of bags, as discussed in section 5.3, throughput rates, describing the behavior of the system, and the work in process have been taken into account (as discussed in appendix B.1.2).

In Figure 22 the effects of factors on throughput can be observed, for the various measured locations. All main effects and the two most influential interaction effects are depicted separately. All other interaction effects are grouped together in the data series “others”. Effects are presented in terms of percentages, in order to make their impacts comparable.

It appears that the selection of a distribution has a relatively high impact on throughput rates. Especially the main stream of baggage items (relating to Induct 1_1 / 1_2 / 1_3 / 2_1 / 2_2 / 2_3, and the sorters) is affected by a change. Sorter throughput is furthermore influenced by lateral assignment. The third sorter

inducts are related to the flow being approved by the level 3 screeners. These are positively influenced by a higher reject rate of level 1 screening machines, and negatively by a higher reject rate of level 3 screening machines. Consequently, a higher reject rate of level 3 screening machines has a positive impact on the throughput at inspection stations. The fourth sorter inducts, corresponding to EBS outputs, are mainly effected by the EBS rate and the EBS control. Manual coding throughput is almost solely affected by the manual coding rate. Surprisingly, recurrence to manual coding of bags that have used the EBS has virtually no impact. Furthermore, it is interesting to see that velocity, window length, and service time have virtually no effect on any throughput rate.

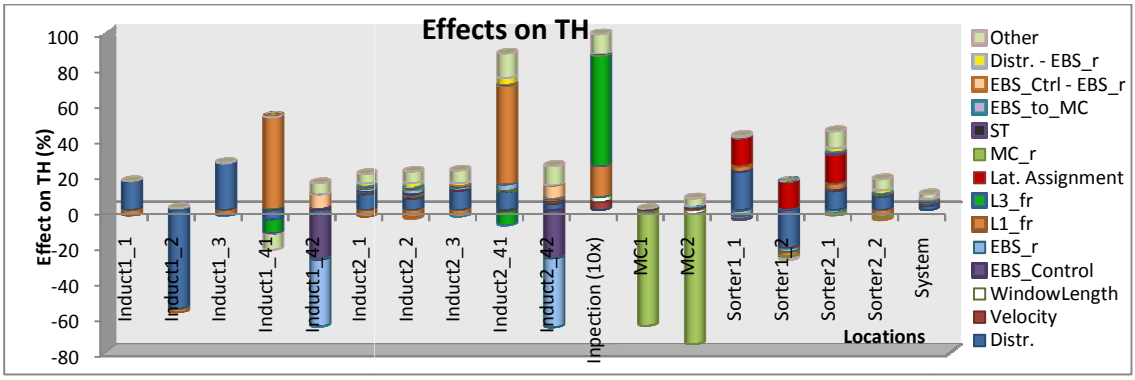



Figure 22: Effects of factors on the throughput at different locations

In Figure 23 the factor effects on the cycle time of standard baggage can be seen, as well as the related 95 % confidence intervals. The figure contains two vertical axes. The left axis indicates the absolute factor effect, while the right vertical axis presents the effects in terms of percentages. The figures only contain the 15 most significant effects. In this setting, significant implies practically significant; the size of the effects is taken into account. Statistical significance is indicated with the use of color. Dark blue implies an effect is statistically significant, while light blue indicates a lack of statistical significance (the confidence interval contains the null value).

From Figure 23 it can be concluded that the impact of alternative conveyor velocities on standard baggage cycle times can be quite large. Other factors that should be taken into account are the input distribution and lateral assignment, and to a lesser extend window length and the EBS rate. Note that also the interactions between these factors have a considerable impact.

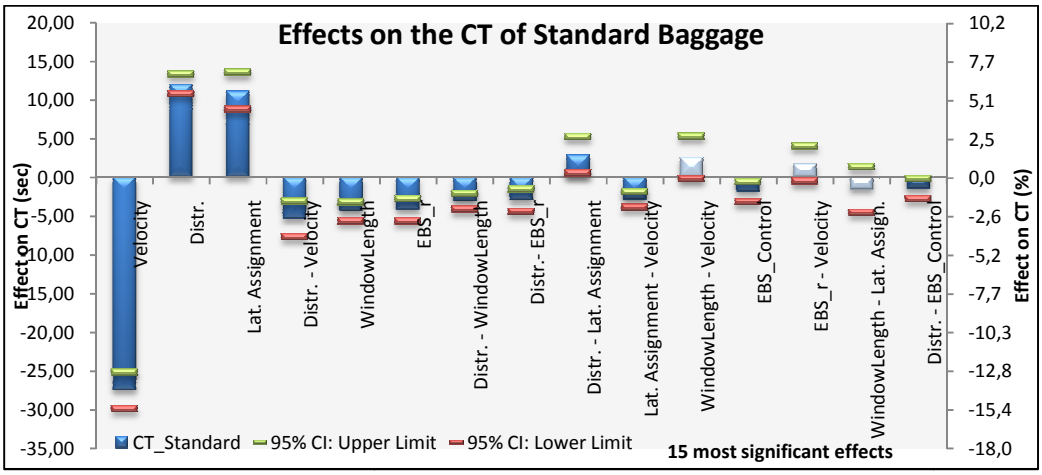



Figure 23: Factor effects on the cycle time of standard baggage

Exact information about factor effects on different responses can be found in appendix B.5.5. Note that, in concordance with the illustrative purpose of this case study, factor groups have not been evaluated in more detail in subsequent stages of the group screening process. Such additional actions might have provided clarity related to the unanticipated effect of window length on cycle times.

One general remark should be made with respect to interpretation of the results. Distinction should be made between the factors of which the alternative levels are based on system data, and the factors of which the alternative levels are based on other sources, such as expert opinions. It is known that the first group is actually present, while the second group of factors only may be present. Thus, the first group can be directly used to explain observed differences, while the second group merely presents clues of what may cause differences observed that cannot be explained otherwise. When this is the case, or these factors appear to be of very high influence on system performance, it may be advisable to pursue retrieval of the real system parameter values. Related to this sensitivity analysis lateral assignment, service times, window lengths, and velocity were not based on real system data.

## 5.5 Discussion

Generally, system output has shown to be quite close to simulation output. However, it appears that differences increase when utilization increases. Comparison of the original simulation results of standard bags related to peak loads (average of 0:03:15) to real system performance based on cycle times under peak loads (average of 0:03:42), an average difference of 27 seconds can be observed (13,8 %). Using the appropriate distribution this difference could be narrowed down to 14 seconds (7,2 %). As an important possible cause of this difference the relatively large number of high outliers contained in the system data can be designated. However, when removing extreme outliers from the system data a difference of 11 seconds remains (5,4 %). Based on the sensitivity analysis it can be said that further investigation of lateral assignments is recommended, in order to find a possible explanation for this difference. Although this might reduce the unexplained dissimilarity even further, a portion cannot be explained by parameter differences, but is due to the level of detail contained in the model and the assumptions made.

As a result of the limited usability of the objective methods for input validation, conclusions should be mainly derived from subjective techniques. Consequently, a certain degree of experience is required in order to be able to conclude whether observed differences are practically significant or not. The sensitivity analysis may aid in deriving conclusions by using the best fitting distributions as an alternative factor level.

Related to the trace-driven output validation, the cycle times of standard bags appeared to consist of two conjoint distributions with contrastive differences. Interpretation of these observed differences in the histogram comparison would be facilitated if it was known in what context the cycle times of both peak probabilities occur. Due to the limited effect of screening reject rates and manual encoding rates these can be excluded as possible causes. Another possibility might be that they are caused by differences in utilization.

Although the general output of standard bags appeared to be reasonably similar, significant differences were found related to parameter settings. Though these differences only had an effect on throughput rates and did not significantly affect the cycle times within this setting, it is expected that they will affect model performance if it is subjected to even higher utilization rates.

A downside of the sensitivity analysis is that it only gives insight in effects on average responses, while, as indicated for output validation, more detailed information may be preferred. In this case alternative levels of factors indicated as important by the sensitivity analysis can be used to generate new simulation results for output validation (with exception of the input distribution).

As a more general finding, based on this case study, it may be said that a model containing relatively little detail may be a viable option for simulating a baggage handling system that is relatively simple and does not experience very high utilization rates. For this type of models selecting appropriate input distributions will likely remain a challenge, since they may be of prime essence in order to keep differences with the real system small. However, in order to increase generality of, and confidence in these findings, more research is required for confirmation. Furthermore, with additional research, it may be possible to relate certain findings to specific assumptions within the assumptions document.

Though, note that this reasoning makes the assumption that the goal of simulation is to develop a model which's performance is as close as possible to the real system, while this is not necessarily true. The main goal of simulation studies at Vanderlande Industries is the reduction of risks. For this purpose a very high accuracy might not always be required; validity relates to a sufficient level of accuracy for the intended purpose of the model and consequently the model may still be valid. Since no appropriate goal was specified for the model of Cairo Terminal 3 (it has been derived of an animation model), no final verdict can be given with respect to its validity.

# 6. Conclusions & Recommendations

In this final chapter the conclusions and recommendations resulting from this Master Thesis are presented. Section 6.1 contains the conclusions, while the recommendations related to future research are discussed in section 6.2.

## 6.1 Conclusions

Within this Master Thesis the most appropriate methods for the logistic simulation models of Vanderlande Industries have been selected to form an operational validation approach. At a high level, this approach consists of maintaining an assumptions document, validating simulation input, validating trace-driven output, and conducting a sensitivity analysis.

An initial selection of appropriate methods has been made by assessing current literature based on evaluation criteria. For input validation and trace-driven output validation, these criteria are: generality, power, objectivity, data, and effort. Additionally, for trace-driven output validation the subject of comparison is taken into account. Other criteria have been used for evaluating methods for conducting a sensitivity analysis, namely efficiency, effectiveness, robustness and ease of use. Furthermore, this selection has been partly based on assumptions about simulation models, which literature claims to be generally true.

Within the case study the applicability and the legitimacy of the conceptual model have been assessed. Because a single case was used, for which a baggage handling system had been selected, the practical findings are limited to its related environment.

The logistic systems simulated at Vanderlande Industries are inherently stochastic. Because random inputs will produce random outputs it is difficult to relate observed differences to specific model characteristics. To isolate the differences in simulation models from a model's own randomness, the system and simulation model, or various model variants, should be evaluated under similar stochastic circumstances. Furthermore, as a result of dynamic, nonstationary input and correlated, nonstationary output, formal statistical techniques have turned out to be difficult to apply, because of violations of their assumptions.

Since a simulation model is a simplification, and consequently merely an approximation of the real system, some differences between the real system and the model are to be expected. As a result, tests that evaluate whether system and model input or output are similar are expected to be false. Therefore, it is more useful to ask whether or not the differences between the model and system are significant enough to affect any conclusions derived from the model. Because validity relates to a sufficient level of accuracy for the intended purpose of a model, no definite criteria can be used in general to determine whether a model is valid or not; the effect of observed differences should be assessed in relation to the objective of the model.

Making the model's assumptions explicit in an assumptions document is important because the model's assumptions and input values determine whether the model is valid, and will remain valid when the real system and its environment will change. The document should be used to make the level of detail contained in the simulation model explicit, as well as its intended purpose. For Vanderlande Industries this assumptions document can have additional value; if the assumptions are conform customer specifications for instance, differences between the real system and the simulation model that can be attributed to these assumptions can be acceptable.

Statistical techniques that evaluate whether system data can be considered a random sample of a predefined distribution generally tend to be too sensitive for simulation purposes. However, the statistics can be used to compare applied distributions to best fitting distributions. For the graphical techniques proposed for input validation experience is required to conclude whether the input is valid or not. In absence of the required level of experience the effect can be further evaluated within the sensitivity analysis.

In order to isolate the differences between a model and the real system from a model's own randomness, output validation should be trace-driven; i.e. model output should be based on real system input. An increase in utilization generally leads to an increase in the difference between simulation output and system output. As such, a difficulty with output validation remains that differences are preferably evaluated per range of utilization. However, for this purpose a high amount of data is required of a relatively stable period per range of utilization, which is difficult to acquire for baggage handling systems. In order to acquire detailed insights in the behavior of differences between a real system and a simulation model, a high utilization range is required. Typically airports experience the most extreme capacity requirements only a few days per year, of which the dates are generally known beforehand. Data of these days would be especially appropriate for operational validation. . Both turned out to be a difficult within the case study. Only a small amount of data was related to higher utilization rates, and in general no utilization rates above 60 % were observed.

A sensitivity analysis can generate insights into what causes the differences observed at the output validation. Furthermore, it can result in an upper boundary in what output may be reached with a simulation model when all the parameters are configured optimally, based on the level of detail it contains. Alternative levels of factors indicated as important by the sensitivity analysis can be used to generate new simulation results for output validation, if a more detailed evaluation of their effects is preferred. Differences between model variants can be isolated from a model's own randomness by using a variance reduction technique called common random numbers.

## *5.2 Recommendations*

The research within this Master Thesis has been bounded to validation methods that could increase the degree of confidence in the simulation models. This in absence of a feedback loop from implemented systems to simulation models. Additional to this research it can be recommended to evaluate possibilities to facilitate the feedback loop itself. A large reduction of time and effort can be gained by simplifying data acquisition and processing. It would be beneficial if system responses and parameter

values could be observed more directly. An important option would be the use of BPI (Business Process Intelligence) to analyze data logging quickly. BPI is Vanderlande's standard software solution for gathering, storing and analyzing data.

The information obtained from the approach can be used to improve future system simulations in a general manner. It is possible to use operational validation to determine which level of detail should be used in simulation models. This can be made clear by discussing the role of accuracy within simulation in more detail. Balci (1998) argues that, within the domain of a model's applicability, it should behave with satisfactory accuracy, consistent with the study objectives. Emphasis is placed upon satisfactory because an increase in accuracy generally relates to an increase in the amount of time and resources required for the simulation. Consequently, dependent of the simulation requirements, a tradeoff can be made between the level of detail used in a model and the cost of performing a simulation, i.e. the complexity of the coding and the additional value of it for the simulation have to be balanced against each other. This tradeoff is shown in Figure 24 as a function of model credibility, development cost and utility. Model credibility can be defined as "the level of confidence in a simulation's results" (Fosset et al., 1991, p. 712). Model accuracy is an important factor of model credibility (Fosset et al., 1991). The validation approach can be used to assess this dilemma; the accuracy of models containing different levels of detail can be determined by comparing the input and output of simulation models with the corresponding real system.

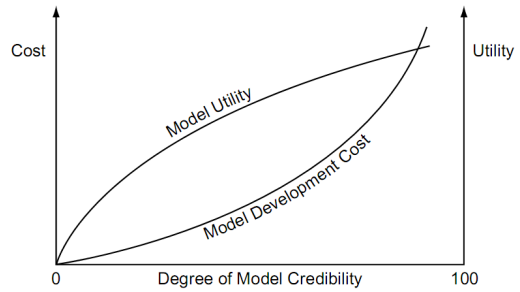



**Figure 24: Degree of model credibility (Balci, 1998)**

Distinguishing different types of simulation models with respect to their objectives, different tolerance limits can be identified that determine whether or not a model is valid. This will aid determining which level of detail results in a sufficient level of accuracy. Subsequently, it can be investigated whether it is possible to adapt statistical techniques that turned out to be too sensitive in this research, such that they allow for differences within the defined limits.

Within this research the focus has been on validation of the total simulation model. However, it is also possible to focus more on the level of accuracy resulting from the level of detail contained in a specific model component, as opposed to the overall system. This would aid in selecting a specific component version when building a model, depending on the purpose of the model.

Airports typically experience the most extreme capacity requirements only a few days per year, of which the dates are generally known beforehand. Because a high utilization range is preferred for output

validation, data of these days would be especially appropriate for operational validation. When planning to compare simulation and system output, these possibilities are recommended.

Finally, within the sections related to practical findings, some indications have been given about suitable software tools that can be used to apply the various methods proposed in this research. Acquisition of an appropriate tool is recommended for conducting the methods discussed. However, it is also amongst the possibilities to program the various statistical methods, in Excel for instance.

## *Glossary of Terms*

| | |
|---|---|
| BHS | Baggage Handling System |
| BPI | Business Process Intelligence |
| cdf | cumulative density function |
| CRN | Common Random Numbers |
| DOE | Design Of Experiments |
| IID | Independent and Identically Distributed |
| pdf | probability density function |
| VI | Vanderlande Industries |

## *References*


Andres, T.H. & Hajas, W.C. (1993). Using iterated fractional factorial design to screen parameters in sensitivity analysis of a probabilistic risk assessment model. In Küster, H., Stein, E. & Werner, W. (Eds), *Proceedings of the Joint International Conference on Mathematical Models and Supercomputing in Nuclear Applications,* 2, pp. 328 – 337

Applied Materials, Inc. (2008). *AutoMod User's Guide*. Version 12.1.3, Santa Carla

Balci, O. (1998). Verification, Validation, and Testing. In Banks, J. (Eds), *Handbook of Simulation*, Wiley: New York, pp. 335 – 392

Banks, J. (2004). *Getting Started with AutoMod.* Second Edition, Brooks Automation: Chelmsford

Bartels, R. (1982). The Rank Version of von Neumann's Ratio Test for Randomness. *Journal of the American Statistical Association,* 77(377), pp. 40 – 46

Bettonvil, B. & Kleijnen, J.P.C. (1996). Searching for important factors in simulation models with many factors: Sequential bifurcation. *European Journal of Operational Research,* 96, pp. 180 – 194

Box, G.E.P., Hunter, J.S. & Hunter, W.G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery,* Second Edition, Wiley: New Jersey

Cooper, D.R. & Schindler, P.S. (2003). *Business Research Methods,* 8th Edition McGraw-Hill: New York

Dean, A.M. & Lewis, S.M. (2002). Comparison of group screening strategies for factorial experiments. *Computational Statistics & Data Analysis,* 30, pp. 287 – 297

Engel, J. (1997). The Multiresolution histogram. *Metrika,* 46,pp. 41 – 57

Farrington, P.A. & Swain, J.J. (1993). Design of Simulation Experiments with Manufacturing Applications. In Evans, G.W., Mollaghasemi M., Russel, E.C. & Biles W.E. (Eds), *Proceedings of the 1993 Simulation Conference,* pp. 69 – 75

Fosset, C.A., Harrison, D., Weintrob, H. & Gass, S.I. (1991). An Assessment Procedure for Simulation Models: A Case Study. *Operations Research*, 39(5), p. 710 – 723

Garson, G.D. (2010). Syllabus for PA766: Advanced Quantitative Research in Public Administration, North Caronlina State University, Retrieved March 10, 2010, from http://faculty.chass.ncsu.edu/garson/PA765/time.htm#stationarity

Gass, S.I. (1983). Decision-Aiding Models: Validation, Assessment, and Related Issues for Policy Analysis. *Operations Research*, 31(4), pp. 603 – 631

Gibbons, J.D. & Chakraborti, S. (2003). *Nonparametric Statistical Inference,* Fourth Edition, Marcel Dekker: New York

Glasserman, P. & Yao, D.D. (1992). Some guidelines and guarantees for common random numbers, *Management Science,* 38(6), pp. 884 – 908

Goossenaerts, J.B.M. & Pels, H.K. (2006). *Methodology & Instruction Materials for the course: Simulation of Operational Processes.* Eindhoven University of Technology: Eindhoven

Green, S.B. & Salkind, N.J. (2004). *Using SPSS for Windows and Macintosh: Analyzing and Understanding Data,* Fourth Edition, Pearson: New Jersey

Greene, W.H. (1993). *Econometric Analysis.* Macmillian Publishing: New York

Groenendaal, W.J.H., van & Kleijnen, J.P.C. (1997). On the assessment of economic risk: factorial design versus Monte Carlo methods. *Journal of Reliability Engineering and System Safety,* 57, pp. 91 – 102

Holcomb, D.R., Montgomery, D.C. & Carlyle, W.M. (2003). Analysis of supersaturated designs. *Journal of quality technology,* 35(1), pp. 13 – 27

Hopp, W.J. & Spearman, M.L. (2008). *Factory Physics*. Third Edition. McGraw-Hill: New York

Horn, R.L. van (1971). Validation of simulation results. *Management Science,* 17(5), pp. 247 – 258

Hsu, D.A. & Hunter, J.S. (1977). Analysis of Simulation-Generated Responses Using Autoregressive Models. *Management Science,* 24, pp. 181 – 190

Ivanova, T., Malone, L.C. & Mollaghasemi, M. (1999). Comparison of a two-stage group-screening design to a standard $2^{k-p}$ design for a whole-line semiconductor manufacturing simulation model. In *Farrington, P.A., Nembhard, H.B., Sturrock, D.T. & Evands, G.W. (Eds), Proceedings of the 1999 Winter Simulation Conference*

Johnston, W.J., Leach, M.P. & Liu, A.H. (1999). Theory-Testing Using Case Studies in Business-to-Business Research. *Industrial Marketing Management*, 28, pp. 201 – 213

Kleijnen, J.P.C. (2009). Sensitivity analysis of simulation models. In Cochran, J.J., Cox, L.A., Keskinocak, P., Kharoufeh, J.P. & Smith, J.C. (Eds), *Wiley Encyclopedia of Operations Research and Management Science,* Wiley: New York

Kleijnen, J.P.C. (2008). *Design and Analysis of Simulation Experiments*, Springer: New York

Kleijnen, J.P.C. (2005). An overview of the design and analysis of simulation experiments for sensitivity analysis. *European Journal of Operational Research*, 164, pp. 287 – 300

Kleijnen, J.P.C. (1999). Validation of models: Statistical techniques and data availability. In Farrington, P.A., Nembhard, H.B., Sturrock, D.T. & Evans, G.W. (Eds), *Proceedings of the 1999 Winter Simulation Conference,* pp. 647 - 654

Kleijnen, J.P.C. (1997). Experimental design for sensitivity analysis, optimization, and validation of simulation models. In Banks, J. (Eds), *Handbook of Simulation: Principles, Methodology, Advances, Application, and Practise,* Wiley: New York, pp. 173 - 224

Kleijnen, J.P.C. (1995). Verification and validation of simulation models. *European Journal of Operational Research*, 82, pp. 145 – 162

Kleijnen, J.P.C. (1992). Sensitivity analysis of simulation experiments: regression analysis and statistical design. *Mathematics and Computers in Simulation,* 32, pp. 297 – 315

Kleijnen, J.P.C., Bettonvil, B. & Persson, F. (2006). Screening for the Important Factors in Large Discrete-Event Simulation Models: Sequential Bifurcation and Its Application. In Dean, A. & Lewis, S. (Eds), *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics,* Springer: New York, pp. 287 – 307

Kleijnen, J.P.C., Sanchez, S.M., Lucas, T.W. & Cioppa, T.M. (2004a). A user's guide to the brave new world of designing simulation experiments. *INFORMS Journal on Computing*

Kleijnen, J.P.C., Bettonvil, B. & Person, F. (2004b). Finding the important factors in large discrete-event simulation: sequential bifurcation and its applications. In Dean, A.M., Lewis, S.M. (Eds), Screening, Springer: New York

Kleijnen, J.P.C., Cheng, R.C.H. & Bettonvil, B. (2001). Validation of Trace-Driven Simulation Models: Bootstrap Tests. *Management Science,* 47(11), pp. 1533 – 1538

Kleijnen, J.P.C., Cheng, R.C.H. & Bettonvil, B. (2000). Validation of Trace-Driven Simulation Models: More On Bootstrap Tests. In Joines, J.A., Barton, R.R., Kang, K. & Fishwick, P.A. (Eds), *Proceedings of the 2000 Winter Simulation Conference,* pp. 882 – 892

Kleijnen, J.P.C. & Sargent, R.G. (2000). A methodology for fitting and validating metamodels in simulation. *European Journal of Operational Research,* 120, pp. 14 – 29

Kleijnen, J.P.C. & Standridge, C.R. (1988). Experimental design and regression analysis in simulation: An FMS case study. *European Journal of Operational Research,* 33, pp. 257 – 261

Law, A.M. (2008). How to build valid and credible simulation models. In Mason, S.J., Hill, R.R., Mönch, L., Rose, O., Jefferson, T. & Fowler, J.W. (Eds), *Proceedings of the 2008 Winter Simulation Conference,* pp. 39 – 47

Law, A.M. (2007). Statistical Analysis of Simulation Output Data: The Practical State of the Art. In Henderson, S.G., Biller, B., Hsieh, M.-H., Shortle, J., Tew, J.D. & Barton, R.R. (Eds), *Proceedings of the 2007 Winter Simulation Conference*, pp. 77 – 83

Law, A.M. & Kelton, W.D. (2000). *Simulation Modeling and Analysis*. Third Edition, McGraw-Hill: New York

Mauro, C.A. (1984). On the Performance of Two-Stage Group Screening Experiments. *Technometrics,* 26(3), pp. 255 – 264

Montgomery, D.C. (1991). *Design and Analysis of Experiments.* Third Edition, Wiley: New York

Montgomery, D.C. & Runger, G.C. (2002). *Applied Statistics and Probability for Engineers.* Third Edition, Wiley: New York

NIST/SEMATECH (2010). *e-Handbook of Statistical Methods,* Retrieved March 5, 1010, from http://www.itl.nist.gov/div898/handbook/

Nyhuis, P., Cieminski, G., von, Fischer, A., & Feldmann, K. (2005). Applying Simulation and Analytical Models for Logistic Performance Prediction. *CIRP Annals – Manufacturing Technology*, 54(1), pp. 417 – 422

Palisade Corporation (2009). Guide to Using @Risk: Risk Analysis and Simulation Add-In for Microsoft Excel. *@Risk Manual,* pp. 705

Philips, P.C.B. & Perron, P. (1988). Testing for a Unit Root in Time Series Regression. *Biometrika,* 75, pp. 335 – 346

Saltelli, A., Andres, T.H., & Homma, T. (1995). Sensitivity analysis of model output: Performance of the iterated fractional factorial design method. *Computational Statistics & Data Analysis,* 20, pp. 387 – 407

Sargent, R.G. (2008). Verification and validation of simulation models. In Mason, S.J., Hill, R.R., Mönch, L., Rose, O., Jefferson, T. & Fowler, J.W. (Eds). *Proceedings of the 2008 Winter Simulation Conference,* pp. 157 – 169

Sargent, R.G. (1996). Some Subjective Validation Methods Using Graphical Displays of Data. *Proceedings of the 1996 Winter Simulation Conference,* pp. 345 – 351

Schipper, A. (n.d.). *Validating and Optimizing System Design with Simulation and Emulation*. Retrieved September 14, 2009, from Vanderlande Industries restricted archives

Shen, H. & Wan, H. (2009). Controlled sequential factorial design for simulation factor screening. *European Journal of Operational Research,* 198, pp. 511 – 519

Statpoint Technologies (2009). *STATGRAPHISCS Centurion XVI User Manual*

Trocine, L. & Malone, L.C. (2001). An overview of newer, advanced screening methods for the initial phase in an experimental design. In Peters, B.A., Smith, J.S., Medeiros, D.J. & Rohrer, M.W. (Eds), *Proceedings of the 2001 Winter Simulation Conference*

Vanderlande Industries (2010). *Product Data Book: Baggage Handling (Part I)*, Retrieved February 2, 2010, from Vanderlande Industries restricted archives

Vanderlande Industries (2009a). *Automated logistics solutions for warehouses and distribution centres*. Retrieved September 15, 2009, from http://www.vanderlande.com/web/Distribution.htm

Vanderlande Industries (2009b). Baggage Handling System Terminal 3: Cairo International Airport. *Functional Specification,* Retrieved February 10, 2010 from Vanderlande Industries restricted archives

Vanderlande Industries (2009c). Baggage Handling Systems*. Baggage Handling Brochure*, Retrieved September 15, 2009, from http://www.vanderlande.com/web/Baggage-Handling.htm

Vanderlande Industries (2009d). Equipment for Automated Baggage Handling*. Equipment Brochure*, Retrieved February 2, 2009, from Vanderlande Industries restricted archives

Vanderlande Industries (2009e). High Tech in the Desert*. Intranet News*, Retrieved November 20, 2009, from Vanderlande Industries restricted archives

Vanderlande Industries (2009f). Parcel and Postal. *Parcel & Postal Brochure*, Retrieved September 15, 2009, from http://www.vanderlande.com/web/Parcel-Postal.htm

Vanderlande Industries (2009g). Services. *Service Brochure*, Retrieved September 15, 2009, from http://www.vanderlande.com/web/Services.htm

Vanderlande Industries (2009h). The Company: Automated Handling Systems. *Company Brochure*, Retrieved September 15, 2009, from http://www.vanderlande.com/web/About-us.htm

Verschuren, P. & Doorewaard, H. (1995). *Het ontwerpen van een onderzoek,* First edition, Lemma: Utrecht

Vonk Noordegraaf, A. (2002). Simulation modeling to support national policy making in the control of bovine herpes virus 1. *Doctorial dissertation,* Wageningen University, Wageningen

Yin, R.K. (1994). *Case Study Research: Design and Methods*. Second edition, Thousand Oaks: Sage Publications

Steppan, D.D., Werner, J. & Yeater, R.P. (1998). *Essential Regression and Experimental Design for Chemists and Engineers.* Retrieved October 8, 2009, from http://www.geocities.com/SiliconValley/Network/1900/

Wan, H. & Ankenman, B.E. (2006). Two-stage controlled fractional factorial screening for simulation experiments. *Journal of Quality Technology,* 39(2), pp. 126 – 139

White, H. (1980). A Heteroskedsticity-Consistant Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica, 48(4), pp. 817 – 838

# Appendix A. Additional Approach Development Information

In this appendix methods presented within the approach development chapter will be discussed in more detail.

## A.1 Additional Information of Input Validation

In this section more detailed information is presented about various methods for input validation.

### A.1.1 Data Evaluation: Data Stationarity & Sample Independence

Linear regression is a statistical technique that tries to fit a linear model to the data, generally by minimizing the sum of squared residuals (Montgomery & Runger, 2002). Linear regression can be used to evaluate changes in mean; the hypothesis can be tested that a horizontal linear line fits the data. However, the method may not be appropriate, due to its limiting assumptions. The assumptions of data independence and homoscedasticity (the variance does not change over time) were already requirement, but in addition it is assumed that the data and residuals are normally distributed (Green & Salkind, 2004). This may be problematic, specifically because arrival processes are typically characterized as exponentially distributed.

Based on linear regression models several tests are available for testing homoscedasticity (Breusch-Pagan test (Greene, 1993) and the absence of autocorrelation (Durbin-Watson coefficient, Dickey-Fuller test, Augmented Dickey-Fuller test, Unit Root Test (Garson, 2010)). However, these tests also rely on the normality of the residuals of a fitted regression model. Related to homoscedasticity, White's test (White, 1980) and the modified Breusch-Pagan test are alternatives that are less sensitive to the assumption of normality (Greene, 1993). Concerning autocorrelation, the Philips-Perron test (Philips & Perron, 1988) is a non-parametric unit root test that may be used.

For practical purposes stationarity can usually be determined from a run sequence plot (NIST/SEMATECH, 2010). Run sequence plots are an easy way to summarize a data set. Shifts in location and scale are typically quite evident. Furthermore, outliers can easily be detected using a run sequence plot (NIST/SEMATECH, 2010). The graph is formed by plotting the response variable on the vertical axis, and the observation index to the horizontal axis. For a constant location and scale the response should appear constant +/- a random error (NIST/SEMATECH, 2010). Although being very useful for changes in mean and variance, the run sequence plot is not suitable for addressing autocorrelation.

Two graphical techniques for assessing sample independence are the autocorrelation plot and the lag plot (NIST/SEMATECH, 2010; Law & Kelton, 2000). The autocorrelation plot is a graph of the sample autocorrelations for data values at varying time lags. A lag is a fixed time displacement: a plot of lag 1 is a plot of the values $Y_i$ versus $Y_{i-1}$ (IST/SEMATECH, 2010). It is important to note, that the indicated sample correlation will not be 0 even when the data is independent. Only if the data differs significantly from 0, strong evidence of correlation exists. Another indication is a specific pattern in the correlation plot, such as a linear trend (Law & Kelton, 2000). The lag plot of lag 1 of the observations $X_1$, $X_2$, …, $X_n$ is a scatter

diagram of the pairs ($X_i$, $X_{i+1}$) for i = 1, 2, …, n – 1. The nature of the acquired scattering will depend on the underlying distributions of the $X_i$'s. A positive correlation generally leads to points lying along a line with positive slope in the first quadrant, whereas negative correlation generally leads to points lying along a line with negative correlation in the first quadrant (Law & Kelton, 2000). A disadvantage of the lag plot is that it only assesses one lag.

Besides these graphical techniques and the methods based on linear regression, several nonparametric statistical tests exist that can be used to test whether the data is independent. One such test is the rank version of the Von Neumann's ratio test (Bartels, 1982). The test assumes that there are no ties in the data; i.e., no observations have the same value. This assumption may not be satisfied if data are recorded with only few decimal places of accuracy. Nevertheless the critical values may still be reasonable accurate in case the number of ties is relatively small (Bartels, 1982). Alternatively run tests can be applied, which only require two subsequent observations not to be tied (Gibbons & Chakraborti, 2003). However, if the assumption holds, the rank von Neumann test is shown to be more powerful than the run tests (Bartels, 1982). Furthermore, two tests are available based on the autocorrelation plot, namely the Box-Ljung test and the Pankrantz criterion (Garson, 2010). The Pankrantz criterion states that the autocorrelation divided by its standard error must be less than 1,25 for the first three lags and less than 1,60 for subsequent lags, in order to conclude the series has no significant autocorrelation. Instead of testing randomness at each distinct lag, the Box-Ljung test assesses the overall randomness based on a number of lags (Garson, 2010).

### *A.1.2 Sample Selection*

The standard method to test the hypothesis that the means among two or more groups are equal is the one-way Analysis of Variance (ANOVA). The analysis of variance decomposes the variability of the observed data into two components: a between-group component and a within-group component. If the estimated variability between groups is significantly larger than the estimated variability within groups, it is evidence that the group means are not all the same (Green & Salkind, 2004). The limitation of this method is that normally distributed samples are assumed. Similarly, methods to derive confidence intervals of an ANOVA, such as Fisher's LSD and Tukey's HSD, are constraint to normally distributed data (Montgomery & Runger, 2002).

Alternatively nonparametric procedures may be used, which compare the sample medians rather than the means. This is especially appropriate if it is suspected that outliers may be present or the data is skewed. These tests are the Kruskal-Wallis test, the Friedman test, and Mood's median test. The Kruskal-Wallis test is most appropriate when each set contains a random sample from its population (the sets have no intrinsic meaning). The Friedman test is more appropriate when each set represents a block, i.e. the level of some other variable, such as a day of the week. A disadvantage is that the method requires different samples to be of exactly the same size (Statpoint Technologies, 2009). Mood's median test is more general than the other tests, but consequently has considerably less power (Green & Salkind, 2004).

In order to test whether various samples originate from the same distribution based on their standard deviations, other tests can be applied. Bartlett's test is a commonly used test for equal variances. Equality of variances across samples is called homogeneity of variance. The test is not robust; it is very sensitive to departures from normality (NIST/SEMATECH, 2010). Levene's test offers a more robust alternative to Bartlett's procedure. This means it will be less likely to reject a true hypothesis of equality of variances just because the distributions of the sampled populations are not normal. If strong evidence exists that data does come from a normal or nearly normal distribution, then Bartlett's test has better performance (NIST/SEMATECH, 2010).

Levene's original method only proposed using the mean. Brown and Forsythe extended Levene's test to use either the median or the trimmed mean in addition to the mean. Their research indicated that using the trimmed mean performed best when the underlying data followed a heavy-tailed distribution and the median performed best when the underlying data followed a skewed distribution. Using the mean provided the best power for symmetric, moderate-tailed, distributions (NIST/SEMATECH, 2010).

Although the optimal choice depends on the underlying distribution, the method based on the median is recommended as the choice that provides good robustness against many types of non-normal data while retaining good power. If detailed knowledge of the underlying distribution of the data is obtained, this may indicate using one of the other choices (NIST/SEMATECH, 2010). Alternatively nonparametric procedures may be used, which compare the sample medians rather than the means.

### *A.1.3 Goodness-of-Fit Tests*

The classical goodness-of-fit hypothesis test is the chi-square test. It tests whether the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. Therefore the entire range of the fitted distribution is divided into intervals (Law & Kelton, 2000). A difficulty with this test is that no definitive prescription can be given about the number and size of the intervals that is guaranteed to produce good results in terms of validity. Depending on the interval setting different conclusions can be reached based on the same data set. A benefit of the chi-square test is that it can be applied to any hypothesized distribution for which you can calculate the cumulative distribution function (NIST/SEMATECH, 2010).

A different goodness-of-fit test is the Kolmogorov-Smirnov test. It compares the empirical distribution function of the sample with the cumulative distribution function of the hypothesized reference distribution, by quantifying their maximum difference (Law & Kelton, 2000). As such, this test does not require the data to be grouped, and consequently no information is lost. An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested (NIST/SEMATECH, 2010). Another advantage of Kolmogorov-Smirnov tests is that they tend to be more powerful than chi-square tests against many alternatives (Law & Kelton, 2000). A drawback is that their applicability is more limited than that of the chi-square tests. The original form of the Kolmogorov-Smirnov test is valid only “if all the parameters of the hypothesized distribution are known and the distribution is continuous; i.e., the parameters cannot have been estimated from the data” (Law & Kelton, 2000, p. 363). Though, the test has been extended to allow for

estimation of the parameters in the case of normal, exponential, Weibull, lognormal, or log-logistic distributions (Law & Kelton, 2000). Furthermore, the required critical values for discrete data can be calculated using advanced formulas. A possible drawback of the test is that it gives the same weight to any difference, whereas many often applied distributions differ primarily in their tails (Law & Kelton, 2000).

The third goodness-of-fit test, the Anderson-Darling test, also compares the empirical distribution function of the sample with the cumulative distribution function of the hypothesized reference distribution, by quantifying their difference. However, a difference is that it is designed to specifically detect discrepancies in the tails, while the Kolmogorov-Smirnov is more sensitive to differences near the center of the distribution (NIST/SEMATECH, 2010). Furthermore, the Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution (NIST/SEMATECH, 2010). These critical values are available for the same distributions as for the Kolmogorov-Smirnov test (Law & Kelton, 2000).

### *A.1.4 Comparing Two Samples*

When a load file (an estimated sample defined by the client) is used in a simulation model instead of an input distribution, input validation will be based on comparing two samples. The most powerful tests for comparing two independent samples are the Student's t-test for equality of means and the F-test for equality of variances (Gibbons & Chakraborti, 2003). A downside is that both tests assume that the populations are normally distributed. This assumption may not hold; arrival rates and service times are typically characterized as an exponential process. However, in case the sample size is sufficiently large, the random variables might be approximately normally distributed due to the central limit theorem. Furthermore, Rasch et al. (2007) state that the t-test is robust against the normality assumption to such an extent that it can be recommended in nearly all applications.

A different test which may be applied, the Wald-Wolfowitz runs test, is an extremely general rank test, consistent with all types of differences in populations, e.g. differences in means, variability, scale, or location (the model is shifted) (Gibbons & Chakraborti, 2003). Though, due to its generality, the test lacks power (Gibbons & Chakraborti, 2003). Similar reasoning is valid for the two-sample variant of the chi square test. Furthermore, as for the one-sample variant, no definitive prescription can be given about the number and size of the intervals. Therefore alternative nonparametric tests, e.g. the two-sample Kolmogorov-Smirnov test and the Mann-Whitney U test, are often applied (Cooper & Schindler, 2003).

The two-sample Kolmogorov-Smirnov test is an adaption from the one-sample test discussed earlier. As for the one-sample version, the test is based on the maximum absolute difference between cumulative distribution functions. If the two samples have been drawn from the same population, the cumulative distributions of the samples should be fairly close to each other, showing only random deviations from the population distribution (Cooper & Schindler, 2003). Analogous to the Wald-Wolfowitz test, the test is nonparametric and sensitive to differences in both location and shape of the empirical cumulative distributions functions of the two samples (Gibbons & Chakraborti, 2003).

The Mann-Whitney U test is an alternative to the t-test without its limiting assumption of normality (Cooper & Schindler, 2003). It is a rank test that evaluates whether the medians on a test variable differ significantly between two samples (Green & Salkind, 2004). It is an extension of the Wilcoxon rank-sum test such that it does not require equal sample sizes (therefore it is also called the Mann-Whitney-Wilcoxon test) (Cooper & Schindler, 2003).

Alternatives to the F-test for equality of variances have already been discussed in appendix A.1.2. These tests are suitable for two or more samples and therefore are also applicable in this situation.

## A.2 Additional Information of Trace-Driven Output Validation

In this section a more in-depth discussion can be found about various methods concerning trace-driven output validation.

### A.2.1 The Correlated Inspection Approach

The first method, called the correlated inspection approach, prescribes to "compute one or more statistics from the real-world observations and corresponding statistics from the model output data, and then to compare the two sets of statistics without the use of a formal statistical procedure" (Law & Kelton, 2000, p. 283). Examples are the sample mean, the sample variance, the sample correlation function and graphical plots (Law & Kelton, 2000). Sargent (1996) and Kleijnen (1995) discuss several graphical methods, namely histograms, box plots and behavior graphs, which can be used for operational validity. These do not require the data to be independent, have no distributional requirements on the data, and can be used with a limited number of observations (Sargent, 1996).

Though the correlated inspection approach does not use a formal statistical procedure to compare real-world and simulation statistics, it may provide valuable insights into the adequacy of a simulation model and it will often be the only feasible statistical approach due to limitations on available data (Law & Kelton, 2000). Due to the lack of a formal, objective procedure to compare the two sets of data, determining whether the model has sufficient accuracy for its intended purpose should be done subjectively. This comparison can be made by the model development team and/or by experts using face validity or Turing tests (Kleijnen, 1995; Sargent, 1996). A Turing test is performed by asking people knowledgeable about the system to examine and identify one or more sets of system data as well as one or more sets of model data without knowing which sets are which (Law & Kelton, 2000). Related to this, a model is said to have face validity when simulation results are consistent with perceived system behavior (Law & Kelton, 2000).

### A.2.2 Confidence-Interval Approach

When it is possible to collect a potentially large amount of data from both the model and the system it is possible to create confidence-intervals based on output differences. This is a more reliable approach for comparing a model with the corresponding real system (Balci, 1998; Law & Kelton, 2000). The

combination of confidence intervals of various output differences is also called the model's range of accuracy (Balci, 1998).

Because the model and system output are dependent (the same input values have been used), the paired-t approach should be used for creating confidence intervals for the differences in responses. This method pairs dependent observations, and therefore requires the amount of observations of system output and model output to be equal. Furthermore, the paired-t method assumes the response differences to be independent and identical (IID) random variables, and normally distributed (Law & Kelton, 2000). It is important to note that the responses should be random variables over entire independent replications (e.g. a single day in a terminating system). As a result the data is IID as required (Law, 2007; Kleijnen, 1995). The method is quite robust for deviations of the normality assumption; the central limit theorem applies (also when autocorrelation exists) (Kleijnen, 1995; Law & Kelton, 2000), which means that the coverage probability will be near $1 - \alpha$ for a large number of observations (with $\alpha$ being the probability of a type I error). Therefore the test may still be applied in case of non-normality (Kleijnen, 1995). In contrast to the classical two-sample-t approach, Var(model) = Var(system) is no prerequisite (Law & Kelton, 2000).

A $100(1 - \alpha)$ percent confidence interval is statistical significant at level $\alpha$ in case the interval does not contain 0. When it does contain 0 any observed difference may be explained by sampling fluctuation (Law & Kelton, 2000). As discussed, differences are to be expected and do not necessarily imply that the model is invalid. Therefore practical significance is defined as the magnitude of the difference being large enough to invalidate any inferences about the system that would be derived from the model (Law & Kelton, 2000). As for the inspection approach, the decision whether the difference between the model and the system is practically significant, is a subjective one, and should be decided on by the model development team or expert.

### *A.2.3 Regression Based Approach*

In Kleijnen (1999) and Kleijnen et al. (1998) two validation approaches for trace-driven simulations are discussed, which are based on a standard regression analysis. The classical approach is to make a scatter plot with x and y simulated and real outputs that use the same input, subsequently fit a line $y = \beta_0 + \beta 1\ x$, and test whether $\beta_1 = 1$ and $\beta_0 = 0$ (Kleijnen, 1999). Concerning this approach two possible tests can be performed. The less stringent variant only requires a strong correlation between system and model output; H0: $\beta_1 \leq 0$. The null-hypothesis is rejected and the simulation model accepted if there is strong evidence that the simulated and real responses are positively correlated (Kleijnen, 1995). Alternatively, the more stringent method requires the means of the simulated and real response to be identical and the correlation to be 1; $H_0$: $\beta_1 = 1$ and $\beta_0 = 0$. The simulation model is valid if the null-hypothesis is accepted (Kleijnen, 1995).

The more stringent classical approach has been applied widely in practice (Kleijnen et al., 1998). Nonetheless it appears that it rejects a valid model too often (Kleijnen, 1999). Therefore a novel approach has been developed in Kleijnen et al. (1998). This approach computes the n differences (d), as well as the n sums ($q_i = x_i + y_i$). Subsequently, a line $d = \gamma_0 + \gamma_1 q$ is fit to the n pairs ($d_i$, $q_i$). The related

null-hypothesis is $H_0$: $\gamma_0 = 0$ and $\gamma_1 = 0$. This hypothesis implies that $\mu_x = \mu_y$. Furthermore, $\gamma_1 = 0$ implies equal variances (Kleijnen, 1999). Standard regression software can be used to test this hypothesis (Kleijnen, 1999).

When serious non-normality exists (for instance in case of short runs), a regression analysis can be created based on bootstrapping (Kleijnen et al., 2001). Bootstrapping generates observations by random resampling with replacement from the original observations (Kleijnen, 2008). Because of the correlation between simulation and system output an observation in this context implies the correlated pair ($x_i$, $y_i$) (Kleijnen et al., 2000). By repeating the resampling many times (typically 100 or 1000 times) the sampling variation can be reduced. Note that, instead of generating responses through bootstrapping, more simulation responses may be generated (Kleijnen et al., 2000). However, this generally requires much more computer time than applying bootstrapping (Kleijnen et al., 2000).

### *A.2.4 Time-Series Approaches*

The spectral-analysis approach is a sophisticated technique that "proceeds by computing the sample spectrum, i.e. the Fourier cosine transformation of the estimated autocovariance function, of each output process and then using existing theory to construct a confidence interval for the difference of the logarithms of the two spectra" (Law & Kelton, 2000, p. 289). It can be used to evaluate the degree of similarity of the two autocorrelation functions, without making assumptions about the distributions of the observations in the time series (Law & Kelton, 2000).

Alternatively the time-series approach of Hsu & Hunter (1977) can be used, which consists of "fitting a parametric time-series model to each set of output data and then applying a hypothesis test to see whether the two models appear to be the same" (Law & Kelton, 2000, p. 290). As discussed before, a downside is that it is a hypothesis-test, which gives no additional insights in case the data appears to be different.

## *A.3 Additional Information of Conducting a Sensitivity Analysis*

In this section a more profound debate of various methods related to a sensitivity analysis can be found.

### *A.3.1 Metamodels*

In a sensitivity analysis, the simulation model is run for the set of factor combinations and the resulting input-output data are analyzed to estimate a metamodel (Kleijnen, 1997). A metamodel is a model or approximation of the implicit input/output function (Kleijnen et al., 2004). If at least some of the input parameters are quantitative, and if a performance measure can be clearly stated, then it is possible to construct metamodels of the performance that describe the I/O relationships in terms of functions of various parameter values (Kleijnen et al., 2004). In a simulation study generally several outputs are distinguished, which correspond to multiple response variables. However, current practice uses metamodels with a single output; a separate metamodel is developed for each output (Kleijnen & Sargent, 2000).

Various different metamodels can be distinguished. Metamodels containing a relatively small amount of complexity, and are the most popular, are linear regression metamodels (or first-order polynomials) (Kleijnen, 2009). Examples of other types are: second-order polynomials, the more complex Kriging models (an interpolation method that predicts unknown values of a random process), neural nets (a type of nonlinear regression), radial basis functions, splines (which partition the domain of applicability into sub domains and fit simple regression models to each of the sub domains), support vector regression, and wavelets (Kleijnen & Sargent, 2000). The model required is dependent on the goal of the metamodel; e.g. understanding, prediction, or optimization. Corresponding to this goal the size of the experimental area or frame, for which the metamodel is valid, varies (e.g. locally or globally). Zeigler (in Kleijnen, 1995, p. 157) defines the experimental frame as "a limited set of circumstances under which the real system is to be observed or experimented with." For instance in case of a nonlinear input/output function, a first-order polynomial metamodel can only be used for a local area. Montgomery (1991) states that the first-order polynomial will work quite well, even when the linearity assumption holds only very approximately. Though, when the area gets bigger, a second-order polynomial might be required, as illustrated in Figure 25. When the experimental area covers the whole area in which the simulation model is valid, global, more complex metamodels become relevant (Kleijnen, 2005).

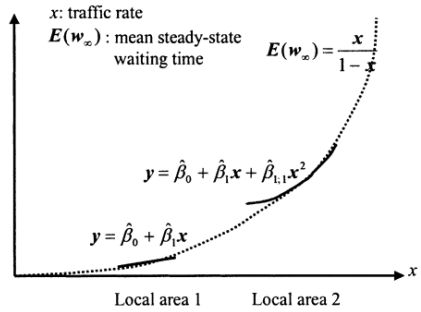



**Figure 25: A first- and second-order polynomial metamodel example of an M/M/1 queue (Source: Kleijnen, 2005)**

For determining how sensitive the output is to changes in the input, a local experiment is sufficient; the aim is to evaluate the effects of relatively small deviations in input variables, due to inaccuracy of the simulation model. Furthermore, the exact value is of minor importance; the relative impact of a factor on the output with regard to other factors is what matters most. Consequently linear regression models may be used.

### *A.3.2 A Full Factorial Design*

For experiments involving the study of the effects of two or more factors, factorial designs are most efficient (Montgomery, 1991). A factorial design is the investigation of all possible combinations of the levels of the factors in each complete trial or replication of the experiment (Box et al., 2005; Montgomery, 1991). It is possible to evaluate the effect on several responses in one experimental run. In contrast to the one-factor-at-a-time approach, the factorial design is capable of taking interactions among factors into account (Law & Kelton, 2000).

The effect of a factor is defined as "the change in response produced by a change in the level of the factor" (Montgomery, 1991, p. 197), and is often called the main effect since it refers to the primary factors of interest in the experiment (Montgomery, 1991). When the difference in response between the levels of one factor is not the same at all levels of the other factors, an interaction between the factors exists (though, small deviations may exist due to randomness in the simulation model). When an interaction is large, the corresponding main effects may have little practical meaning (Montgomery, 1991). Magnitude and direction of the factor effects are examined in order to determine which variables are likely to be important (Box et al., 2005).

The factorial design associated with a first-order polynomial is an experiment with two levels for all k factors. This is called a $2^k$ factorial design due to the $2^k$ possible factor-level combinations. A specific combination is also called a design point (Law & Kelton, 2000). The statistical model would contain $2^k - 1$ effects; k main effects, $\binom{k}{2}$ two-factor interactions, $\binom{k}{3}$ three-factor interactions, …, and one k-factor interaction (Montgomery, 1991). Factorial designs can also be used for second-order polynomials. This results in a $3^k$ design or a central composite design, which adds a central combination to the factorial design (Kleijnen, 1999).

How the factors are varied over the different runs is specified in a design matrix. Generally a + and a – are used to denote the high and low level of a factor, respectively (Montgomery, 1991). For these matrices standard lay-outs exist which are called standard orders or Yates orders (Box et al., 2005). No specific prescription can be given how the levels should be specified. Generally, one should select reasonable levels for the objective of the study, which can be based upon expert opinions (Law & Kelton, 2000).

A downside of the full factorial design is that the amount of runs required tends to become large, when testing an increasing amount of factors. A relatively large part of these runs is required for determining the many degrees of freedom that are associated with higher interactions, which are often negligible (Montgomery, 1991). For example a complete $2^6$ design requires 64 runs, of which 6 of the 63 degrees of freedom are related to main effects, 15 degrees of freedom correspond to two-factor interactions, and the remaining 42 degrees of freedom are associated with three-factor and higher interactions (Montgomery, 1991).

### *A.3.3 Fractional Factorial Design*

A $2^{k-p}$ fractional factorial design contains a $\left(\frac{1}{2}\right)^p$ fraction of the $2^k$ design (Montgomery, 1991). Several important fractional factorial designs have been classified (Box et al., 2005; Law & Kelton, 2000; Montgomery, 1991).

- Resolution III designs (denoted as $2^{k-p}_{III}$), which are defined as "designs in which no main effects are aliased with any other main effect, but main effects are aliased with two-factor interactions and two-factor interactions may be aliased with each other" (Montgomery, 1991, p. 339). A resolution III design can be used to investigate up to k = N – 1 factors, in only N runs, where N is a multiple of 4 (Montgomery, 1991).

- Resolution IV designs (denoted as $2_{IV}^{k-p}$). These are "designs in which no main effect is aliased with any other main effect or with any two-factor interaction, but two-factor interactions are aliased with other" (Montgomery, 1991, p. 339). Any resolution IV design must contain at least 2k runs (Montgomery, 1991). It can be obtained by adding the mirror image to a resolution III design; each run is an opposite of an earlier run (Kleijnen, 1997).
- Resolution V designs (denoted as $2_{V}^{k-p}$), which can be defined as "designs in which no main effect or two-factor interaction is aliased with any other main effect or two-factor interaction, but two-factor interactions are aliased with three-factor interactions" (Montgomery, 1991, p. 339).

It should be noted that not all resolution designs are possible for every k (Law & Kelton, 2000).

### *A.3.4 Plackett-Burman Designs*

Plackett-Burman designs, experimental arrangements derived by Plackett and Burman, make it possible to investigate k factors in as few as N = k + 1 runs, provided that N is divisible by 4 (Box et al., 2005; Law & Kelton, 2000; Montgomery, 1991). Fractional factorials can be used when the number of runs are a power of 2 (4, 8, 16, 32, 64, etc.). Plackett-Burman designs can be used at intermediate values; as indicated they are available for any N that is a multiple of 4 (in particular for N = 12, 20, 24, 28, 26) (Box et al., 2005). An important downside of these designs is that they have very complicated alias structures (Box et al., 2005; Montgomery, 1991). As a result it is for instance not possible, in case a main effect appears to be important, to know which specific interaction terms it is aliased with (Montgomery, 1991; Steppan et al., 1998). Note that, when k + 1 is a power of 2, the Plackett-Burman design coincides with the $2_{III}^{k-p}$ fractional factorial design (Kleijnen, 2005; Law & Kelton, 2000).

In case a first order polynomial metamodels is used, these designs are called saturated designs. In a saturated design the number of factor combinations (n) equals the number of metamodel parameters. Thus for a first order polynomial metamodel a saturated design means n = k + 1 (Kleijnen et al, 2004a).

### *A.3.5 Supersaturated Designs*

Designs are called supersaturated when they aim to estimate more effects than they have runs (Kleijnen, 2005; Trocine & Malone, 2001). Examples are the random-balanced supersaturated design, and the systematic supersaturated design (Law & Kelton, 2000). Holcomb et al. (2003) conclude that supersaturated designs should be used with caution. Trocine & Malone (2001) add to this that you might get good results when you are lucky: i.e. the right design needs to be chosen, the factors and columns in the design need to match the factors, and the right analysis method should be chosen and used in the right way. This is due to the complexity of the confounding effects and the insufficient degrees of freedom to apply traditional analysis such as regression (Trocine & Malone, 2001).

### A.3.6 Additional Group Screening Designs

#### A.3.6.1 Two-Stage Group Screening

For two-stage group screening the experimenter uses experience and knowledge of the problem and the factors to arrange the factors into logical groups (Trocine & Malone, 2001). A fractional factorial design is run on the groups in order to indentify the important ones. Subsequently, a new fractional design is run on the factors or subgroups within an important group until the important factors are identified. The method is iterative since the results of the first stage are used in the second stage. Note that interactions between factors in different groups are not measured and if they exist may confound the results of the groups (Trocine & Malone, 2001).

In order to avoid cancellation of factors and to detect as many of the effective factors as possible Ivanova et al. (1999) identified several rules of thumb.

- A factor with an unknown direction of effect should be placed alone in a group.
- Factors with assumed important positive effects should be placed in one group.
- Factors with assumed small effects and the same direction should be placed in a group.
- Factors with possible effects and the same direction should be placed in a group.
- Resolution IV designs should be used to calculate main effects unbiased by possible second-order interactions.

#### A.3.6.2 Sequential Bifurcation

Sequential bifurcation (SB) is, as the name states, a sequential design; it is constructed one point at a time using feedback from all prior points to direct the search for important factors (Trocine & Malone, 2001). It starts with only two scenarios, namely one with all individual factors at the level -1, and a scenario with all factors at +1. If the difference between the responses is considered to be significant, it means that some factors in the group are significant and the procedure continues (Trocine & Malone, 2001). In the next step SB splits (bifurcates) a significant group halfway. Groups, and all its individual factors, are eliminated from further experimentation as soon as the group effect is statistically unimportant. As SB proceeds the groups get smaller and it stops when the first-order effects of all important individual factors are estimated (Kleijnen, 2005). Several heuristic rules exist to decide on how to assign factors to groups (Kleijnen et al., 2004b).

Sequential bifurcation is a very efficient method for screening important factors (Kleijnen, 2005; Trocine & Malone, 2001). Though it is vulnerable to violations of the prerequisite that all the effects of the factors within a group should have the same direction (all effects are positive or all effects are negative) (Trocine & Malone, 2001). Though in principal this method is not capable of identifying interactions, an extension is available, called CSB-X, which takes second order effects into account as well (Wan & Ankenman, 2006).

### *A.3.6.3 Iterated Fractional Factorial Design*

The iterated fractional factorial design starts with the construction of a fractional factorial design and with factors randomly assigned to groups. The method introduces a third intermediate level on which 25 % of the runs is put, while the other runs are equally split to the low and high level. Subsequently the process is repeated with factors assigned again randomly to the groups. After several iterations the data is analyzed using forward stepwise regression (Trocine & Malone, 2001). According to Andres and Hajas (1993) the method is designed for very large problems and works best when a very small number of factors dominate. Trocine & Malone (2001) confirm this and state that the method may not be efficient for as few as 100 factors.

### *A.3.6.4 Controlled Screening Designs*

Controlled screening methods are not driven by a predetermined amount of runs to be performed but by the desired type I error (the probability of declaring a factor important when it is not) and power (the probability of correctly declaring a factor important). It is called controlled screening due to the statistical control imposed on the experiments (Wan & Ankenman, 2006). Given the amount of statistical control the number of replications is minimized.

Controlled screening methods are developed for fractional factorial designs and sequential bifurcation. For example concerning fractional factorial designs, the total number of runs can be decreased by only replicating certain rows of the factorial design (additional to a small amount of base replications) for which the responses have high variance (Wan & Ankenman, 2006). The research of Wan & Ankenman (2006) indicated that the controlled screening method based on fractional factorial designs requires less prior knowledge of factor effects and is more efficient than the method based on sequential bifurcation when the percentage of important factors is 5% or higher.

## *A.3.7 Variance Reduction Techniques*

Common random numbers (CRN) strives to compare alternative configurations under similar experimental conditions, so that any observed difference in performance is due to differences in the system configurations rather than to fluctuations of the experimental conditions (Law & Kelton, 2000; Farrington & Swain; 1993). CRN tries to induce a positive correlation by generating corresponding random variables across simulations from the same random numbers (Glasserman & Yao, 1992).

In order to illustrate the effect, consider two systems consisting of random variables X and Y. The effort required to obtain a valid estimate of difference depends on the variance of f(X) – g(Y), where

$$\mathrm{Var}[f(X) - g(Y)] = \mathrm{Var}[f(X)] + \mathrm{Var}[g(Y)] - 2\,\mathrm{Cov}[f(X), g(Y)]$$ (Glasserman & Yao, 1992).

Var [f(X)] and Var[g(Y)] are determined by the individual distribution of X and Y. When X and Y are simulated independently their covariance is zero. CRN attempts to reduce the variance by introducing a positive dependence between f(X) and g(Y) (Glasserman & Yao, 1992; Farrington & Swain; 1993).

From the above equation it can also be concluded that it is possible for CRN to backfire; i.e. when CRN induces a negative correlation it leads to an increase in the variance (Law & Kelton, 2000; Glasserman & Yao, 1992). According to Glasserman & Yao (1992, p. 904) "variance reduction is guaranteed (in comparing throughputs and in some cases sojourn times and queue lengths) whenever changing the order of some events does not radically change the evolution of the systems. This is the case for most standard queuing systems with a single class of jobs and a first-come-first-served discipline, but not for most multiclass networks or queues with, e.g., pre-emptive disciplines."

Another possible drawback to CRN is that formal statistical analysis can be complicated by the induced correlation (Law & Kelton, 2000). In Kleijnen (2008) several alternatives are evaluated. As is appropriate in case no CRN is used, one can calculate the ordinary least square (OLS) estimators. Alternatively, estimated generalized least squares (EGLS) can be applied, which may give better point estimates of the factor effects, but requires many replicates.

Furthermore, for CRN to work, the random numbers across the different system configurations on a particular replication should be properly synchronized (Law & Kelton, 2000). "Ideally, a specific random number used for a specific purpose in one configuration is used for exactly the same purpose in all other configurations" (Law & Kelton, 2000, p. 586). It is generally not enough to start the simulations of all configurations with the same seed of random number stream; multiple streams should be used with specific streams dedicated to producing the random numbers for each particular type of input random variate (Law & Kelton, 2000).

# Appendix B. Additional Case Study Information

Within this appendix additional results will be presented related to the case study.

## B.1 Cairo Terminal 3

In this section more details will be presented about the case study. The first part will focus upon the configuration of the model itself and the simulation. Subsequently, the standard simulation results of the case study will be discussed.

### B.1.1 Model Configuration

The simulation model of the terminal was built for the purpose of animation; e.g. to illustrate the baggage handling of the designed system. Therefore the level of detail contained in the model is relatively low. It was particularly important that the model's appearance met the client's expectations. The model has been verified for animation purposes. This means that it handles bags correctly. Though, model parameter adequacy was of lesser importance. Therefore the model has to be verified again with respect to its parameters before validation methods can be applied, due to changes within the objective of the model. In order to evaluate whether parameters are specified properly, information will be combined of various sources; the Terminal 3 BHS design (Vanderlande Industries, 2009b), the product database (Vanderlande Industries, 2010) and opinions of simulation engineers at Vanderlande.

#### System Components

In general, not all parameters of system components can be acquired by evaluating design specifications. The missing parameters may be calculated by using a formula.

The capacity of system components is determined by two main factors; the conveyor speed and the window length of a conveyor (the space reserved for a bag). This relation is defined in the following formula, where the velocity of the belt is multiplied by 3600 due to a difference in units of measurements (conversion from seconds to hours).

$Capacity\ (b/h) = \frac{3600 \cdot V_{Belt}}{L_{Window}}$ (Vanderlande Industries, 2010)

LWindow = LBag + LGap (m)
VBelt = speed of the conveyor (m/s)
LWindow = the window length of a conveyor (m)
LBag = baggage length (m)
LGap =minimum gap between baggage for sorting (m)

Within the simulation model all components are set at a speed of 1 m/s, with a window of 2 m, relating to a capacity of 1800 bags per hour. This can be compared to the values of the design and the product database. In Table 17 the capacity, velocity, and window length can be observed per component as specified within the system design and product database. Note that window lengths are given in neither source. These have to be calculated by applying the formula described above. The product database

specifies ranges in some instances, relating to different application areas of the component and environmental conditions.

| | Design | | | Product Database | | |
|---|---|---|---|---|---|---|
| Component | Cap. (b/h) | V (m/s) | $L_{Window}$ (m) | Cap. (b/h) | V (m/s) | $L_{Window}$ (m) |
| Belt Floorveyor | 1500 | - | - | - | 0,33 - 2 | - |
| Belt In Tracking | 1200 | - | - | - | - | - |
| Stop-Start Conveyor | - | - | - | - | 1 | - |
| Belt Curve | - | - | - | 1500 | 1 | - |
| 90% Transfer with Straight Conveyors | - | - | - | 1200 – 1800 | 1 | - |
| L1 Screener | 1200 | - | - | - | - | - |
| L3 Screener | 300 | - | - | - | - | - |
| Vertisorter | 1800 | - | - | 1650 - 2100 | 1 | - |
| 90% Induct | - | - | - | 1200 – 1800 | 1 | - |
| Flat Triplanar Sorter | 1500 | - | - | - | 0,5 – 1,33 | - |
| Divert Parallel Pushers | 2000 | - | - | 1600-1800 | - | - |
| EBS | >120 m. | - | - | - | - | - |

**Table 17: Relevant BHS component characteristics**

The standard conveyors (belt floorveyors) can be set to a velocity of 0,33 to 2 meters per second. Though, the maximum speed of a mainline at a merge for instance is 1,25 meter per second (Vanderlande Industries, 2010). As is the case for this model, it is common to use a speed of 1 meter per second within simulation. Applying this speed and a capacity of 1500 bags per hour to the formula discussed, results in a window length of 2,4 meters.

A part of the belt Floorveyors is equipped with a system for tracking baggage items. Tracking generally requires larger gaps between bags, resulting in larger windows (Vanderlande Industries, 2010). Though capacity levels up to 1800 bags per hour are still supported, the capacity levels of belts in tracking are set lower than the regular Floorveyors (Table 17). The lower capacity, in combination with a speed of 1 meter per second, relates to a window length of 3 meters. Within the material flow diagram, Floorveyors in red or pink are kept in tracking.

The above specified parameters are in line with stop-start conveyors (among others used for buffering) and curve, transfer, and induct restrictions (Table 17).

The design states that level 1 and level 3 screeners have a capacity of respectively 1200 and 300 bags per hour. For level 1 this means that the capacity is equal to the surrounding Floorveyors in tracking. Though, this capacity level relates to dissimilar speeds and window lengths. Screening machines generally use a low velocity to make accurate scans of baggage items. The speed is however constrained by the amount of stop-start conveyors directly in front of it; the maximum speed step between two conveyors is 0,25 meter per second (Vanderlande Industries, 2010). With two stop-start conveyors being placed in front of the screeners this means that the minimum speed is 0,25 meters per second. Reduced speed generally implies that windows lengths may be smaller (due to a reduced probability of sliding) (Vanderlande Industries, 2010). Though, window lengths are constrained by bag lengths (1 meter in this case). Combining this information leads a velocity of 0,33 meter per second and a window length of 1 meter for level 1 screening, and a velocity of 0,25 and a window length of 3 meters for level 3 screening. Note that a window length of 3 meters is actually large following above reasoning. In practice it might be

that a bag is required to be put to a hold for some time, relating to a smaller window length while maintaining a capacity of 300 bags per hour. This option however will not be applied within the simulation, due to some added complexity that this implies, while yielding similar results.

The Flat Triplanar Sorter is configured similar to the Belt Floorveyors without tracking. The Divert Parallel Pushers installed aside of it have a capacity of 1600 to 1800, or 2000 bags per hour, depending on the source of information. It is important to note that either way, the pushers have a higher capacity than the sorter. Consequently, every bag on the sorter could be diverted by the same pusher; no recirculation of bags is required within the simulation model. It is assumed that the capacity of container loading at a lateral is always sufficient.

Concerning the EBS, the design specifies two Early Bag Stores (left and right), that both contain two storage lanes of minimal 60 meters (Vanderlande Industries, 2009b). “When baggage is received from the sorter it will be accumulated with small gaps to maximize storage capacity” (Vanderlande Industries, 2009b, p. 48). The stopping space for baggage items in the EBS in the simulation model is 1,5 meters. Gaps of 0,5 meters are not considered very small. Therefore this size is decreased to 0,2 meter. The effect of such a reduction will be evaluated within the sensitivity analysis.

Vertisorter capacity levels depend on the amount of position switches that are required (directing a bag up or down). In Figure 26 the capacity (bags per hour) of a Vertisorter can be observed, as a function of the amount of switches required and the average bag length. Additional switches and larger average bag lengths reduce the capacity of the sorter. Typically a capacity level of 1800 bags per hour is assumed. With a bag length of 1 meter and a redirected flow of 5 % (see subsequent section) this assumption seems justified. With a speed of 1 meter per second this implies a window length of 2 meters.

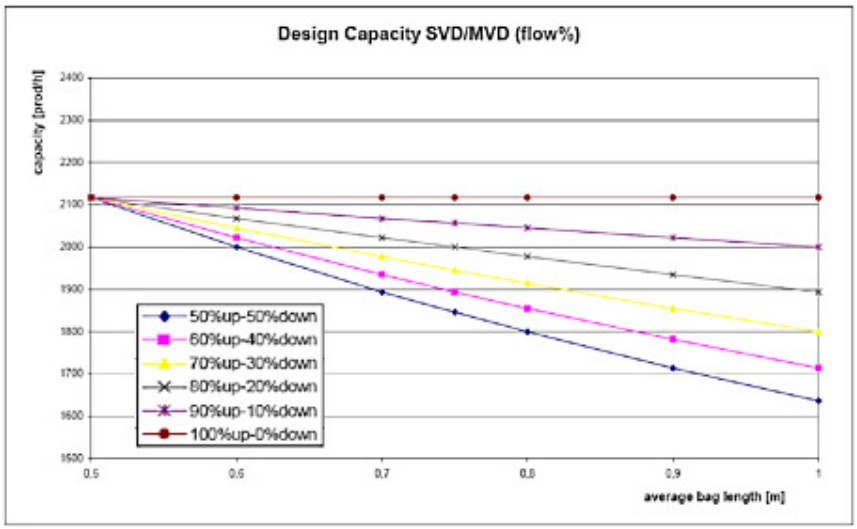



Figure 26: Design capacity of a vertisorter / vertimerge (Vanderlande Industries, 2010)

Above information is summarized in Table 18, which shows the capacity, velocity, and window length per component, as set within the updated simulation model. Curves, transfers and inducts are not shown. They are set identical to surrounding Floorveyor speeds.

| Component | Simulation Model | | |
|---|---|---|---|
| | Cap. (b/h) | V (m/s) | $L_{Window}$ (m) |
| Belt Floorveyor | 1500 | 1 | 2,4 |
| Belt In Tracking | 1200 | 1 | 3 |
| Speed Reduction 1 | 1200 | 0,75 | 2,25 |
| Speed Reduction 2 | 1200 | 0,5 | 1,5 |
| L1 Screener | 1200 | 0,33 | 1 |
| L3 Screener | 300 | 0,25 | 3 |
| Vertisorter | 1800 | 1 | 2 |
| Flat Triplanar Sorter | 1500 | 1 | 2,4 |
| Divert Parallel Pushers | >1500 | - | - |
| EBS Lane | >100 bags | 1 | 1,2 (V=0) |

**Table 18: Parameter specifications of the simulation model**

Some remarks can be made. A component not yet assessed is the ACRS. No information was available of this component. It is assumed that its capacity is non-restricting. Furthermore, as discussed, the check-in desks and collector belts are out of scope. In order to give a complete overview of the system they are kept within the simulation model. However the capacity levels are set to match the base speeds of the subsequent belt conveyors, such that the possibility of interference is minimized. Finally, besides the parameters evaluated so far, also conveyor acceleration and deceleration are relevant, particularly for stop-start conveyors. These are kept constant at the standard 0,3 meter / second$^2$ for all conveyors, as is common within simulation models.

### Bag Dimensions

Standard baggage dimensions can be found in Table 19; maximum, average, and minimum bag length, width and height. Baggage that exceeds these dimensions is considered out of gauge; too large in comparison to airport specifications, but it can still be handled (Vanderlande Industries, 2009b). In the simulation model bag dimensions are held constant at length 1000 mm; width 700 mm; height 400 mm. It is common to use these dimensions in a simulation model; in baggage handling systems often fixed windows are used, in which case standard bag dimensions do not affect system performance. This also holds for the Cairo Terminal 3.

| | Length (mm) | Width (mm) | Height (mm) |
|---|---|---|---|
| Maximum | 1000 | 750 | 650 |
| Average | 700 | 500 | 400 |
| Minimum | 300 | 300 | 50 |

**Table 19: Standard baggage dimensions (Vanderlande Industries, 2009e)**

### Baggage Flows

Baggage flow parameters determine how baggage items are directed through the system. These are expected to have a major impact on the system performance; a different route implies significant different cycle time. Original and adapted flow specifications of Cairo Terminal 3 can be observed in Table 20. Reject rates are set for combinations of two levels; baggage is redirected based on the decision of both a screening machine and an operator. The reject rates of level 1 / 2 and level 3 / 4, as well as the probability that manual encoding is required, have been reduced to standard levels. They were set higher than normal for animation purposes.

| Flow | Original model<br>Value | Updated model<br>Value |
|---|---|---|
| L1/2 reject rate | 10 % | 5 % |
| L3/4 reject rate | 15 % | 1 % |
| Manual encoding required | 5 % | 1 % |
| EBS required | 5 % | 5 % |
| EBS lane assignment | 1 :60 %, 2:40 % | 1 :60 %, 2:40 % |
| EBS time between flushes | 15 minutes | 15 minutes |
| EBS flush-back rate (lane 1) | 25 % | 25 % |
| EBS flush-back lane assignement (lane 1) | 1 :100 %, 2:0 % | 1 :100 %, 2:0 % |
| EBS flush-back rate (lane 2) | 100 % | 100 % |
| EBS flush-back lane assignement (lane 2) | 1 :75%, 2:25 % | 1 :75 %, 2:25 % |
| Lateral assignment | Random | Random |

**Table 20: Original and updated model flow specifications**

### Interarrival times

Because of the original purpose of the model no appropriate distribution is used to generate arrivals. Within regular simulation projects, an input distribution or a load file (a data set based upon expected check-in profiles and flight schedules) is specified by the customer. Because this information is not available for Terminal 3, a rough approximation based on system data will be used, as discussed in the assumptions document.

### Service Times

Within baggage handling systems service times generally only play a minor role. Service times can be observed at check-in desks, manual encoding areas, and manual inspection. The service times at check-in desks are out of the scope of the system, while manual encoding and inspection only handle a minor flow (Table 20). Consequently, the effect of an increase in the duration of a service time is expected to be marginal (baggage items are rarely waiting to be handled). As a result, no detailed information about service times is available. Within the model of Terminal 3, a constant service time of 20 seconds has been used for both manual encoding and inspection. Whether the assumption about the small effects of service times can be considered valid, will be examined within the sensitivity analysis.

## *B.1.2 Simulation Configuration*

In this section it will be discussed how the various simulation parameters are configured.

### Responses and Measuring Locations

Normally within a simulation model one, or typically more, output parameters (called responses) are built in to give insight in the model's performance. The most important output parameters for a BHS are the amount of baggage items that arrives too late at a lateral (when it is already closed), the amount of bags that a sorted to a wrong destination and the cycle time of a bag through the system. Since no detailed flight information is available, the first two parameters cannot be taken into account. The cycle time (CT) of a bag will be selected as the main response.

Other important output parameters of a logistic system are throughput (TH) and work in process (WIP) (Hopp & Spearman, 2008). Throughput rates will not only be given of the whole system, but also of

various locations within the system. This enables us to compare the behavior of bags in the system within different settings.

In order to measure these parameters specific registration locations need to be identified. Recall that the first system registration takes place at the first ACRS, while the last registrations take place when a bag enters a lateral. The system identifies a bag as being diverted to the lateral when the bag is not detected by the photo eye on the sorter just after a lateral. However, both the Automatic Code Reader Stations and the photo eyes are not modeled within the simulation model. Based on the AutoCAD drawing, it can be said that an ACRS is placed two meters in front of a level 1 screening machine. Unfortunately, the photo eyes on the sorter cannot be found in the AutoCAD drawing either. It is assumed that they are placed half a meter behind a lateral.

The cycle time of a baggage item, as well as the work in process of the system, are based on the measuring locations of bags entering and leaving the system.

Measuring locations of throughput rates are depicted such that information can be registered of the behavior of every type of flow within the system (Figure 27).

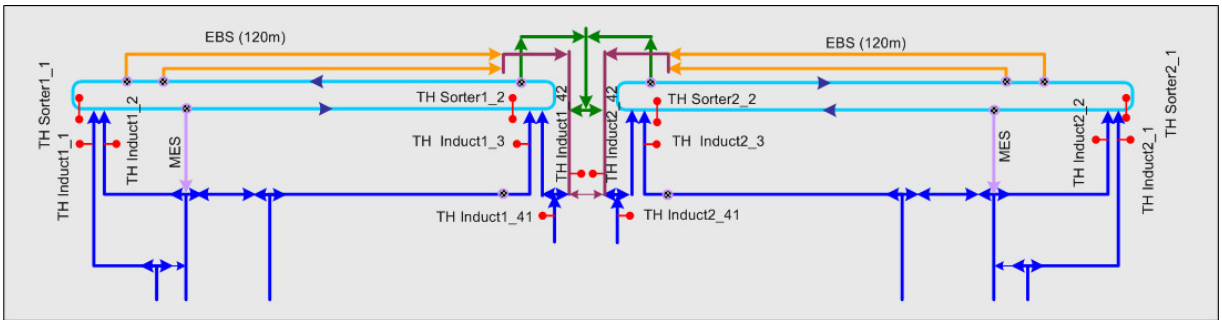



**Figure 27: Measuring locations of throughput rates**

### Warm-up Period, Run Length & Replications

Before the simulation can be conducted, the warm-up period, run length and amount of replications should be determined. For this assessment characteristics of the sensitivity analysis will be taken into account, such that the values can also be used for the design of experiments.

A warm-up analysis is used to estimate how long it takes a system to reach a steady state. If statistics are gathered during the warm-up period, they might not reflect the steady state of the system. The initial conditions of a model and a system, determine whether a model needs to warm up. If the model starts up empty and must reach a loaded state before representing the real system, then the model requires warm-up time (Applied Materials, Inc., 2008). This holds for the model of Cairo Terminal 3; the model starts empty while the real system (almost) never is.

The most general technique for determining the warm-up period is Welch's graphical procedure (Law & Kelton, 2000). The technique is based on plotting a simulation response and eyeing when it reaches a

steady state. Due to the inherent variability of the process, it is difficult to determine the warm-up period on a single replication. As a result the Welch procedure uses the average over more replications. High-frequency oscillations can be smoothened by using a moving average (Law & Kelton, 2000).

The Work In Process (WIP) of the system is recommended as a good response; for instance utilization and throughput almost always reach steady state and tend to reach it more quickly (Applied Materials, Inc., 2008). This behavior was also observed for Terminal 3 while experimenting with several responses. It takes significant additional time to fill all EBS lanes properly. This directly affects the WIP, whereas the effect on other responses is much less noticeable. It should be noted though, that in case a load file is used, a steady state will never be reached due to the nonstationary input process.

In Figure 28 the WIP levels can be observed as a function of time. Three replications are used for both the standard and the alternative distributions, determined in section 5.2.4. This number of replications was thought to be sufficient because the variation among replications was relatively small. Furthermore, WIP levels are measured with an interval of 1 minute. Since oscillations are reasonable small, no moving average has been applied. Based on this figure the warm-up time has been set to 90 minutes; at this time the WIP level appears to have reached a reasonable steady state.

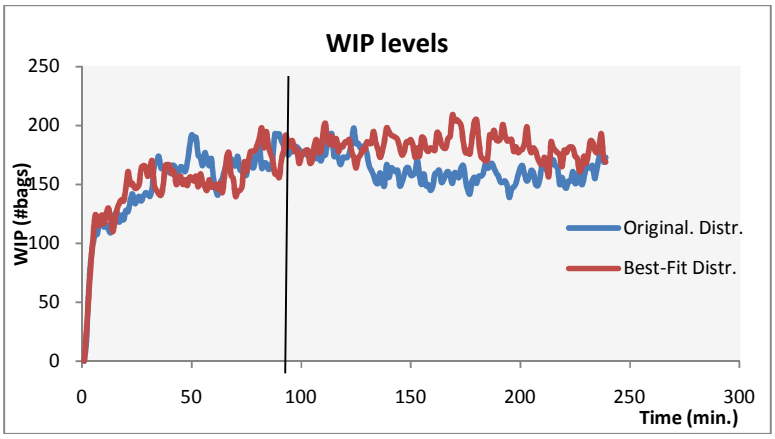



Figure 28: Warm-up period determination based on simulation WIP levels

Concerning the run length and replications, typically a total run time of 24 hours is used for BHS simulations. Due to the application of CRN not too many replications are required in order to get sufficiently small confidence intervals. Therefore the amount of replications has been set to 3. When combining this with a total time of 24 hours, this would imply a run length of 8 hours. However, this might be considered somewhat short, since in the alternative configuration the time-to-flush of an EBS lane is 30 minutes (determined in appendix B.5.2). Though, both lanes are desynchronized, implying that the effect of flushing a lane is experienced once every 15 minutes (per sorter). In order to insure that the effect of the EBS is sufficiently taken into account the run length is set to 12 hours, resulting in a total run time of 36 hours (exclusive warm-up period).

### *B.1.3 Simulation Model Results*

In Table 21 the output results of the standard simulation model can be observed. The responses correspond to the responses identified in the previous section. The cycle time of standard bags is somewhat more than 3 minutes. As expected, the cycle time of bags using the EBS is much larger, while the cycle time of bags to inspection is smaller, since they are removed from the system more upstream. Taking these cycle times into account, the average cycle time of the total amount of bags appears relatively low. This is due to amount of bags requiring EBS is small (5 %). Approximately 168 bags are on average in the system at a given moment in time. The throughput rate of the complete system is approximately 2159 bags per hour. From the throughput rates at different sections within the system it can be observed that they symmetrical left and right side of the system behave very similarly.

Standard deviations are typically very small compared to the average. This is due to the use of common random numbers. As a result of the small standard deviations the 95 % confidence intervals of the average values are relatively small, even though only three replications are used.

| | CT_Standard (sec) | CT_EBS (sec) | CT_Inspection (sec) | CT_Total (sec) | WIP (# bags) | TH_System (bags/h) | TH_Inpection (bags/h) | |
|---|---|---|---|---|---|---|---|---|
| Average | 195,2167 | 1749,90 | 163,663 | 274,007 | 165,007 | 2167,777 | 1,0823 | |
| Std. Dev. | 0,7520 | 72,12 | 3,426 | 6,617 | 4,482 | 7,113 | 0,3643 | |
| CI Low | 193,3487 | 1570,74 | 155,152 | 257,568 | 153,873 | 2150,107 | 0,1773 | |
| CI High | 197,0846 | 1929,05 | 172,174 | 290,445 | 176,141 | 2185,447 | 1,9873 | |
| | TH_Induct1_1 | TH_Induct1_2 | TH_Induct1_3 | TH_Induct1_41 | TH_Induct1_42 | TH_Sorter1_1 | TH_Sorter1_2 | TH_MC1 |
| Average | 341,083 | 340,8067 | 346,223 | 52,2233 | 104,667 | 220,167 | 361,427 | 11,877 |
| Std. Dev. | 2,621 | 0,3443 | 4,516 | 0,7104 | 7,292 | 4,304 | 2,807 | 1,101 |
| CI Low | 334,572 | 339,9514 | 335,005 | 50,4587 | 86,553 | 209,476 | 354,453 | 9,142 |
| CI High | 347,594 | 341,6619 | 357,442 | 53,9880 | 122,781 | 230,857 | 368,400 | 14,611 |
| | TH_Induct2_1 (bags/h) | TH_Induct2_2 (bags/h) | TH_Induct2_3 (bags/h) | TH_Induct2_41 (bags/h) | TH_Induct2_42 (bags/h) | TH_Sorter2_1 (bags/h) | TH_Sorter2_2 (bags/h) | TH_MC2 (bags/h) |
| Average | 340,473 | 338,9433 | 341,250 | 52,803 | 103,817 | 217,443 | 364,850 | 9,8077 |
| Std. Dev. | 5,475 | 0,7564 | 2,349 | 3,043 | 2,505 | 3,055 | 2,877 | 0,6254 |
| CI Low | 326,872 | 337,0643 | 335,414 | 45,244 | 97,593 | 209,854 | 357,702 | 8,2541 |
| CI High | 354,075 | 340,8223 | 347,086 | 60,362 | 110,041 | 225,033 | 371,998 | 11,3612 |

**Table 21: The average, standard deviation and 95 % confidence intervals of simulation responses**

The system bottleneck relates to the component(s) with the highest utilization. For this system configuration the bottlenecks appear to be the sorters. The capacity of the sorter itself is 1500 bags per hour. Though, due to the configuration of the system the sorters are able to handle additional bags. Inducts are placed at both ends of the sorter, while diverts are distributed evenly. This implies that a significant amount of bags inserted at the first two inducts will be diverted before reaching the next two inducts, making space available for additional bags. Combining this information with the setting that bags are directed to a random lateral, results in a theoretical maximum capacity of 1,33 times the basic capacity level (based upon interviews of simulation engineers), which equals a total sorter capacity of 2000 bags per hour. Combining the throughput rates of all sorter inducts, results in a total infeed rate of 1182,18 bag per hour for the left sorter, and 1173,7 bags per hour for the right sorter, which leads to utilization rates of respectively 59,1 % and 58,7 %.

The sorter utilization rates can be verified by calculating the utilization rates at specific locations on the sorters, namely just after two inducts (where the scanning stations are placed). Throughput rates at these points can be calculated by summing what is on the sorter just before the set of inducts and what is added to the sorter at the set of inducts. For the left sorter, the throughput rate at the left ACRS is 338,12 + 341,56 + 218,68 = 898,35 bags per hour. The throughput rate at the right ACRS is 343,14 + 51,75 + 107,61 + 362,69 = 865,19 bags per hour. This results in utilization rates of respectively 59,9 % and 57,7 %. Similarly utilization rates can be calculated of the right sorter. Throughput rates of 892,47 (right) and 859,17 bags per hour (left) relate to utilization rates of respectively 59,5 % and 57,3 %. As such the overall sorter utilization of approximately 59 % seems accurate.

To which degree this simulated system performance matches real system performance is evaluated within the subsequent chapters.

## *B.2 Maintaining an Assumptions Document*

In this section assumptions used while conducting the case study are listed as an example of an assumptions document. The assumptions are specified per section. The purpose of the assumptions document is to present an overview of the assumptions made. More information can be acquired in the respective section.

### *Model Configuration*

- The level of detail contained in the simulation model is sufficiently accurate to make findings derived from its results valid for the real system
- The parameter values accurately describe the real system as specified in appendix B.1.1
- In the simulation model bag dimensions are assumed to be constant at length 1000 mm; width 700 mm; height 400 mm
- No capacity information was available of an ACRS; it is assumed that its capacity is non-restricting
- The stopping space in the EBS is equal to the bag size plus a small gap (200 mm)
- The system is assumed to have an availability of 100%
- The system uses fixed window lengths
- Service times only have a minor effect on the cycle time
- The real system is implemented as drawn in the AutoCAD file
- Back-up lines are not used; no cases traverse both the right and the left subsystem
- The capacity at container loading areas is non-restricting and can be considered endless
- Photo eyes placed on the sorter for registering that bags are diverted, are located half a meter behind a chute

### *Input Validation*

- Interarrival times of cases can be considered continuously distributed, and characterized as a triangle distribution with a minimum of 0, a maximum of 27, and a most likely value of 3
- The effect of input distribution tales is negligible
- The system can handle variation of interarrival times within some minutes
- All 6 scanners have an identical input distribution

- The selected system samples are stationary and IID

### *Trace-Driven Output Validation*

- Nonstationary input data results in nonstationary output differences
- Outliers are not inherent to the system but caused by anomalies such as unavailability
- Bags that are not identified within the tracking system (no LPC), follow a similar pattern as the other baggage items
- An increase in utilization generally leads to an increase in the difference between simulation output and system output

### *Sensitivity Analysis*

- Bags that are not identified within the tracking system (no LPC), follow a similar pattern as the other baggage items
- The left and right EBS are configured identically
- The first EBS flush-backs are representative for the small amount of additional flush-backs
- A linear underlying model can be used to assess factor effects

Factor interactions of the third order, or higher, are negligible

## B.3 Additional Input Validation Results

### B.3.1 System Input Capacity Requirements

In this section graphs can be observed of the system input capacity requirements of day 22, 23, 24 and 25, based on a moving average of 100 bags. The peak capacity requirements of the different days appear to be fairly similar.

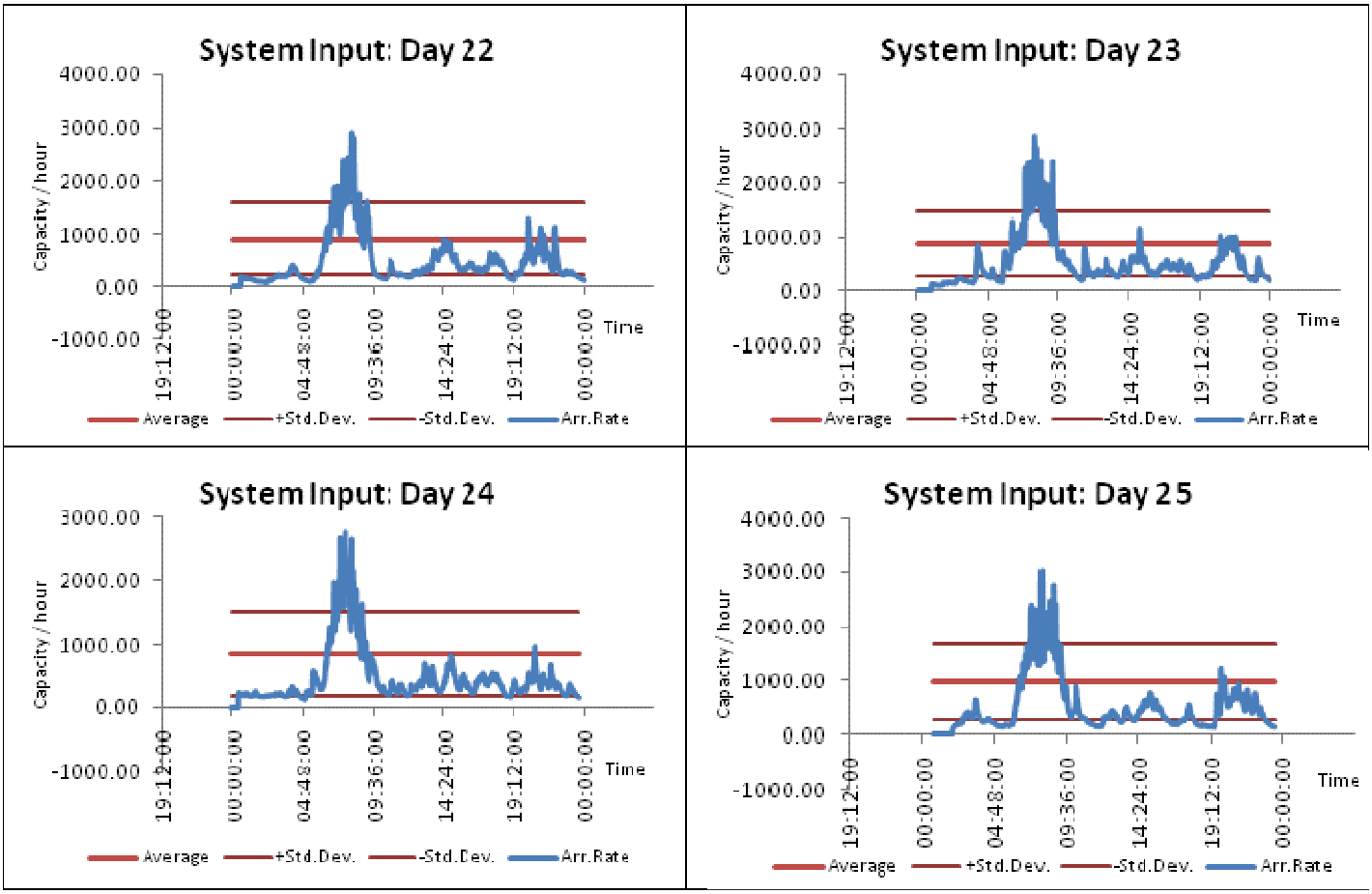



Figure 29: System input capacity requirements, based on a moving average of 100 bags

### *B.3.2 Examples of the effect of a reduced sample size*

In this section additional examples can be found that indicate the difference between the stationarity of samples based on 2250 or 1250 bags.

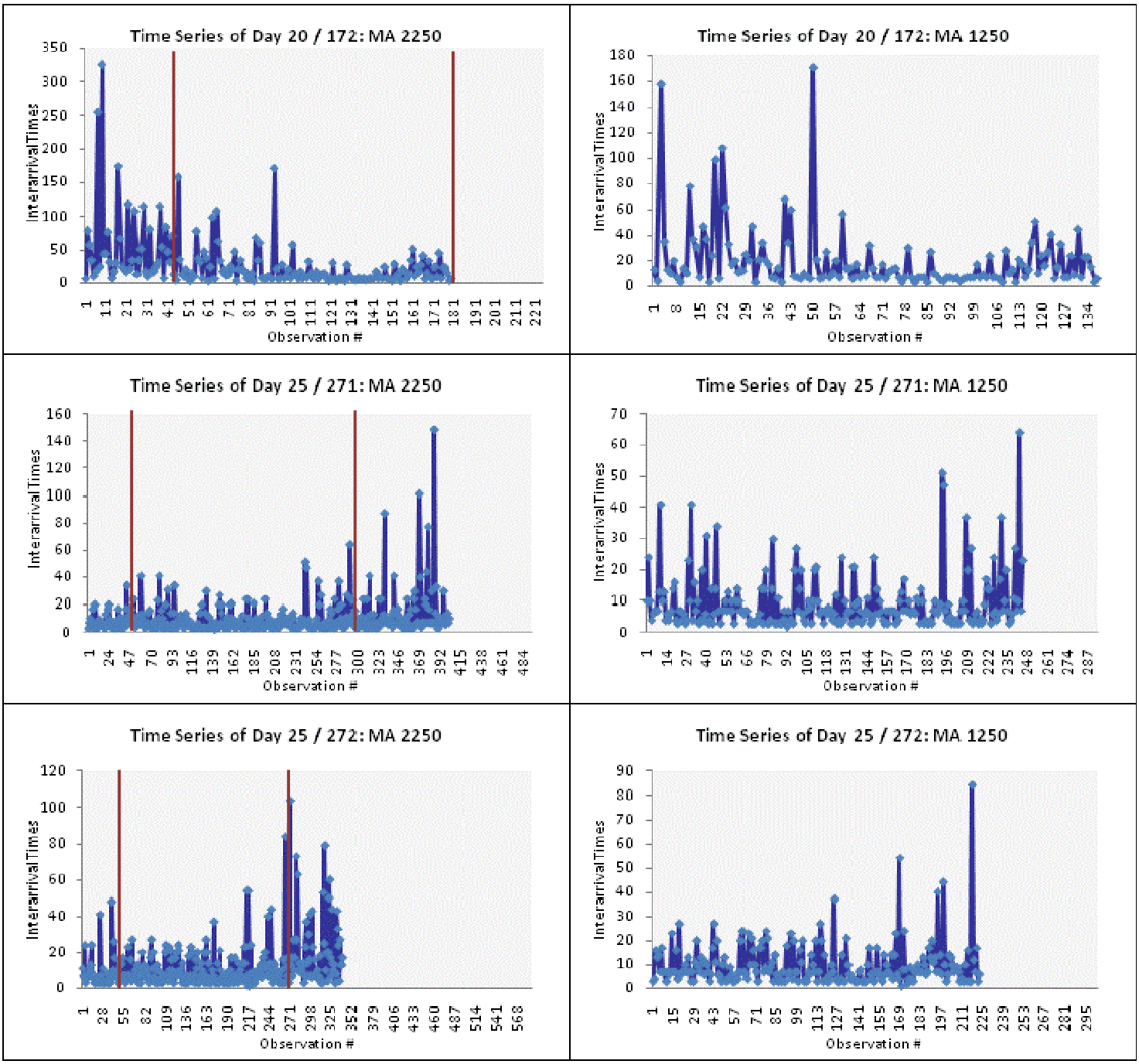



Figure 30: Time series comparison based on a sample size of 2250 and 1250

### *B.3.3 Time Series for System Input Sample Selection*

Within this section run sequence plots can be observed of the input samples per ACRS per day. Note that not all samples appear to be sufficiently stationary. These samples will be excluded within the subsequent steps.

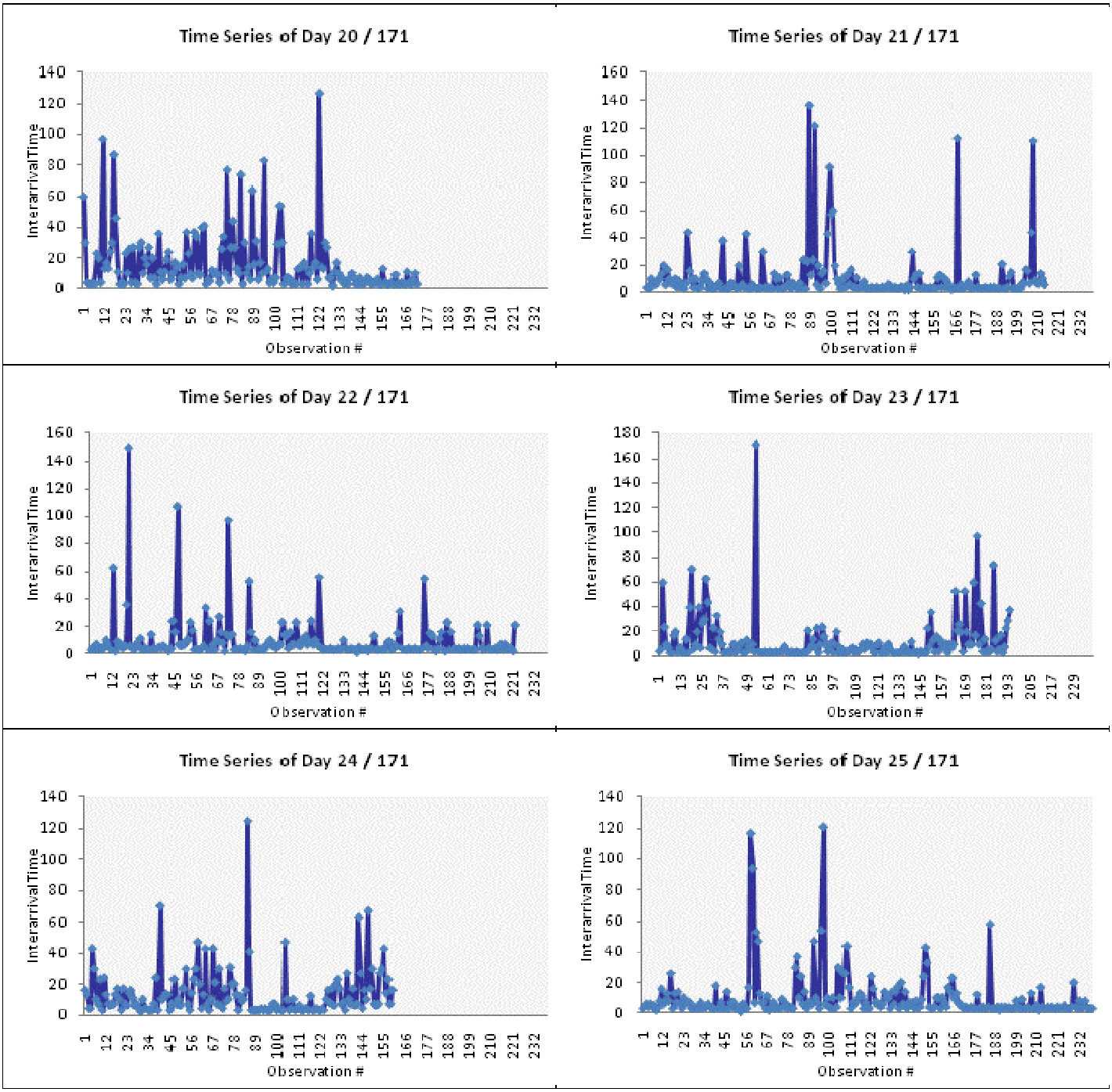



Figure 31: Interarrival times of ACRS 171

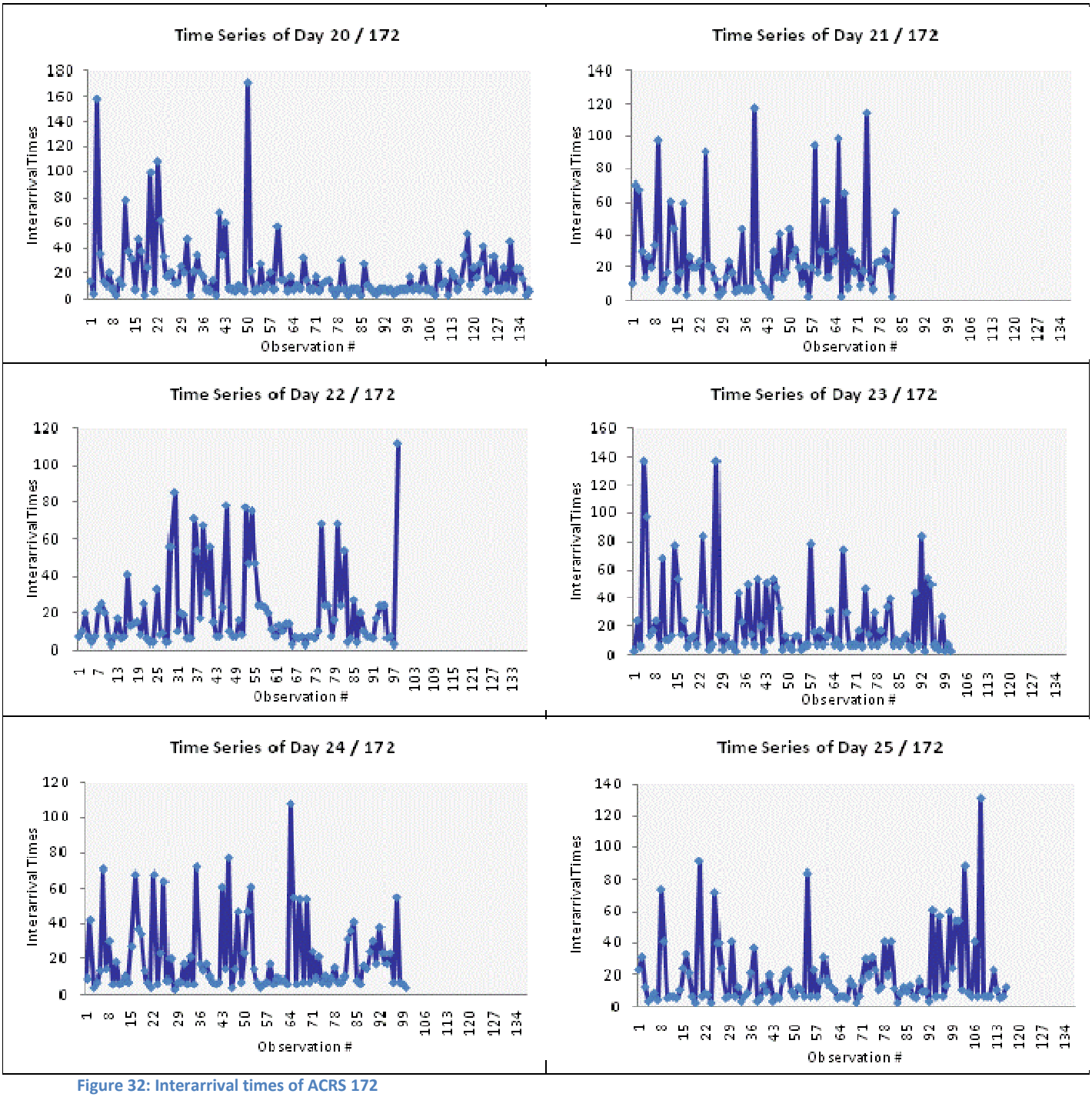



Figure 32: Interarrival times of ACRS 172

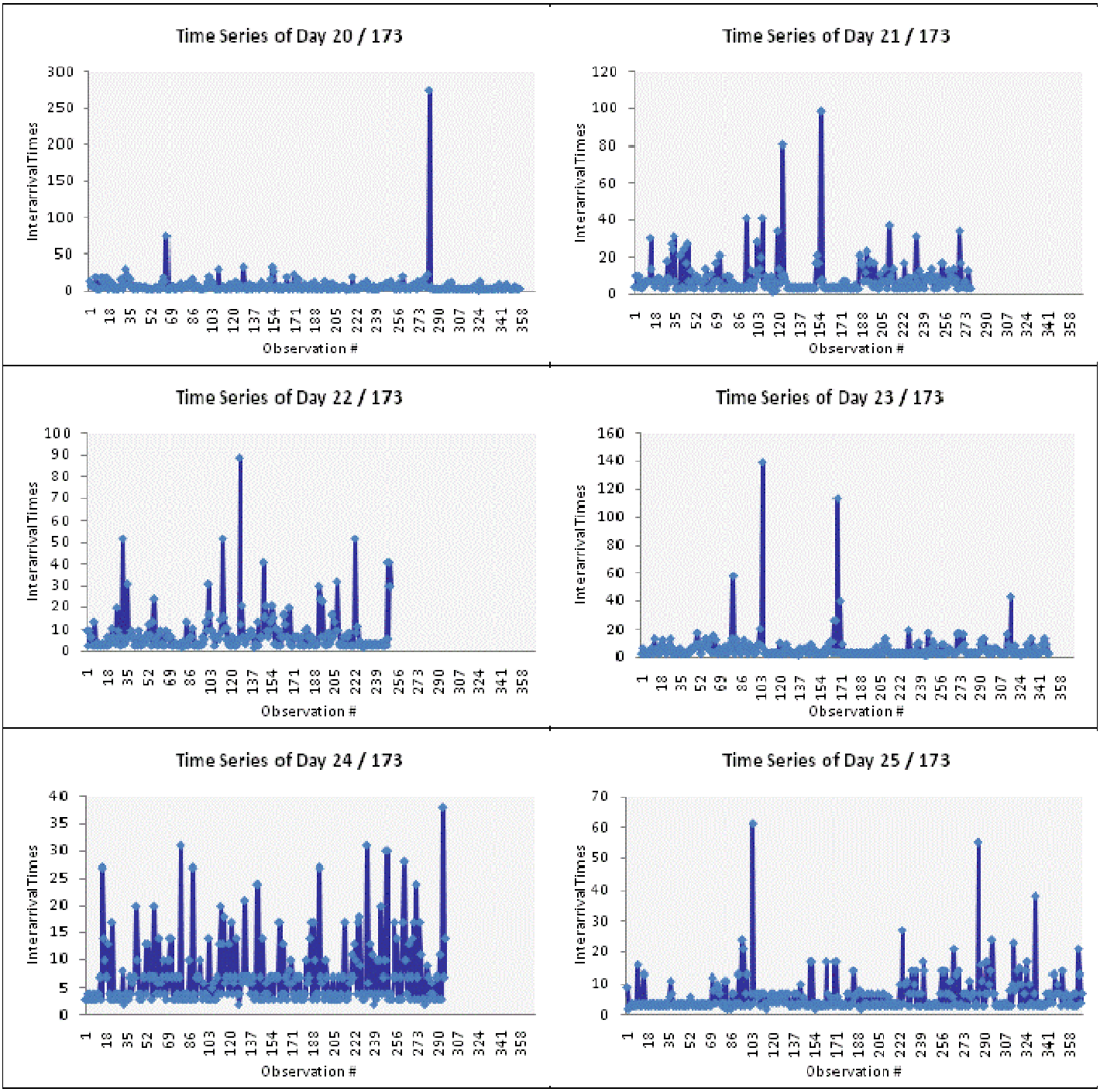



Figure 33: Interarrival times of ACRS 173

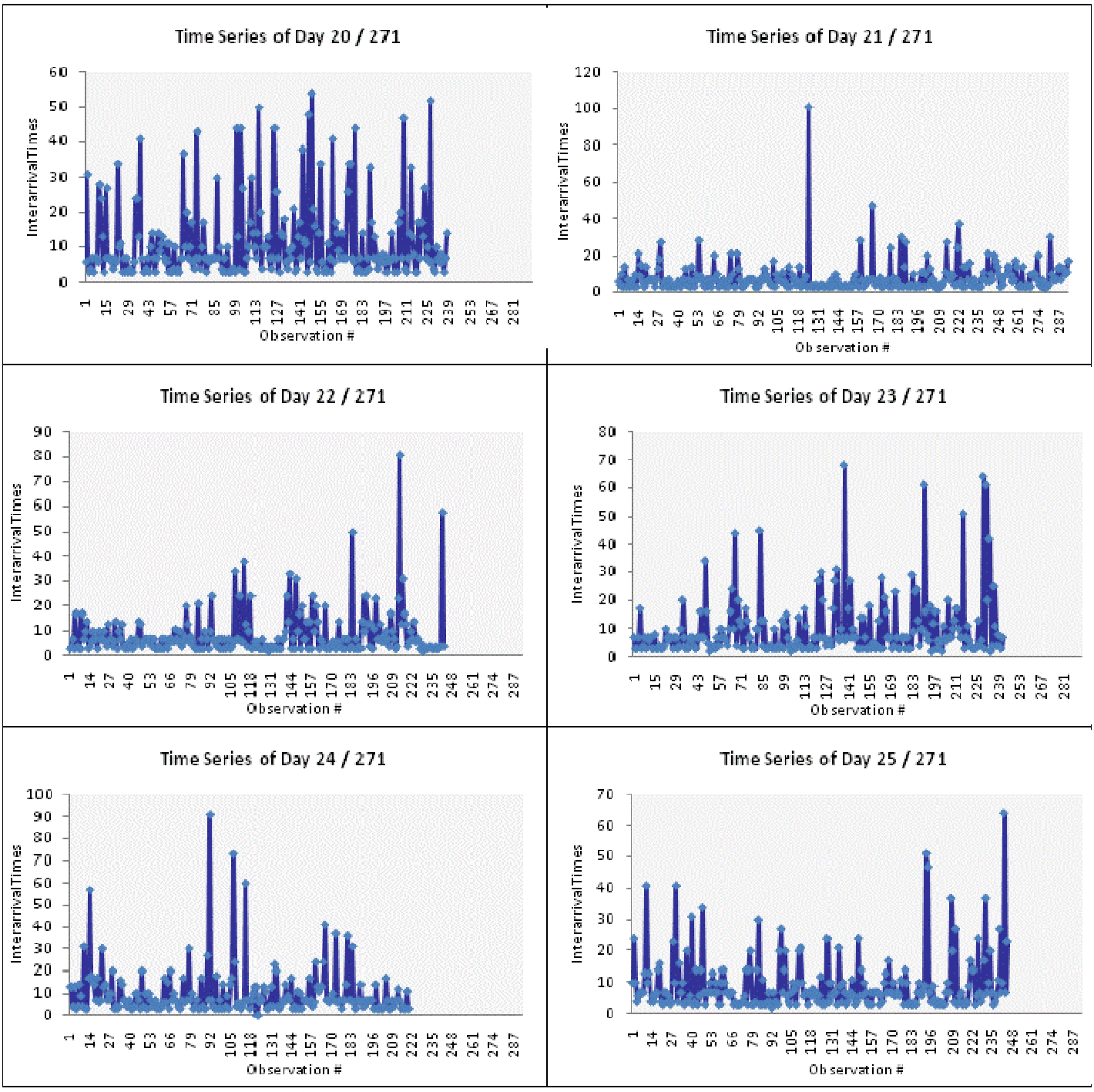



Figure 34: Interarrival times of ACRS 271

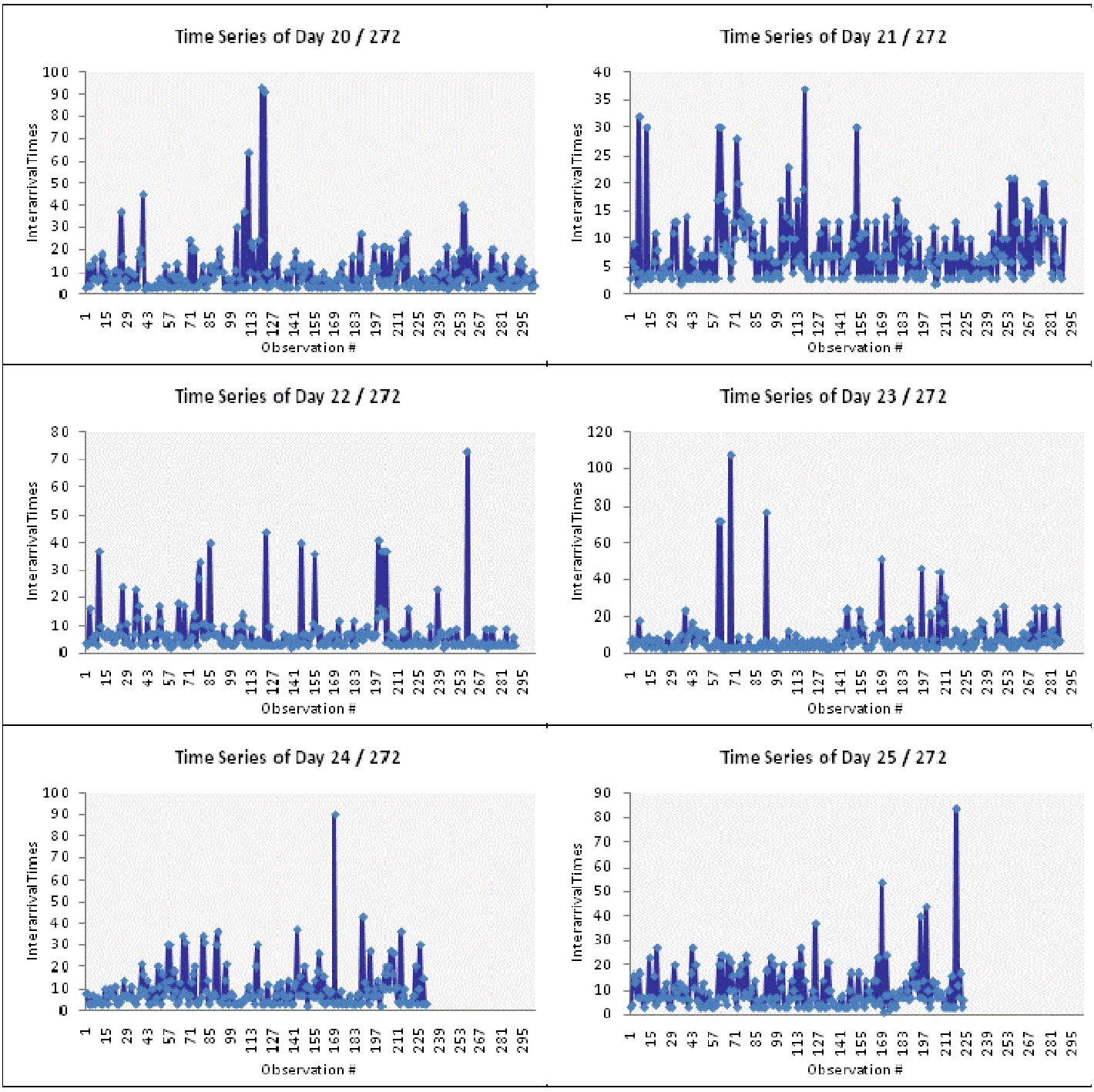



Figure 35: Interarrival times of ACRS 272

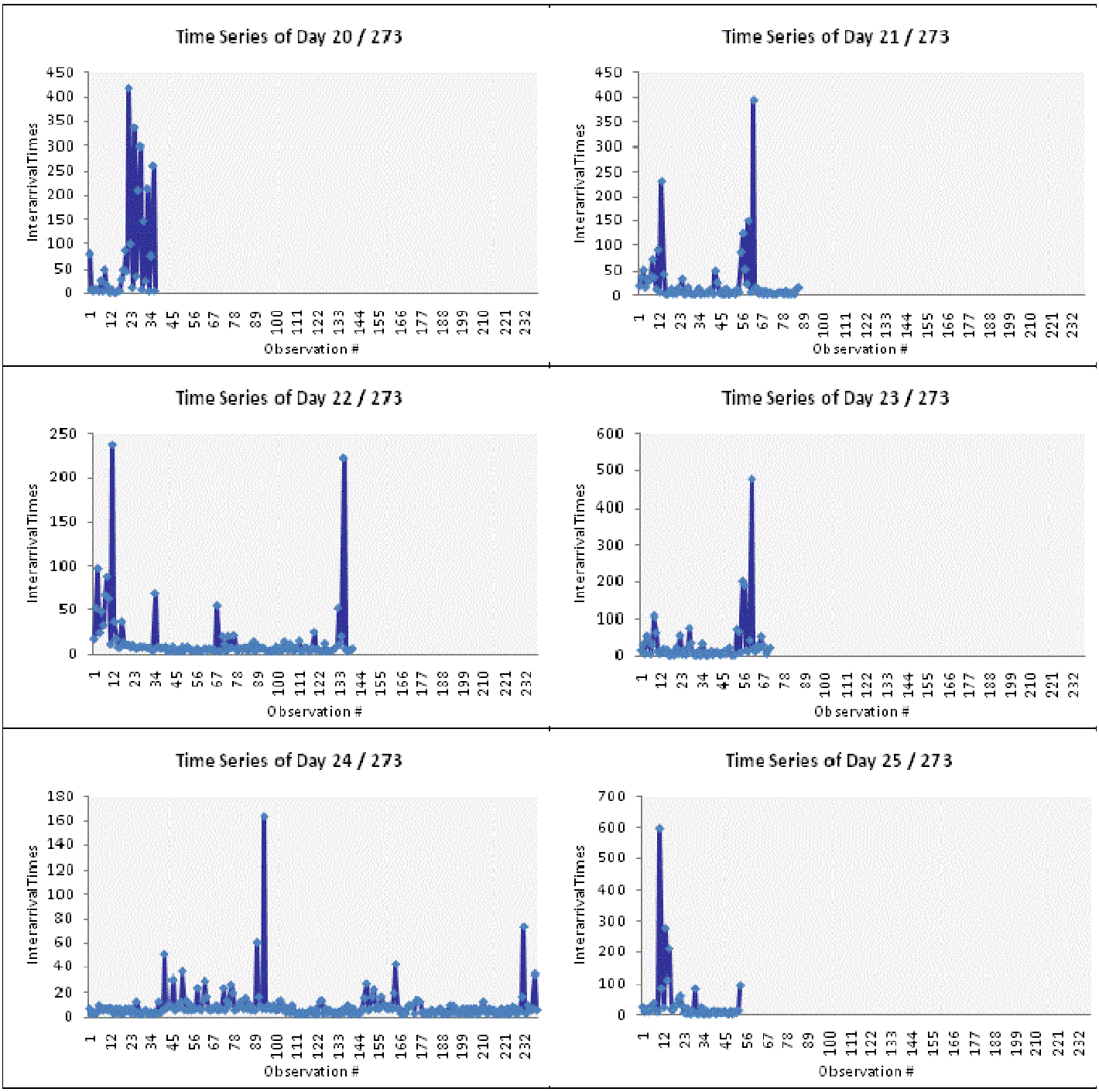



Figure 36: Interarrival times of ACRS 273

### *B.3.4 Data Autocorrelation*

Autocorrelation plots can be observed in Figure 37. Except for ACRS 273, observed values are generally not very high and appear follow no specific pattern. However, the autocorrelation tables indicate that also the autocorrelation of ACRS 273 is generally within the required limits.

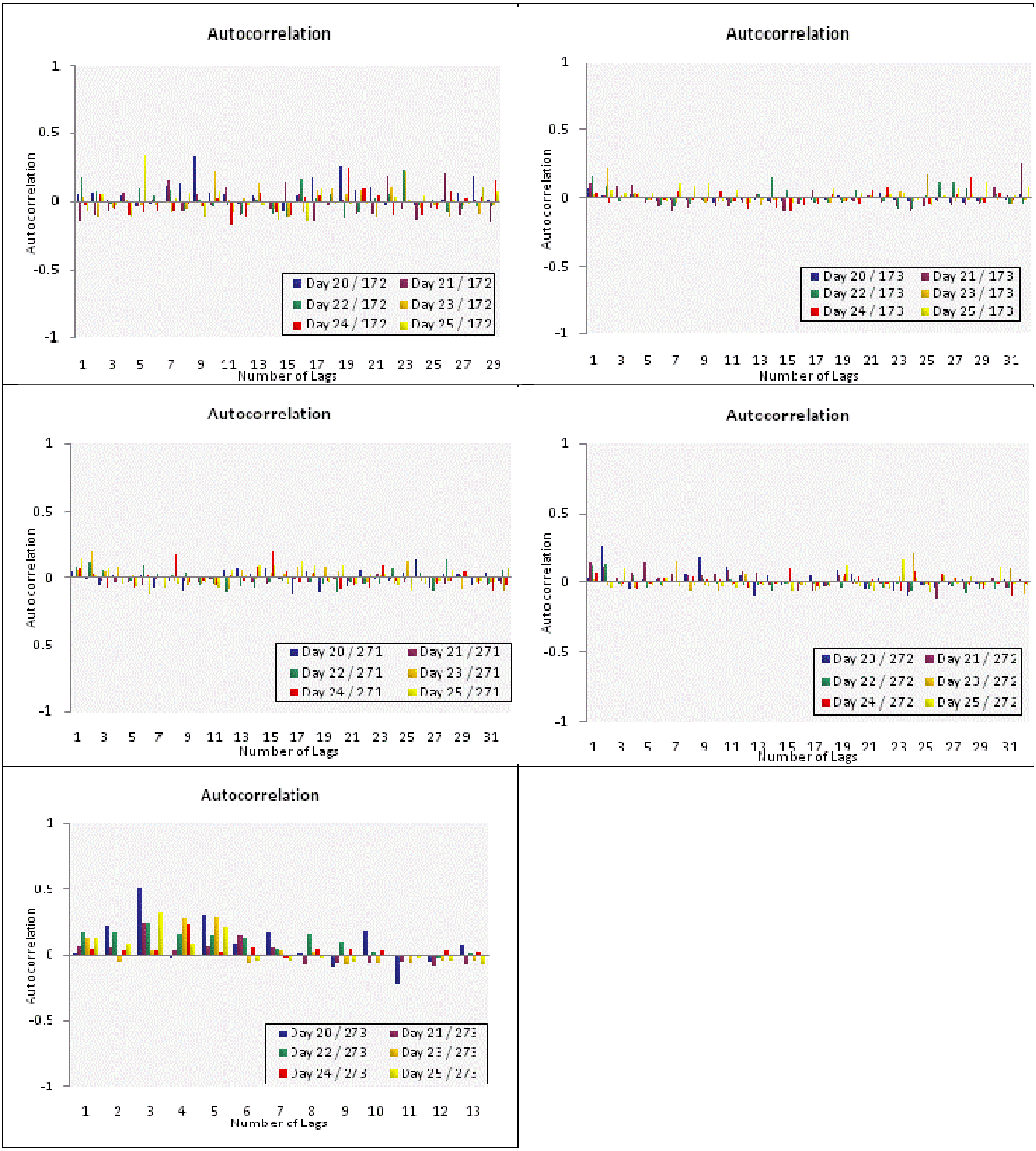



Figure 37: Autocorrelation figures of ACRS 172 / 173 / 271 / 272 / 273

| *Autocorrelation Table* | Day 20<br>171 | Day 21<br>171 | Day 22<br>171 | Day 23<br>171 | Day 24<br>171 | Day 25<br>171 | Day 20<br>172 | Day 21<br>172 | Day 22<br>172 | Day 23<br>172 | Day 24<br>172 | Day 25<br>172 | Day 20<br>173 | Day 21<br>173 | Day 22<br>173 | Day 23<br>173 | Day 24<br>173 | Day 25<br>173 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Number of Values** | 173 | 215 | 223 | 195 | 160 | 239 | 137 | 83 | 98 | 101 | 100 | 116 | 357 | 277 | 250 | 348 | 296 | 369 |
| **Standard Error** | 0,0760 | 0,0682 | 0,0670 | 0,0716 | 0,0791 | 0,0647 | 0,0854 | 0,1098 | 0,1010 | 0,0995 | 0,1000 | 0,0928 | 0,0529 | 0,0601 | 0,0632 | 0,0536 | 0,0581 | 0,0521 |
| **Lag #1** | 0,1118 | **0,2201** | 0,1032 | 0,1022 | **0,1867** | **0,3868** | 0,0610 | -0,1467 | 0,1834 | 0,0377 | -0,0257 | -0,0706 | 0,0734 | 0,1105 | **0,1621** | 0,0241 | 0,0445 | 0,0687 |
| **Lag #2** | 0,0990 | 0,1262 | 0,0001 | **0,1449** | 0,0057 | 0,1073 | 0,0644 | -0,0989 | 0,0803 | -0,1137 | 0,0602 | 0,0569 | 0,0179 | 0,0154 | 0,0969 | **0,2254** | -0,0343 | 0,0611 |
| **Lag #3** | 0,0430 | **0,2532** | -0,0537 | 0,0126 | -0,0238 | 0,1278 | 0,0193 | -0,0704 | -0,0182 | -0,0472 | -0,0563 | -0,0233 | -0,0004 | 0,0899 | -0,0225 | 0,0146 | 0,0101 | 0,0356 |
| **Lag #4** | 0,1435 | 0,0366 | -0,0072 | 0,0659 | 0,0383 | 0,0886 | 0,0512 | 0,0653 | -0,0105 | 0,0049 | -0,1042 | -0,1116 | 0,0254 | 0,1067 | 0,0336 | 0,0375 | 0,0262 | 0,0378 |
| **Lag #5** | -0,0226 | 0,0115 | 0,0008 | 0,0863 | 0,0133 | 0,0634 | -0,0350 | -0,0408 | 0,0984 | -0,0237 | -0,0755 | **0,3465** | 0,0023 | -0,0367 | -0,0172 | -0,0171 | -0,0145 | 0,0439 |
| **Lag #6** | **0,1556** | -0,0179 | -0,0225 | 0,0160 | 0,0836 | -0,0642 | -0,0212 | 0,0009 | 0,0447 | -0,0240 | -0,0691 | -0,0059 | -0,0239 | -0,0643 | -0,0504 | -0,0039 | -0,0125 | -0,0262 |
| **Lag #7** | 0,0815 | 0,0252 | 0,0078 | 0,1027 | 0,0816 | -0,0440 | 0,1230 | 0,1572 | 0,0875 | -0,0752 | -0,0660 | 0,0236 | -0,0167 | -0,0920 | -0,0665 | -0,0297 | 0,0512 | **0,1119** |
| **Lag #8** | 0,0623 | 0,1263 | 0,1201 | 0,0873 | 0,0418 | -0,0138 | 0,1409 | -0,0687 | -0,0647 | -0,0560 | -0,0135 | 0,0695 | -0,0106 | -0,0799 | -0,0480 | -0,0163 | -0,0066 | 0,0936 |
| **Lag #9** | 0,0426 | 0,0887 | -0,0536 | **0,1842** | 0,0390 | -0,0179 | **0,3390** | 0,0586 | 0,0111 | -0,0408 | -0,0328 | -0,1127 | 0,0122 | -0,0157 | -0,0224 | -0,0311 | -0,0217 | **0,1154** |
| **Lag #10** | 0,0938 | **0,1386** | -0,0390 | 0,0416 | 0,0258 | 0,0106 | 0,0654 | -0,0243 | -0,0333 | **0,2284** | 0,0222 | 0,0750 | -0,0303 | -0,0671 | 0,0024 | -0,0201 | 0,0502 | -0,0260 |
| **Lag #11** | -0,0475 | **0,1580** | 0,0068 | 0,0052 | 0,0298 | -0,0019 | 0,0560 | 0,1049 | -0,0246 | 0,0104 | -0,1762 | -0,0760 | -0,0144 | -0,0655 | -0,0295 | -0,0214 | -0,0213 | 0,0565 |
| **Lag #12** | 0,0327 | 0,1100 | 0,0655 | 0,0176 | 0,0236 | 0,0908 | -0,0160 | -0,0975 | -0,0875 | 0,0214 | -0,1073 | -0,0290 | -0,0145 | -0,0376 | 0,0110 | -0,0411 | -0,0819 | -0,0350 |
| **Lag #13** | 0,0414 | 0,0701 | -0,0129 | -0,0173 | 0,0245 | 0,0833 | 0,0467 | 0,0302 | 0,0108 | 0,1388 | 0,0689 | -0,0314 | -0,0035 | 0,0286 | 0,0280 | -0,0594 | 0,0321 | -0,0002 |
| **Lag #14** | 0,0362 | -0,0089 | -0,0318 | -0,0253 | -0,0476 | 0,0484 | -0,0003 | -0,0623 | -0,0936 | -0,0545 | -0,0800 | -0,1274 | -0,0179 | -0,0358 | **0,1570** | -0,0092 | -0,0778 | -0,0148 |
| **Lag #15** | 0,0629 | -0,0236 | -0,0076 | 0,0391 | 0,0284 | 0,0185 | -0,0669 | 0,1549 | -0,1133 | -0,1088 | -0,0967 | -0,0043 | -0,0183 | -0,0961 | 0,0642 | 0,0012 | -0,0970 | -0,0368 |
| **Lag #16** | 0,1128 | -0,0446 | -0,0506 | 0,0736 | 0,0163 | -0,0370 | 0,0466 | 0,0568 | 0,1713 | -0,0766 | 0,0369 | -0,1388 | 0,0072 | -0,0440 | -0,0052 | 0,0003 | -0,0574 | 0,0103 |
| **Lag #17** | 0,0899 | -0,0447 | 0,0032 | -0,0309 | 0,0261 | -0,0235 | **0,1852** | -0,1379 | 0,0258 | 0,0913 | 0,0487 | 0,0953 | -0,0023 | 0,0585 | -0,0363 | -0,0226 | -0,0433 | 0,0102 |
| **Lag #18** | -0,0237 | -0,0580 | -0,0679 | 0,0314 | 0,1055 | -0,0315 | -0,0010 | -0,0253 | 0,0612 | 0,0940 | -0,0082 | -0,0144 | -0,0134 | -0,0055 | -0,0298 | -0,0307 | 0,0290 | 0,0716 |
| **Lag #19** | 0,1247 | -0,0237 | -0,0122 | 0,0189 | **0,2128** | -0,0522 | **0,2636** | 0,0135 | -0,1203 | 0,0577 | **0,2573** | -0,0194 | 0,0150 | -0,0166 | -0,0367 | -0,0239 | -0,0241 | -0,0075 |
| **Lag #20** | 0,1003 | -0,0623 | -0,0376 | 0,0255 | -0,0507 | -0,0179 | 0,0829 | -0,0869 | -0,0749 | 0,0878 | 0,0934 | -0,0213 | -0,0098 | -0,0269 | 0,0572 | -0,0061 | -0,0424 | 0,0439 |
| **Lag #21** | 0,0354 | -0,0718 | 0,0127 | 0,0427 | -0,0846 | -0,0131 | 0,1046 | -0,0862 | 0,0254 | -0,1139 | 0,0426 | -0,0072 | 0,0110 | 0,0140 | -0,0537 | 0,0198 | 0,0598 | 0,0083 |
| **Lag #22** | -0,0411 | -0,0162 | -0,0109 | -0,0237 | 0,0375 | 0,0236 | -0,0264 | 0,1904 | 0,0590 | 0,1068 | -0,0998 | 0,0374 | 0,0358 | -0,0337 | -0,0182 | 0,0202 | 0,0847 | 0,0285 |
| **Lag #23** | 0,0312 | -0,0746 | 0,0373 | 0,0508 | 0,0361 | -0,0092 | -0,0069 | -0,0544 | **0,2344** | **0,2273** | -0,0048 | 0,0015 | -0,0134 | -0,0698 | -0,0856 | 0,0547 | 0,0079 | 0,0390 |
| **Lag #24** | 0,0480 | -0,0229 | 0,0198 | 0,0366 | -0,1334 | 0,0636 | -0,0242 | -0,1282 | -0,0303 | -0,0511 | -0,1023 | 0,0425 | -0,0216 | -0,0925 | -0,0807 | -0,0191 | 0,0012 | -0,0089 |
| **Lag #25** | -0,0488 | -0,0383 | **0,2193** | 0,0308 | -0,0561 | 0,0783 | -0,0059 | -0,0517 | 0,0060 | -0,0284 | -0,0574 | -0,0315 | 0,0057 | -0,0623 | 0,0214 | **0,1726** | -0,0426 | -0,0524 |
| **Lag #26** | -0,0137 | -0,0911 | 0,0320 | 0,0440 | 0,0549 | 0,0436 | 0,0109 | 0,2125 | -0,0827 | -0,1135 | 0,0733 | 0,0124 | -0,0105 | -0,0261 | 0,1236 | 0,0029 | 0,0518 | 0,0214 |
| **Lag #27** | -0,0095 | -0,0735 | 0,1219 | 0,1292 | 0,0524 | 0,0387 | 0,0633 | -0,0986 | -0,0598 | -0,0239 | 0,0278 | -0,0150 | -0,0287 | -0,0518 | 0,1202 | -0,0232 | 0,0249 | 0,0714 |
| **Lag #28** | 0,1141 | -0,0451 | 0,0280 | 0,0222 | -0,0240 | -0,0431 | **0,1890** | 0,0043 | -0,0362 | -0,0911 | 0,0369 | 0,1134 | -0,0285 | -0,0516 | 0,0729 | -0,0101 | **0,1556** | 0,0310 |
| **Lag #29** | 0,0071 | -0,0629 | 0,0283 | 0,0767 | -0,0651 | -0,0403 | 0,0107 | -0,1472 | -0,0350 | -0,0224 | 0,1579 | 0,0768 | -0,0226 | -0,0454 | -0,0350 | -0,0051 | -0,0330 | **0,1257** |
| **Lag #30** | 0,0396 | -0,0567 | -0,0254 | -0,0164 | -0,1341 | -0,0197 | | | | | | | 0,0002 | 0,0858 | 0,0334 | -0,0135 | 0,0426 | 0,0122 |
| **Lag #31** | -0,0281 | -0,0515 | -0,0118 | 0,0518 | -0,1312 | -0,0441 | | | | | | | -0,0006 | 0,0189 | -0,0452 | -0,0492 | -0,0046 | 0,0142 |
| **Lag #32** | 0,0332 | -0,0946 | 0,0185 | -0,0083 | -0,0187 | -0,0541 | | | | | | | 0,0288 | **0,2626** | -0,0447 | -0,0081 | 0,0084 | 0,0829 |

**Table 22: Autocorrelation table of ACRS 171 / 172 / 173 (significant correlations are bold)**

| *Autocorrelation Table* | Day 20 271 | Day 21 271 | Day 22 271 | Day 23 271 | Day 24 271 | Day 25 271 | Day 20 272 | Day 21 272 | Day 22 272 | Day 23 272 | Day 24 272 | Day 25 272 | Day 20 273 | Day 21 273 | Day 22 273 | Day 23 273 | Day 24 273 | Day 25 273 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Number of Values** | 238 | 293 | 243 | 242 | 220 | 245 | 303 | 290 | 290 | 288 | 230 | 223 | 36 | 86 | 140 | 70 | 238 | 52 |
| **Standard Error** | 0,0648 | 0,0584 | 0,0642 | 0,0643 | 0,0674 | 0,0639 | 0,0574 | 0,0587 | 0,0587 | 0,0589 | 0,0659 | 0,0670 | 0,1667 | 0,1078 | 0,0845 | 0,1195 | 0,0648 | 0,1387 |
| **Lag #1** | 0,0525 | 0,0101 | 0,0833 | 0,0598 | 0,0762 | **0,1421** | 0,0319 | **0,1446** | **0,1180** | 0,0240 | 0,0720 | -0,0271 | 0,0193 | 0,0626 | 0,1659 | 0,1301 | 0,0503 | 0,1275 |
| **Lag #2** | -0,0144 | -0,0059 | 0,1149 | **0,1976** | 0,0337 | 0,0157 | **0,2681** | 0,1107 | **0,1321** | -0,0181 | 0,0089 | -0,0453 | 0,2205 | 0,0565 | **0,1696** | -0,0495 | 0,0367 | 0,0933 |
| **Lag #3** | -0,0479 | -0,0212 | 0,0627 | 0,0478 | -0,0693 | 0,0697 | 0,0848 | 0,0283 | -0,0022 | -0,0288 | -0,0077 | 0,1028 | **0,5072** | **0,2576** | **0,2576** | 0,0342 | 0,0361 | **0,3249** |
| **Lag #4** | 0,0254 | -0,0357 | 0,0684 | 0,0845 | 0,0149 | -0,0454 | -0,0513 | 0,0749 | 0,0461 | -0,0380 | -0,0558 | 0,0085 | -0,0141 | 0,0376 | 0,1593 | **0,2784** | **0,2474** | 0,0933 |
| **Lag #5** | 0,0033 | -0,0301 | -0,0239 | 0,0238 | -0,0711 | -0,0619 | 0,0155 | **0,1406** | -0,0448 | -0,0122 | -0,0053 | -0,0187 | 0,3058 | 0,0639 | 0,1476 | **0,2946** | 0,0255 | 0,2162 |
| **Lag #6** | 0,0255 | -0,0471 | 0,0903 | 0,0075 | 0,0186 | -0,1237 | 0,0253 | 0,0290 | -0,0232 | -0,0315 | 0,0304 | 0,0262 | 0,0908 | 0,1538 | 0,1283 | -0,0538 | 0,0525 | -0,0369 |
| **Lag #7** | -0,0768 | -0,0084 | 0,0300 | 0,0127 | 0,0120 | -0,0733 | 0,0006 | 0,0563 | 0,0037 | **0,1513** | 0,0053 | -0,0307 | 0,1682 | 0,0597 | 0,0456 | 0,0414 | -0,0018 | -0,0375 |
| **Lag #8** | -0,0261 | 0,0108 | 0,0240 | -0,0179 | **0,1794** | -0,0366 | 0,0573 | 0,0467 | 0,0009 | -0,0582 | 0,0399 | -0,0199 | 0,0166 | -0,0613 | 0,1610 | 0,0274 | 0,0420 | -0,0177 |
| **Lag #9** | -0,0968 | -0,0159 | 0,0425 | -0,0537 | -0,0362 | 0,0026 | **0,1858** | 0,0507 | 0,0180 | -0,0172 | 0,0174 | -0,0269 | -0,0977 | -0,0523 | 0,0980 | -0,0684 | 0,0428 | -0,0409 |
| **Lag #10** | 0,0063 | -0,0272 | -0,0501 | -0,0420 | -0,0171 | -0,0298 | 0,0097 | 0,0619 | -0,0115 | -0,0622 | 0,0217 | -0,0270 | 0,1821 | -0,0605 | 0,0279 | -0,0540 | 0,0319 | 0,0018 |
| **Lag #11** | -0,0012 | -0,0102 | -0,0427 | -0,0554 | -0,0513 | -0,0692 | 0,1085 | 0,0906 | 0,0220 | -0,0169 | -0,0009 | -0,0370 | -0,2200 | -0,0447 | 0,0039 | -0,0596 | 0,0030 | -0,0036 |
| **Lag #12** | 0,0657 | -0,0375 | -0,1033 | -0,0880 | 0,0209 | 0,0593 | 0,0499 | 0,0811 | -0,0179 | 0,0577 | -0,0412 | -0,0416 | -0,0442 | -0,0835 | -0,0084 | -0,0301 | 0,0354 | -0,0381 |
| **Lag #13** | 0,0714 | 0,0098 | -0,0585 | 0,0630 | -0,0261 | 0,0210 | -0,0982 | 0,0682 | -0,0196 | -0,0044 | -0,0033 | -0,0256 | 0,0754 | -0,0610 | 0,0125 | -0,0301 | 0,0253 | -0,0639 |
| **Lag #14** | 0,0355 | -0,0327 | -0,0731 | -0,0152 | 0,0808 | 0,0901 | 0,0462 | -0,0178 | -0,0611 | 0,0019 | -0,0135 | -0,0109 | | | | | | |
| **Lag #15** | 0,0725 | -0,0458 | -0,0357 | 0,0444 | **0,1942** | 0,0973 | -0,0170 | 0,0069 | -0,0152 | -0,0063 | 0,1051 | -0,0595 | | | | | | |
| **Lag #16** | -0,0054 | -0,0098 | -0,0260 | 0,0325 | 0,0495 | -0,0416 | -0,0494 | -0,0571 | -0,0074 | -0,0175 | 0,0090 | -0,0167 | | | | | | |
| **Lag #17** | -0,1249 | -0,0001 | 0,0005 | 0,0840 | -0,0262 | 0,1230 | 0,0535 | -0,0566 | -0,0002 | -0,0666 | -0,0323 | -0,0468 | | | | | | |
| **Lag #18** | 0,0484 | -0,0342 | -0,0352 | 0,0310 | 0,0400 | 0,0974 | -0,0296 | -0,0283 | -0,0208 | -0,0332 | 0,0046 | 0,0059 | | | | | | |
| **Lag #19** | -0,1064 | 0,0051 | -0,0402 | 0,0875 | -0,0177 | -0,0356 | 0,0874 | 0,0462 | -0,0391 | 0,0420 | 0,0562 | 0,1208 | | | | | | |
| **Lag #20** | -0,0055 | -0,0146 | -0,1079 | 0,0554 | -0,0859 | 0,0900 | 0,0602 | -0,0022 | 0,0209 | -0,0041 | 0,0397 | -0,0301 | | | | | | |
| **Lag #21** | -0,0624 | -0,0181 | -0,0281 | 0,0027 | -0,0521 | -0,0397 | -0,0519 | -0,0130 | -0,0512 | -0,0315 | 0,0213 | -0,0622 | | | | | | |
| **Lag #22** | 0,0640 | -0,0366 | -0,0455 | -0,0354 | -0,0733 | 0,0298 | 0,0349 | -0,0039 | -0,0446 | -0,0370 | -0,0056 | -0,0469 | | | | | | |
| **Lag #23** | -0,0053 | 0,0338 | -0,0376 | 0,0187 | 0,0880 | 0,0315 | -0,0658 | 0,0114 | -0,0003 | 0,0458 | -0,0583 | **0,1637** | | | | | | |
| **Lag #24** | -0,0207 | -0,0069 | 0,0691 | -0,0390 | -0,0179 | -0,0535 | -0,1016 | -0,0719 | -0,0626 | **0,2141** | 0,0848 | 0,0293 | | | | | | |
| **Lag #25** | 0,0402 | -0,0329 | -0,0119 | 0,1237 | 0,0069 | -0,0891 | -0,0247 | -0,0192 | -0,0171 | -0,0119 | -0,0188 | -0,0748 | | | | | | |
| **Lag #26** | **0,1336** | 0,0069 | 0,0426 | -0,0203 | 0,0012 | -0,0417 | -0,0401 | **-0,1249** | -0,0103 | 0,0050 | 0,0584 | 0,0523 | | | | | | |
| **Lag #27** | -0,0747 | 0,0146 | -0,0897 | -0,0345 | -0,0431 | -0,0196 | -0,0222 | -0,0106 | -0,0261 | 0,0349 | 0,0259 | -0,0075 | | | | | | |
| **Lag #28** | 0,0316 | -0,0439 | **0,1380** | -0,0244 | -0,0222 | 0,0606 | 0,0213 | -0,0484 | -0,0818 | -0,0096 | -0,0156 | 0,0409 | | | | | | |
| **Lag #29** | 0,0294 | 0,0351 | 0,0186 | -0,0794 | 0,0466 | 0,0174 | -0,0076 | -0,0011 | -0,0556 | -0,0105 | -0,0541 | -0,0212 | | | | | | |
| **Lag #30** | -0,0547 | 0,0151 | **0,1467** | -0,0352 | -0,0460 | -0,0235 | 0,0095 | 0,0348 | -0,0469 | -0,0038 | -0,0016 | 0,1133 | | | | | | |
| **Lag #31** | 0,0385 | -0,0487 | -0,0429 | -0,0272 | -0,0920 | -0,0340 | 0,0233 | -0,0380 | -0,0367 | 0,0989 | -0,1022 | -0,0036 | | | | | | |
| **Lag #32** | -0,0251 | -0,0451 | 0,0602 | -0,0905 | -0,0502 | 0,0761 | 0,0040 | 0,0241 | 0,0046 | -0,0878 | -0,0148 | -0,0206 | | | | | | |

**Table 23: Autocorrelation table of ACRS 271 / 272 / 273 (significant correlations are bold)**

### *B.3.5 ACRS Sample Selection*

Within this section it will be evaluated which samples can be combined, of automated code scanners other than ACRS 171. For this purpose summary statistics and Box-Whisker plots can be observed, as well as the results of hypothesis tests based on the median and the standard deviation.

| | Day 20 | Day 21 | Day 22 | Day 23 | Day 24 | Day 25 |
|---|---|---|---|---|---|---|
| ***One Variable Summary*** | **172** | **172** | **172** | **172** | **172** | **172** |
| **Mean** | 19,64 | 27,71 | 21,02 | 23,50 | 21,05 | 19,56 |
| **Variance** | 631,20 | 705,96 | 496,76 | 745,95 | 442,31 | 463,31 |
| **Std. Dev.** | 25,12 | 26,57 | 22,29 | 27,31 | 21,03 | 21,52 |
| **Median** | 10,00 | 20,00 | 11,00 | 13,00 | 13,00 | 11,00 |
| **Mode** | 7,00 | 7,00 | 7,00 | 7,00 | 7,00 | 7,00 |
| **Minimum** | 3,00 | 3,00 | 3,00 | 3,00 | 3,00 | 3,00 |
| **Maximum** | 171,00 | 118,00 | 112,00 | 138,00 | 108,00 | 131,00 |
| **Count** | 137 | 83 | 98 | 101 | 100 | 116 |
| | **173** | **173** | **173** | **173** | **173** | **173** |
| **Mean** | 7,68 | 8,444 | 8,252 | 6,66 | 7,311 | 6,160 |
| **Variance** | 241,98 | 96,741 | 89,362 | 112,47 | 34,967 | 34,569 |
| **Std. Dev.** | 15,56 | 9,836 | 9,453 | 10,61 | 5,913 | 5,880 |
| **Median** | 4,00 | 6,000 | 6,000 | 4,00 | 6,000 | 4,000 |
| **Mode** | 3,00 | 3,000 | 3,000 | 3,00 | 3,000 | 3,000 |
| **Minimum** | 1,00 | 1,000 | 2,000 | 2,00 | 2,000 | 2,000 |
| **Maximum** | 275,00 | 99,000 | 89,000 | 139,00 | 38,000 | 61,000 |
| **Count** | 357 | 277 | 250 | 348 | 296 | 369 |
| | **271** | **271** | **271** | **271** | **271** | **271** |
| **Mean** | 11,50 | 7,973 | 8,432 | 9,69 | 9,79 | 9,269 |
| **Variance** | 121,31 | 68,232 | 78,337 | 112,12 | 117,79 | 75,911 |
| **Std. Dev.** | 11,01 | 8,260 | 8,851 | 10,59 | 10,85 | 8,713 |
| **Median** | 7,00 | 6,000 | 6,000 | 7,00 | 7,00 | 7,000 |
| **Mode** | 7,00 | 3,000 | 3,000 | 3,00 | 7,00 | 7,000 |
| **Minimum** | 3,00 | 2,000 | 2,000 | 2,00 | 0,00 | 2,000 |
| **Maximum** | 54,00 | 101,00 | 81,000 | 68,00 | 91,00 | 64,000 |
| **Count** | 238 | 293 | 243 | 242 | 220 | 245 |
| | **272** | **272** | **272** | **272** | **272** | **272** |
| **Mean** | 9,08 | 7,714 | 7,117 | 8,23 | 9,409 | 9,843 |
| **Variance** | 102,10 | 29,790 | 61,502 | 107,50 | 88,260 | 82,295 |
| **Std. Dev.** | 10,10 | 5,458 | 7,842 | 10,37 | 9,395 | 9,072 |
| **Median** | 6,00 | 6,000 | 4,000 | 6,00 | 7,000 | 7,000 |
| **Mode** | 3,00 | 3,000 | 3,000 | 3,00 | 3,000 | 3,000 |
| **Minimum** | 2,00 | 2,000 | 2,000 | 2,00 | 2,000 | 1,000 |
| **Maximum** | 93,00 | 37,000 | 73,000 | 108,00 | 90,000 | 84,000 |
| **Count** | 303 | 290 | 290 | 288 | 230 | 223 |
| | **273** | **273** | **273** | **273** | **273** | **273** |
| **Mean** | 74,08 | 24,56 | 14,36 | 32,36 | 9,03 | 40,13 |
| **Variance** | 11701,3 | 2829,2 | 934,33 | 4246,15 | 176,14 | 8697,4 |
| **Std. Dev.** | 108,17 | 53,19 | 30,57 | 65,16 | 13,27 | 93,26 |
| **Median** | 20,00 | 7,00 | 6,00 | 13,00 | 7,00 | 11,00 |
| **Mode** | 10,00 | 3,00 | 3,00 | 3,00 | 7,00 | 3,00 |
| **Minimum** | 3,00 | 3,00 | 3,00 | 2,00 | 3,00 | 3,00 |
| **Maximum** | 420,00 | 394,00 | 238,00 | 479,00 | 164,00 | 596,00 |
| **Count** | 36 | 86 | 140 | 70 | 238 | 52 |

Table 24: Summary statistics of ACRS 172 / 173 / 271 / 272 / 273

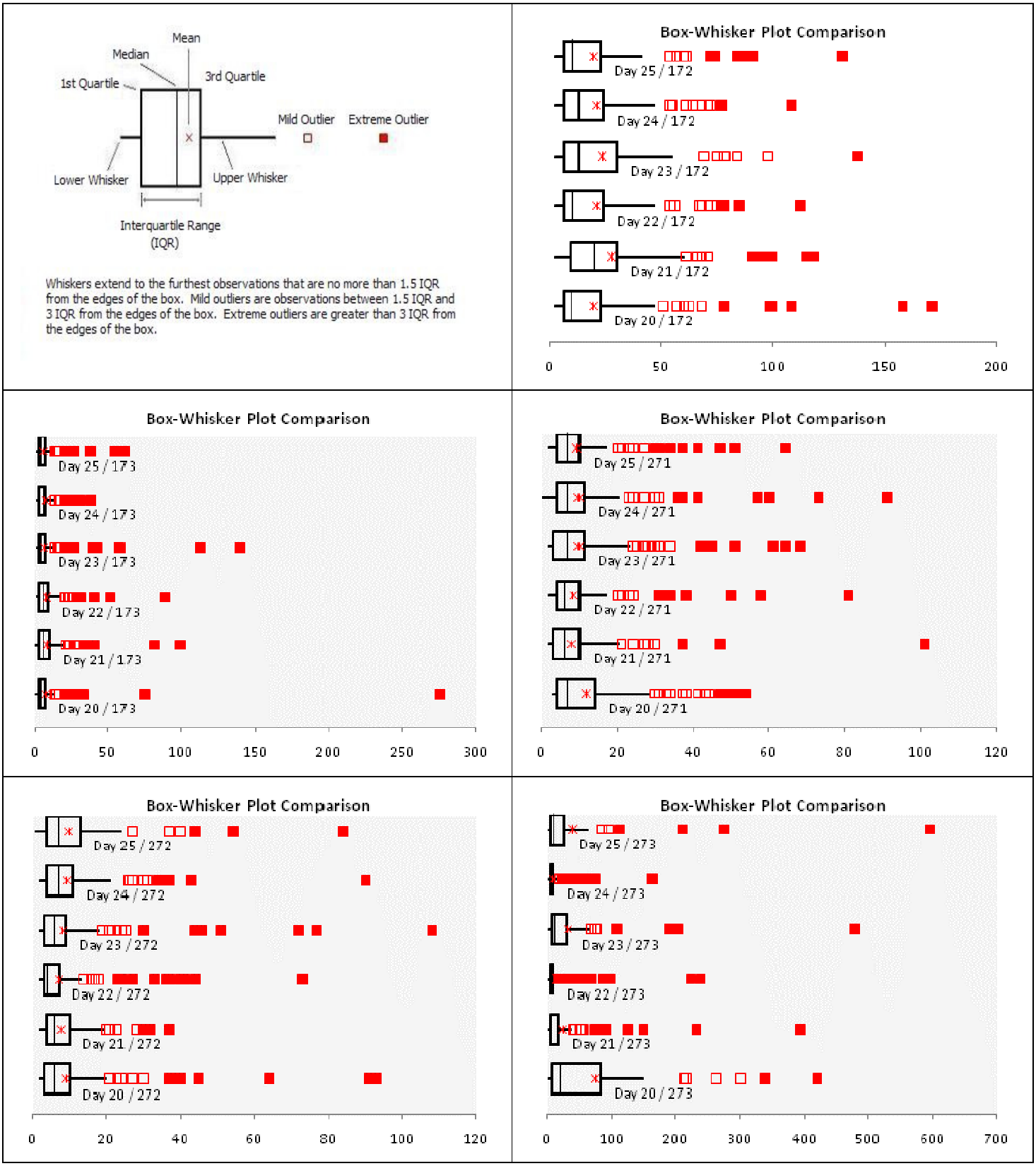



Figure 38: Box-Whisker Plots of ACRS 172 / 173 / 271 / 272 / 273

| Included Samples | Kruskal Wallis test | | Mood's Median test | | |
|---|---|---|---|---|---|
| ACRS 172 | Test Statistic | P-Value | Test Statistic | P-Value | Comments |
| Day 20 / 21 / 22 / 23 / 24 / 25 | 12,6602 | 0,0268 | 17,2682 | 0,004 | Day 21 has the most offset median and mean |
| *Day 20 / 22 / 23 / 24 / 25* | *0,8308* | *0,9343* | *0,6813* | *0,9536* | *Accept null hypothesis* |
| ACRS 173 | Test Statistic | P-Value | Test Statistic | P-Value | Comments |
| Day 20 / 21 / 22 / 23 / 24 / 25 | 32,5402 | 4,6434E-6 | 35,6387 | 1,1216E-6 | Day 25 has the most offset median, mean and variance |
| Day 20 / 21 / 22 / 23 / 24 | 21,225 | 2,857E-4 | 23,5718 | 9,731E-5 | Day 23 has the most offset median and mean |
| *Day 20 / 21 / 22 / 24* | *3,1202* | *0,3735* | *1,9184* | *0,5895* | *Accept null hypothesis* |
| *Day 23 / 25* | *0,0215* | *0,8833* | *0,0534* | *0,8172* | *Alternative combination - Accept null hypothesis* |
| ACRS 271 | Test Statistic | P-Value | Test Statistic | P-Value | Comments |
| Day 20 / 21 / 22 / 23 / 24 / 25 | 22,4118 | 4,371E-4 | 12,776 | 0,0255 | Day 20 has the most offset median and mean |
| Day 21 / 22 / 23 / 24 / 25 | 8,314 | 0,0807 | 4,1981 | 0,3798 | Day 21 has the most offset median and mean |
| *Day 22 / 23 / 24 / 25* | *4,4813* | *0,21396* | *3,1627* | *0,3672* | *Accept null hypothesis* |
| *Day 21 / 22* | *0,1088* | *0,7415* | *0,0608* | *0,8053* | *Alternative combination – Accept null hypothesis* |
| ACRS 272 | Test Statistic | P-Value | Test Statistic | P-Value | Comments |
| Day 20 / 21 / 22 / 23 / 24 / 25 | 38,6127 | 2,8416E-7 | 34,6403 | 1,7749E-6 | Day 22 has the most offset median and mean |
| Day 20 / 21 / 23 / 24 / 25 | 15,0214 | 0,0047 | 12,0266 | 0,0172 | Day 21 has the most offset median and mean |
| Day 20 / 23 / 24 / 25 | 14,0877 | 0,0027 | 11,7661 | 0,0082 | Day 23 has the most offset median and mean |
| *Day 20 / 24 / 25* | *4,2071* | *0,1220* | *1,9738* | *0,3727* | *Accept null hypothesis* |
| *Day 21 / 23* | *2,4222* | *0,1196* | *1,1801* | *0,2773* | *Alternative combination – Accept null hypothesis* |
| ACRS 273 | Test Statistic | P-Value | Test Statistic | P-Value | Comments |
| Day 20 / 21 / 22 / 23 / 24 / 25 | 67,5392 | 0,0 | 70,0045 | 0,0 | Day 20 has the most offset median and mean |
| Day 21 / 22 / 23 / 24 / 25 | 45,9822 | 2,4840E-9 | 55,9504 | 2,054E-11 | Day 25 has the most offset mean |
| Day 21 / 22 / 23 / 24 | 33,9985 | 1,9826E-7 | 38,6461 | 2,0627E-8 | Day 23 has the most offset median and mean |
| Day 21 / 22 / 24 | 5,9084 | 0,0521 | 7,4277 | 0,0244 | Day 21 has the most offset median and mean |
| *Day 22 / 24* | *0,054* | *0,8162* | *0,3138* | *0,5753* | *Accept null hypothesis* |

**Table 25: Comparing sample medians of ACRS 172 / 173 / 271 / 272 / 273**

| Included Samples | Levene's Test | | Comments |
|---|---|---|---|
| | Test Statistic | P-Value | |
| ACRS 172: Day 20 / 22 / 23 / 24 / 25 | 0,6371 | 0,6362 | Accept null hypothesis |
| ACRS 173: Day 20 / 21 / 22 / 24 | 0,5496 | 0,6485 | Accept null hypothesis |
| ACRS 173: Day 23 / 25 | 0,7061 | 0,4008 | Accept null hypothesis |
| ACRS 271: Day 22 / 23 / 24 / 25 | 0,9112 | 0,4349 | Accept null hypothesis |
| ACRS 271: Day 21 / 22 | 0,2733 | 0,6013 | Accept null hypothesis |
| ACRS 272: Day 20 / 24 / 25 | 0,0060 | 0,9941 | Accept null hypothesis |
| ACRS 272: Day 21 / 23 | 1,9349 | 0,1648 | Accept null hypothesis |
| ACRS 273: Day 22 / 24 | 5,8809 | 0,0158 | Reject null hypothesis |

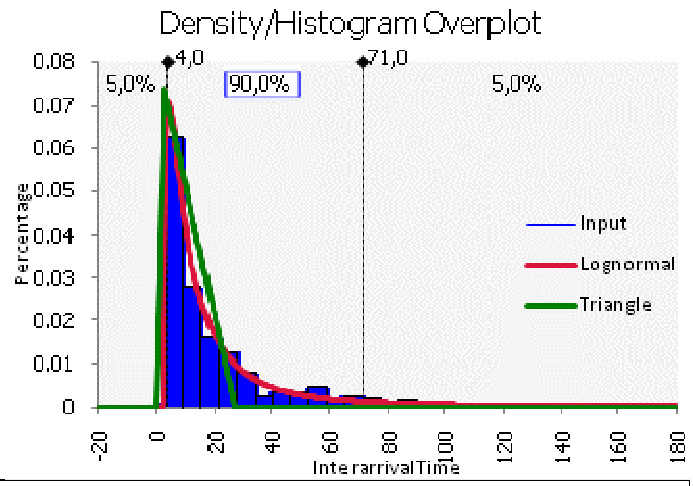
**Table 26: Comparing sample standard deviations of ACRS 172 / 173 / 271 / 272 / 273**

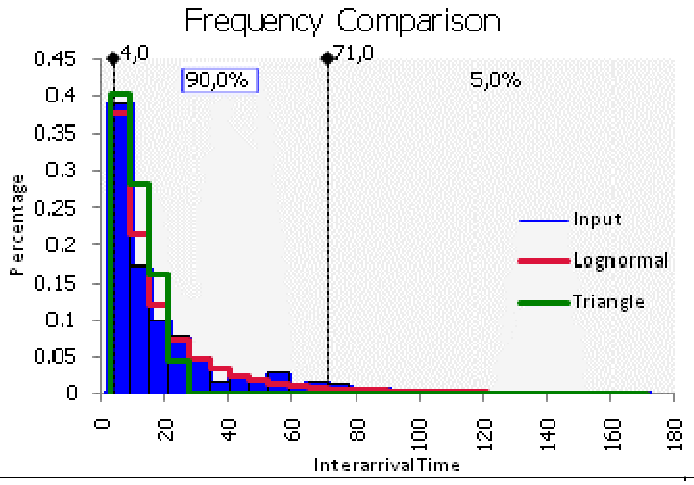
From Table 25 and Table 26 it can be derived which samples may be combined. Only for ACRS 273 it is not possible to combine samples. Consequently, the comparison between the distribution and the real system sample will only be based on a very limited sample size.

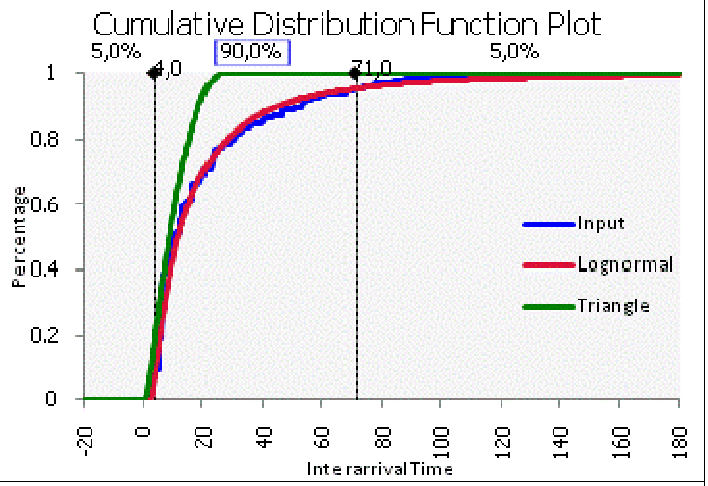
### *B.3.6 Comparing ACRS samples to input distributions*

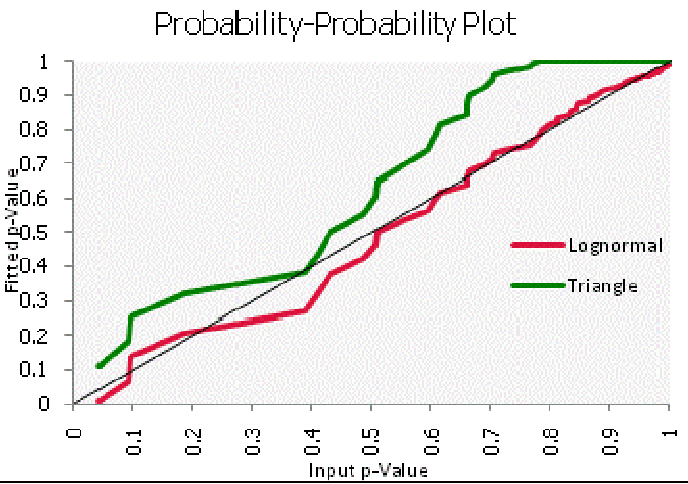
In this section comparisons can be observed between real system samples and input distributions, as well as best fitting distributions. These comparisons consist of graphical techniques as well as goodness-of-fit tests.

| ACRS 172 | Data *(Cr. Value @0,05)* | Triangle Distr. | Lognormal Distr. |
|---|---|---|---|
| Minimum | 3 | 0 | 2,3680 |
| Maximum | 171 | 27 | +Infinity |
| Mean | 20,8297 | 10 | 21,7050 |
| Mode | 3 | 3 | 4,7376 |
| Median | 11 | 9 | 11,9727 |
| Std. Dev. | 23,5967 | 6,0415 | 33,7890 |
| *K-S* | *0,0575** | *0,2670* | *0,1207* |
| *A-D* | *2,4920** | *+Infinity* | *5,8502* |
| *Chi-Sq* | *33,9244** | *602,0417* | *532,6667* |



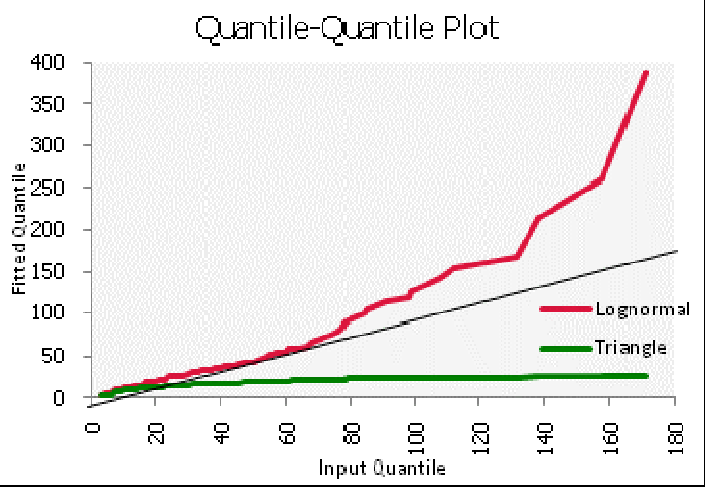
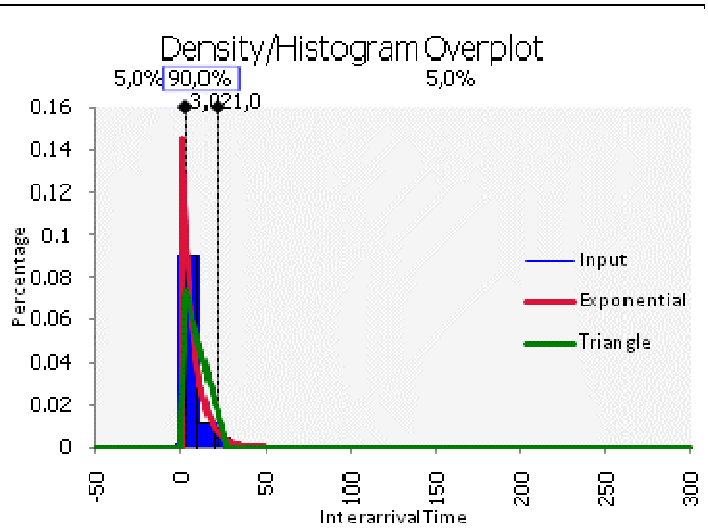
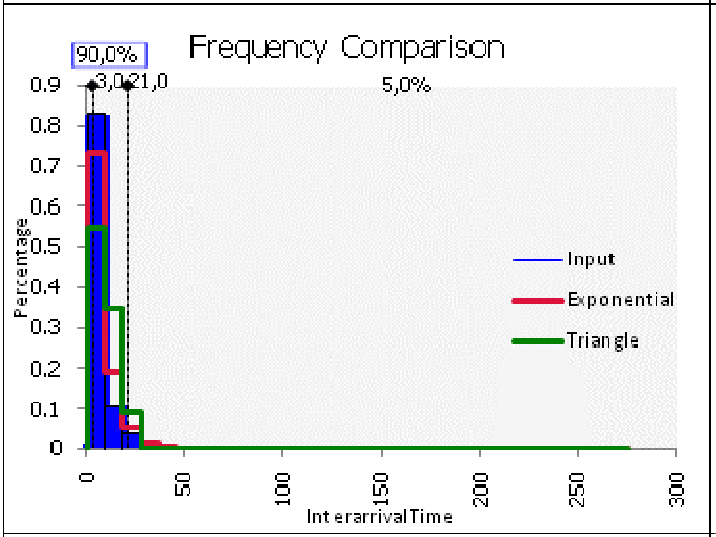
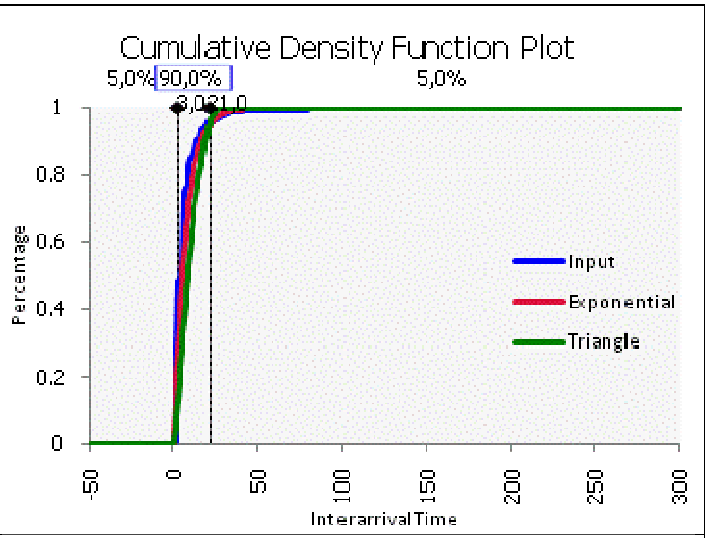



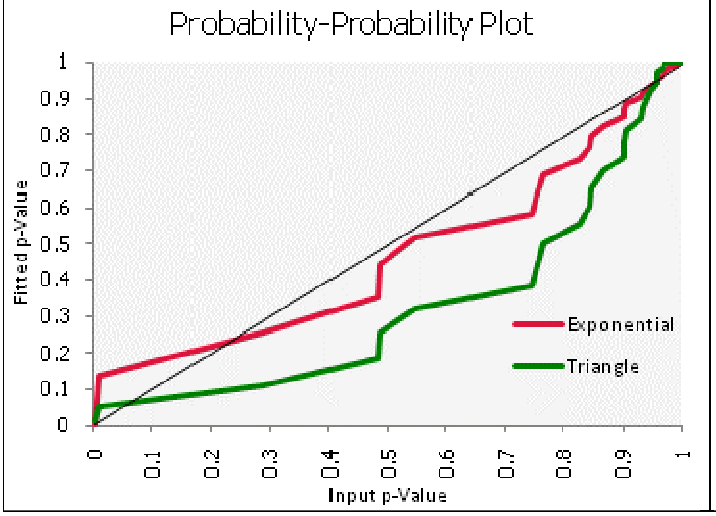

Table 27: Graphical comparisons for ACRS 172 – Sample: Day 20 / 22 / 23 / 24 / 25

| ACRS 173 | Data *(Cr. Value @0,05)* | Triangle Distr. | Exponential Distr. |
|---|---|---|---|
| Minimum | 1 | 0 | 0,9942 |
| Maximum | 275 | 27 | +Infinity |
| Mean | 7,8890 | 10 | 7,8832 |
| Mode | 3 | 3 | 0,9942 |
| Median | 6 | 9 | 5,7693 |
| Std. Dev. | 11,1145 | 6,0415 | 6,8890 |
| *K-S* | *0,0394** | *0,3639* | *0,2433* |
| *A-D* | *2,4920** | *+Infinity* | *+Infinity* |
| *Chi-Sq* | *43,7730** | *4977,6610* | *4978,8593* |



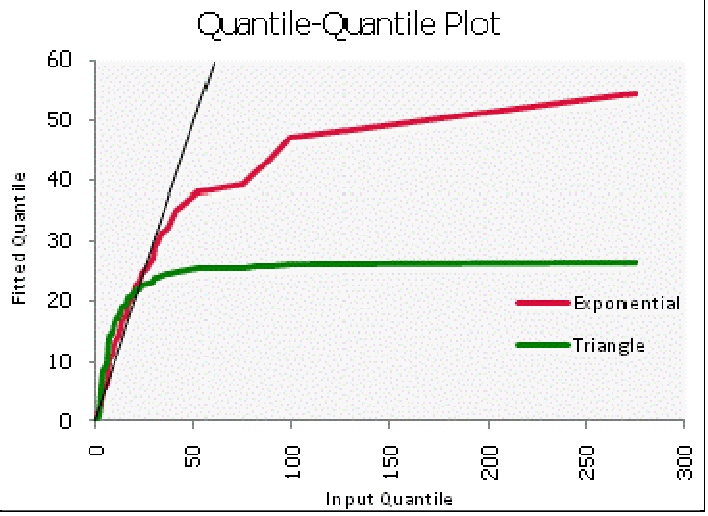
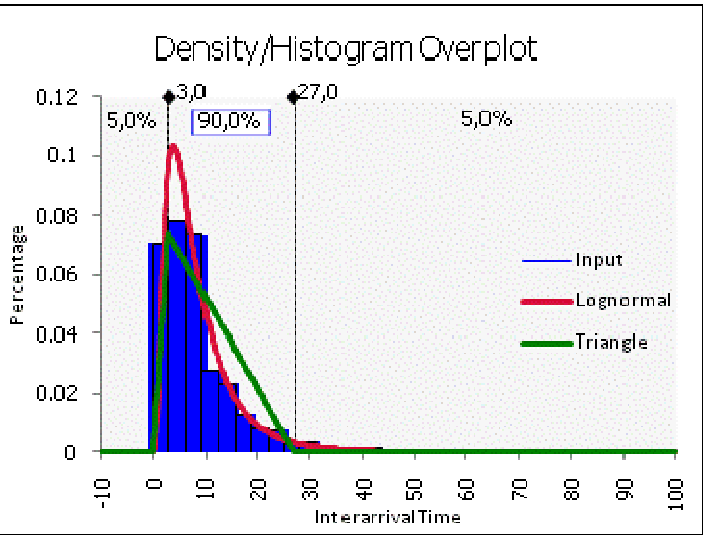
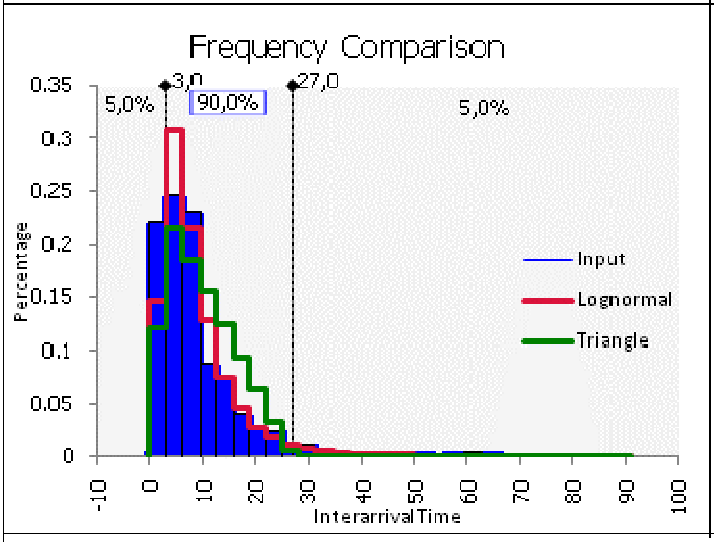
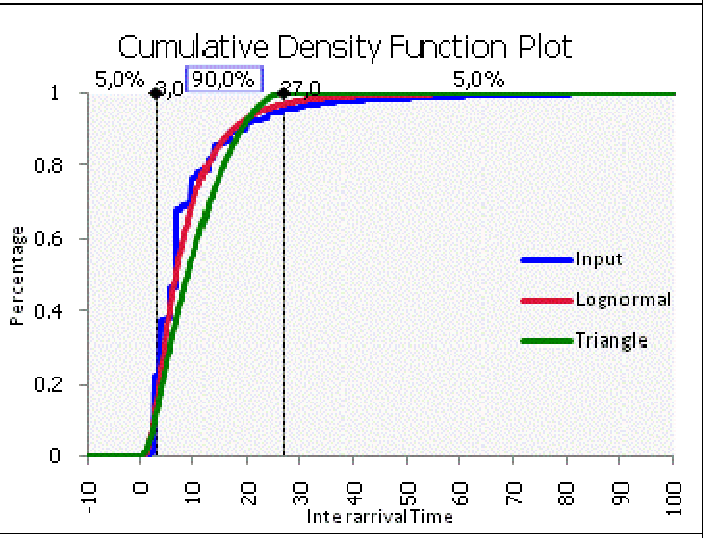



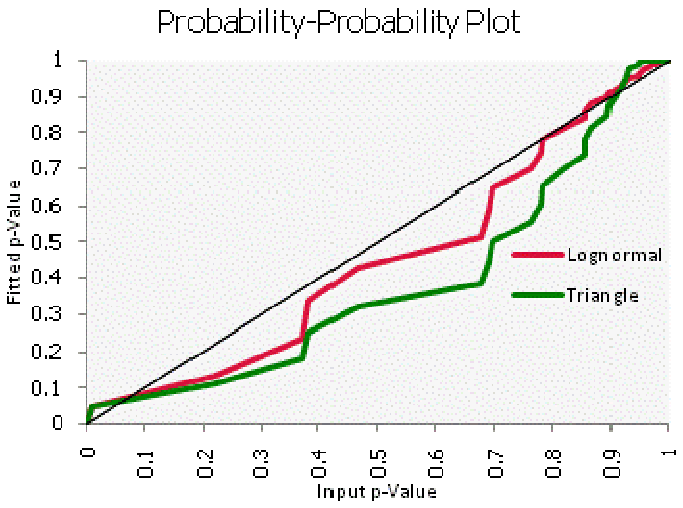

Table 28: Graphical comparisons for ACRS 173 – Sample: Day 20 / 21 / 22 / 24

| ACRS 271 | Data *(Cr. Value @0,05)* | Triangle Distr. | Lognormal Distr. |
|---|---|---|---|
| Minimum | 0 | 0 | -0,1407 |
| Maximum | 91 | 27 | +Infinity |
| Mean | 9,2821 | 10 | 8,8773 |
| Mode | 3 | 3 | 4,0270 |
| Median | 7 | 9 | 6,8315 |
| Std. Dev. | 9,7690 | 6,0415 | 7,3976 |
| *K-S* | *0,0478** | *0,2962* | *0,1657* |
| *A-D* | *2,4920** | *+Infinity* | *22,2255* |
| *Chi-Sq* | *41,3371** | *2724,8105* | *2732,2063* |



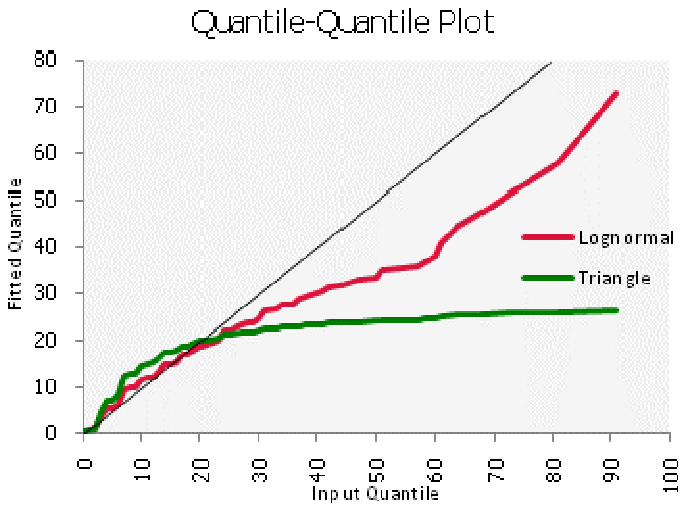
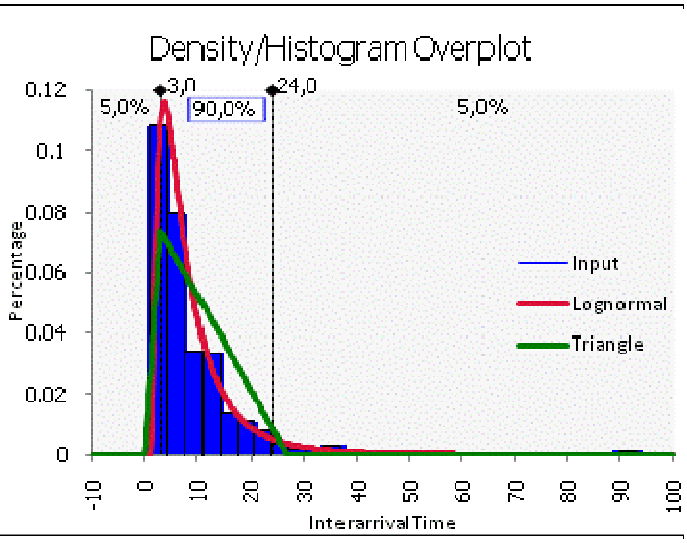
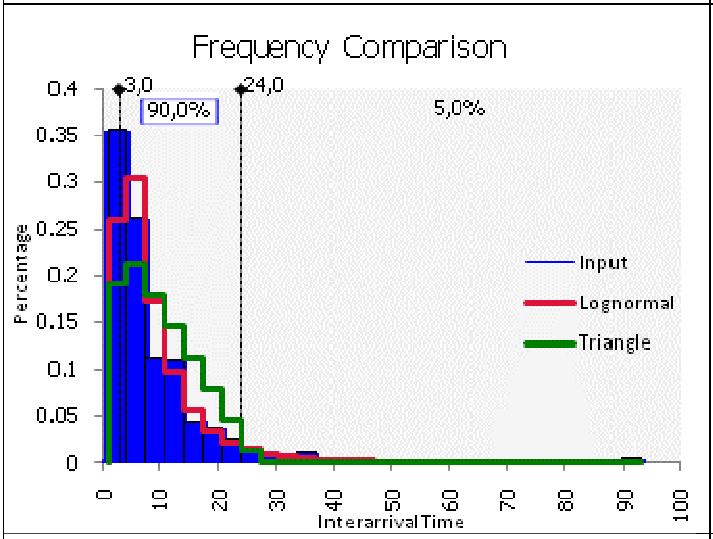
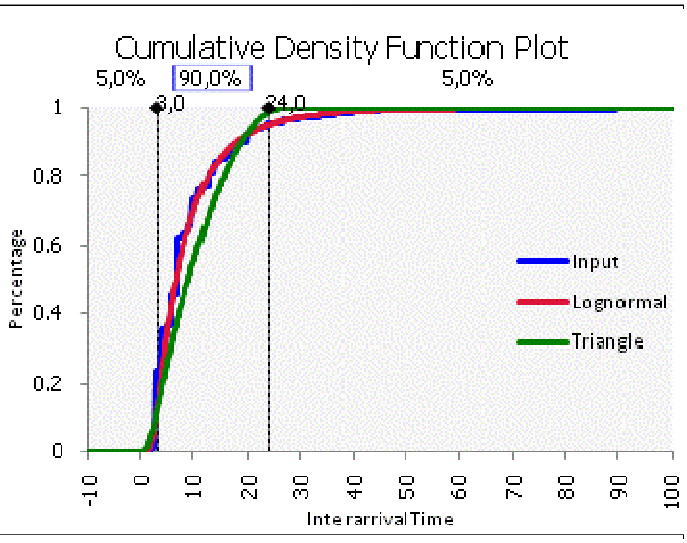



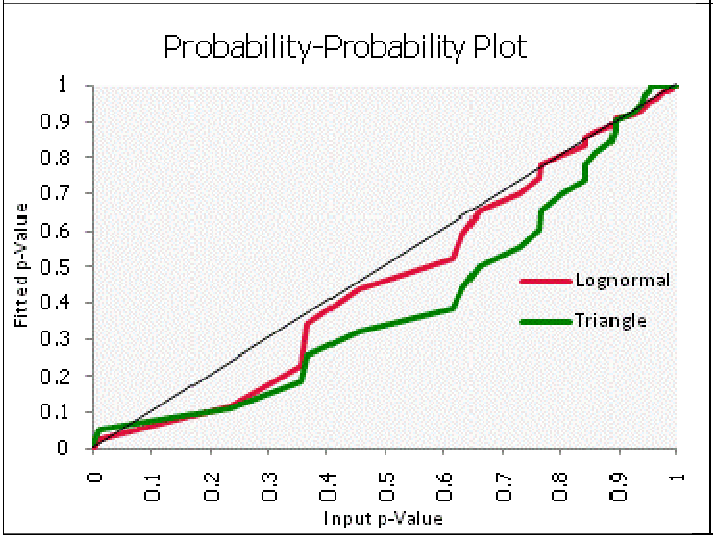

Table 29: Graphical comparisons for ACRS 271 – Sample: Day 22 / 23 / 24 / 25

| ACRS 272 | Data *(Cr. Value @0,05)* | Triangle Distr. | Lognormal Distr. |
|---|---|---|---|
| Minimum | 1 | 0 | 0,8843 |
| Maximum | 93 | 27 | +Infinity |
| Mean | 9,4061 | 10 | 9,1911 |
| Mode | 3 | 3 | 3,7589 |
| Median | 7 | 9 | 6,7162 |
| Std. Dev. | 9,5867 | 6,0415 | 8,4255 |
| *K-S* | *0,0492** | *0,2363* | *0,1291* |
| *A-D* | *2,4920** | *+Infinity* | *12,7089* |
| *Chi-Sq* | *37,6525** | *1550,1984* | *1560,4074* |



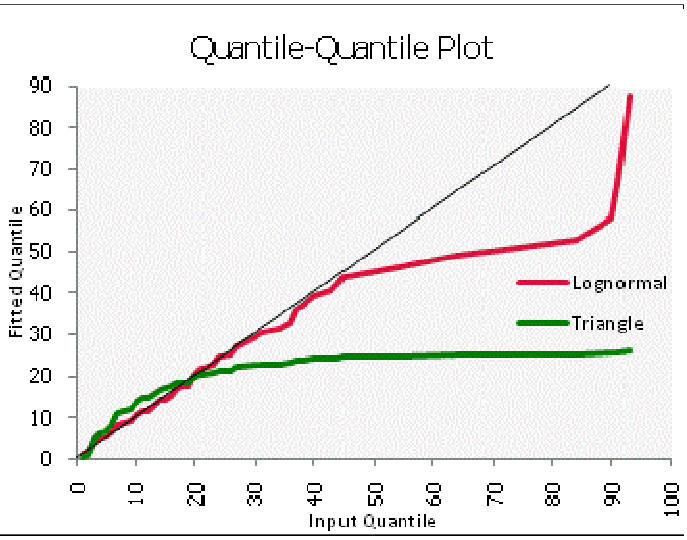
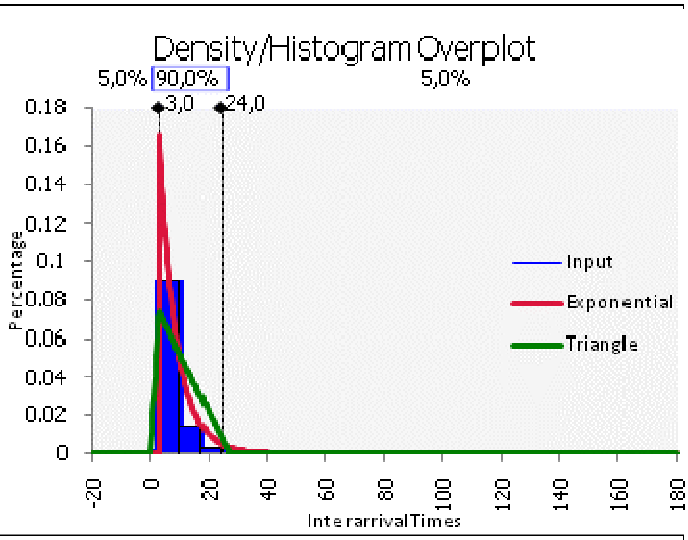
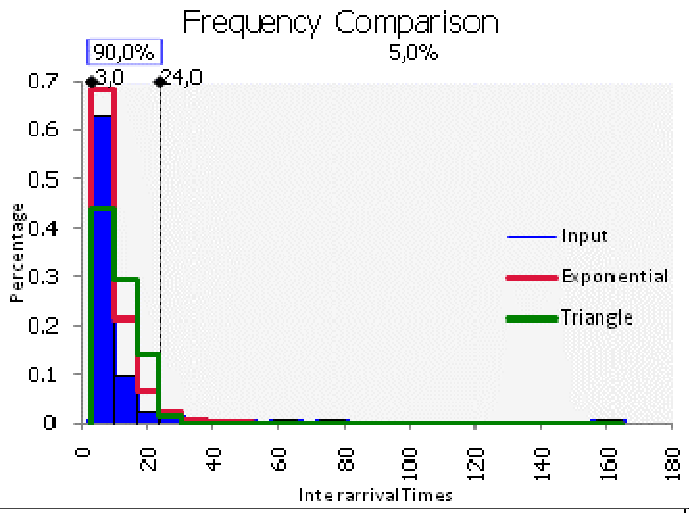
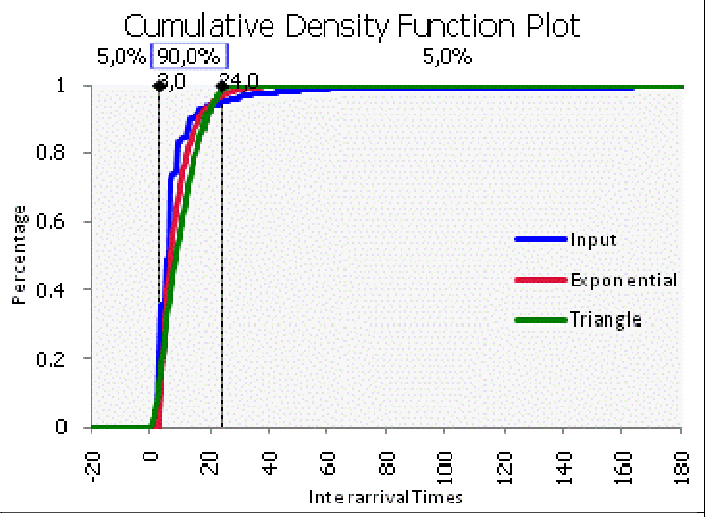



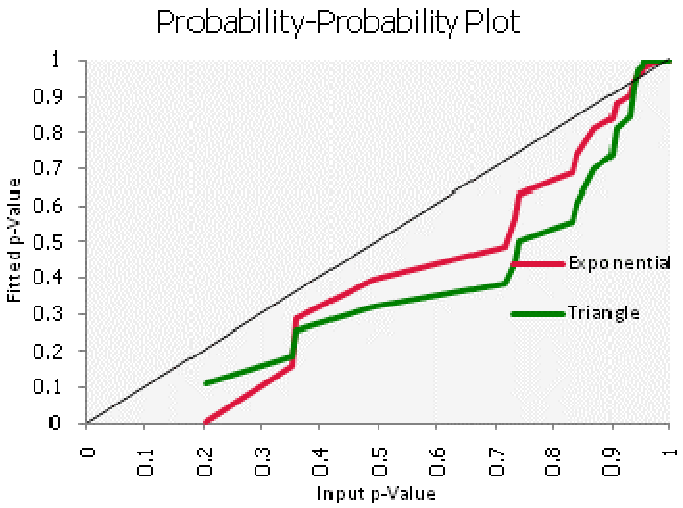

Table 30: Graphical comparisons for ACRS 272 – Sample: Day 20 / 24 / 25

| ACRS 273 | Data *(Cr. Value @0,05)* | Triangle Distr. | Exopnential Distr. |
|---|---|---|---|
| Minimum | 3 | 0 | 2,9746 |
| Maximum | 164 | 27 | +Infinity |
| Mean | 9,0336 | 10 | 9,0082 |
| Mode | 3 | 3 | 2,9746 |
| Median | 7 | 9 | 7,1568 |
| Std. Dev. | 13,2719 | 6,0415 | 6,0336 |
| *K-S* | *0,0873** | *0,3358* | *0,2316* |
| *A-D* | *2,4920** | *+Infinity* | *36,4022* |
| *Chi-Sq* | *24,9958** | *344,2857* | *338,4034* |



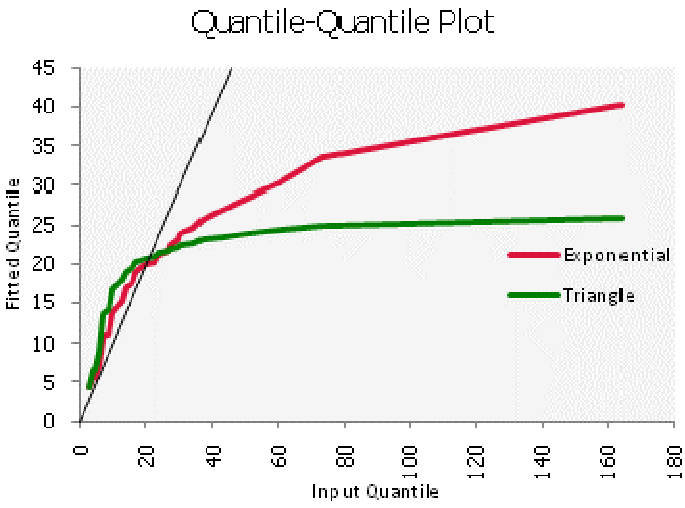




Table 31: Graphical comparisons for ACRS 273 – Sample: Day 24

## B.4 Additional Trace-Driven Output Validation Results

This section contains additional trace-driven output validation results. For bags requiring the EBS and bags going to inspection, summary statistics Box-Whisker plots and histogram comparisons can be observed. For the right subsystem a correlation table and a behavior graph are presented.

### B.4.1 Summary Statistics

| System | Run 1 | Run 2 | Run 3 |
|---|---|---|---|
| Correlation | -0,03 | -0,0563 | 0,0469 |
| P-Value | 0,6110 | 0,3401 | 0,4273 |

Table 32: Spearman rank correlations between the system and various runs, based on the right subsystem

| Statistic | System | Run 1 | Run 2 | Run 3 |
|---|---|---|---|---|
| Count | 1616 | 3198 | 3136 | 3198 |
| Mean | 00:34:50 | 00:26:32 | 00:26:29 | 00:26:41 |
| 5% Trimmed mean | 00:32:36 | 00:25:06 | 00:25:01 | 00:25:14 |
| Median | 00:30:38 | 00:22:07 | 00:22:09 | 00:22:17 |
| Standard deviation | 00:21:33 | 00:15:31 | 00:15:44 | 00:15:33 |

Table 33: Summary statistics of bags requiring the EBS

| Statistic | System | Run 1 | Run 2 | Run 3 |
|---|---|---|---|---|
| Count | 106 | 29 | 37 | 27 |
| Mean | 00:03:04 | 00:02:36 | 00:02:36 | 00:02:38 |
| 5% Trimmed mean | 00:02:58 | 00:02:35 | 00:02:35 | 00:02:37 |
| Median | 00:02:57 | 00:02:34 | 00:02:34 | 00:02:34 |
| Standard deviation | 00:00:52 | 00:00:13 | 00:00:16 | 00:00:15 |

Table 34: Summary statistics of bags going to inspection

| Statistic | System |
|---|---|
| Count | 5359 |
| Mean | 00:05:59 |
| 5% Trimmed mean | 00:05:15 |
| Median | 00:05:30 |
| Standard deviation | 00:04:26 |

Table 35: Summary statistics of bags switching subsystems

### B.4.2 Box-Whisker Plots

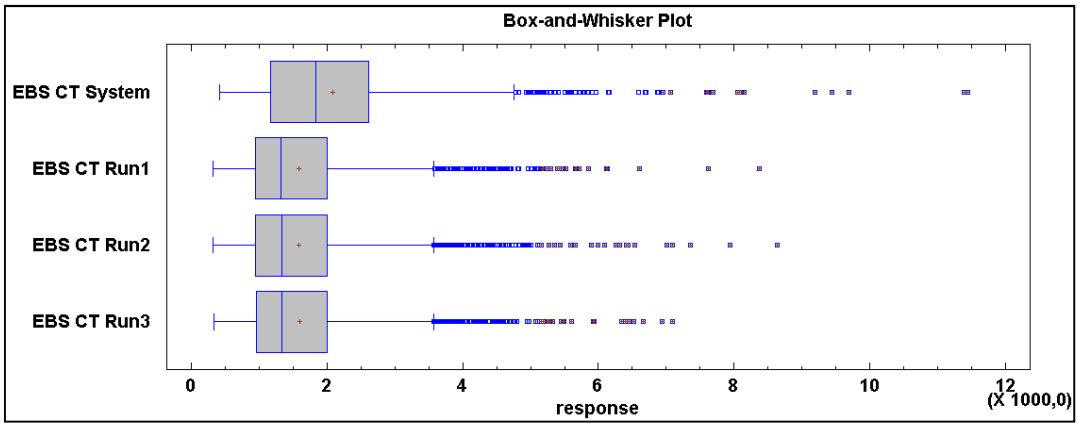



Figure 39: Box-Whisker plots of bags requiring the EBS

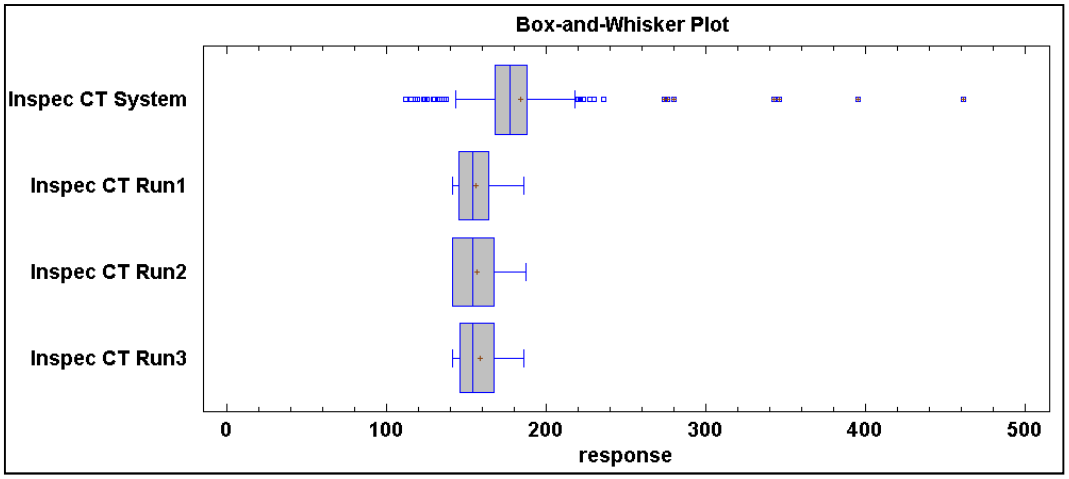



Figure 40: Box-Whisker plots of bags going to inspection

### B.4.3 Histograms

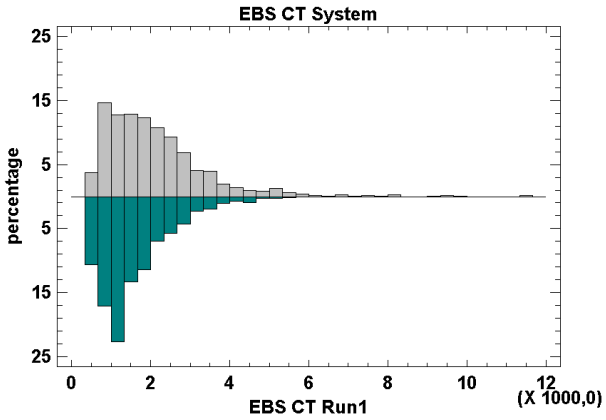



Figure 41: Histograms of bags requiring the EBS

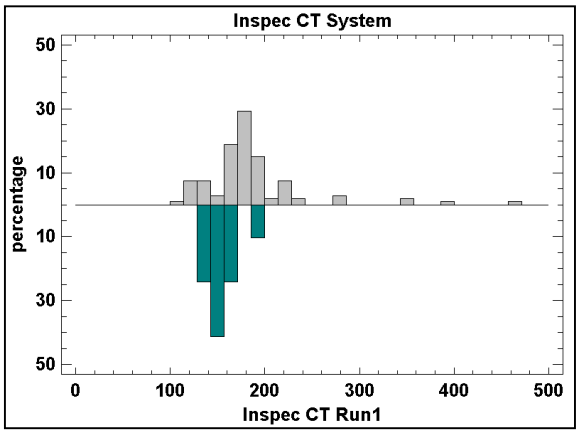



Figure 42: Histograms of bags going to inspection

### *B.4.4 Behavior Graphs*

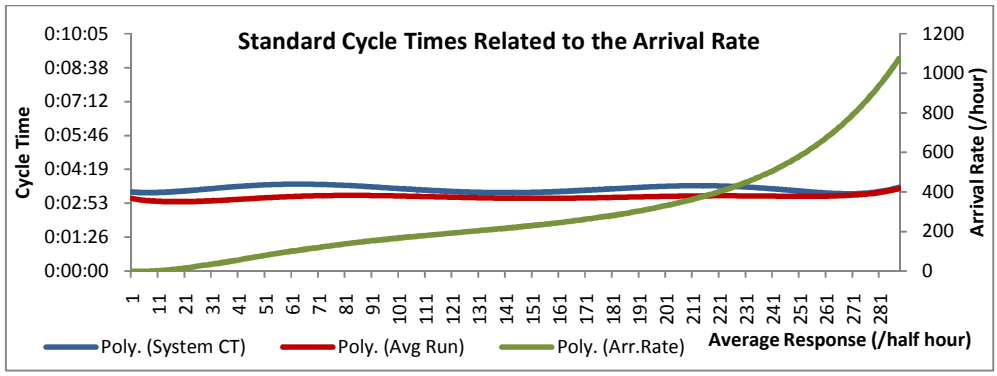



Figure 43: Standard cycle times related to the arrival rate, based on the right subsystem

Note that no behavior graphs have been constructed of the cycle times of bags requiring the EBS or going to inspection. The influence of an increased arrival rate is on bags requiring the EBS is relatively very small. Consequently, this effect is practically insignificant. Related to bags going to inspection, the number of bags traveling to the inspection zone is too small to observe a trend related to the arrival rate.

## B.5 Sensitivity Analysis

### B.5.1 Cumulative Distribution Functions

*Triangular Distribution* (Palisade Corporation, 2009): with minimum, most likely and maximum value.

$$F_{(X)} = \frac{(X - min)^2}{(m.likely - min)(\max - min)} \qquad \min \leq X \leq m.likely$$

$$F_{(X)} = 1 - \frac{(max - X)^2}{(max - m.likely)(max - min)} \qquad \text{m.likely} \leq X \leq max$$

*Exponential Distribution* (Palisade Corporation, 2009)

$$F_{(X)} = 1 - e^{-X\lambda}$$

*Lognormal Distribution* (Palisade Corporation, 2009)

$$F_{(X)} = \Phi\left(\frac{\ln x - \mu'}{\sigma'}\right)$$

$$with\ \mu' = \Phi\left[\frac{\mu^2}{\sqrt{\sigma^2+\mu^2}}\right]\ and\ \sigma' = \sqrt{ln\left[1 + \left(\frac{\sigma}{\mu}\right)^2\right]}$$

Where Φ(z) is the cumulative distribution function of a Normal(0,1) distribution

### B.5.2 Determination of Factor Levels

The factor "distribution" assesses the system's sensitivity related to changes within input parameters. Two suitable distributions have been evaluated already in section 5.2.4; a standard triangular distribution for all input lanes, and lane specific best fitting distribution. These input distributions are stochastic and therefore will require many replications of the experiment, in order to acquire an accurate estimate of factor effects.

The standard reject rates of level 1/2 and level 3/4 screening are respectively 5 % and 1 %. The reject rates as perceived in the real system can be observed in Table 36. It shows the amount of screened and rejected bags, as well as the resulting reject rate of the different days of the various screening levels. Alternative values of 7,7 % and 7,1 % will be used for level 1/2 and level 3/4, as those deviate the most from the standard values.

| | L1/2 Screening | | | L3/4 Screening | | |
|---|---|---|---|---|---|---|
| | #Screened | #Rejected | Reject Rate | #Screened | #Rejected | Reject Rate |
| Day 20 | 8235 | 509 | 0,0618 | 350 | 8 | 0,0229 |
| Day 21 | 9085 | 560 | 0,0616 | 397 | 16 | 0,0403 |
| Day 22 | 10125 | 779 | 0,0769 | 560 | 20 | 0,0357 |
| Day 23 | 10996 | 704 | 0,0640 | 493 | 32 | 0,0649 |
| Day 24 | 9906 | 476 | 0,0481 | 297 | 9 | 0,0303 |
| Day 25 | 10153 | 452 | 0,0445 | 294 | 21 | 0,0714 |
| Total | 58500 | 3480 | 0,0595 | 2391 | 106 | 0,0443 |

Table 36: Real system screening reject rates

In Table 37 the amount of unique bags per day can be observed passing the MES or EBS (thus exclusive flush-backs). In order to convert these values to rates, they are being compared to the amount of bags leaving the system at the various laterals. Since the rates applied in the simulation model are 1 % for MES and 5 % for the EBS, the most deviating values are respectively 0,4 % and 2,6 % (rounded).

| | #MES | #Lateral | Rate | #EBS | #Lateral | Rate |
|---|---|---|---|---|---|---|
| Day 20 | 53 | 8305 | 0,0064 | 264 | 8305 | 0,0318 |
| Day 21 | 40 | 8745 | 0,0046 | 263 | 8745 | 0,0301 |
| Day 22 | 74 | 9690 | 0,0076 | 475 | 9690 | 0,0490 |
| Day 23 | 52 | 9816 | 0,0053 | 387 | 9816 | 0,0394 |
| Day 24 | 47 | 9655 | 0,0049 | 312 | 9655 | 0,0323 |
| Day 25 | 42 | 9719 | 0,0043 | 248 | 9719 | 0,0255 |
| Total | 308 | 55930 | 0,0055 | 1949 | 55930 | 0,0348 |

**Table 37: Real system MES and EBS rates**

Recall that EBS control is constituted of the lane assignment when entering the EBS for the first time, the probability to return to the EBS after a flush (flush-back), the probability to change lane in case a bag flushes back, and the time between flushes. It should be noted that the assumption made within the simulation model that both Early Back Stores are configured identically, will be maintained. Therefore it is assumed that deviations between both stores are the result of real world circumstances that are not taken into account in the simulation model.

In Table 38 the amount of unique bags can be found that require the usage of the EBS, for both the left and the right EBS. Furthermore, it can be observed how these bags are divided over the two available lanes per EBS. Comparing the percentages devoted to lane 1 to the 60 % used within the standard simulation model, an alternative level of 56% will be selected.

| | EBS lane 1 | | EBS Lane 2 | |
|---|---|---|---|---|
| | Amount | Percentage | Amount | Percentage |
| Left EBS | 480 | 0,6275 | 285 | 0,3725 |
| Right EBS | 668 | 0,5642 | 516 | 0,4358 |
| Total | 1148 | 0,5890 | 801 | 0,4110 |

**Table 38: System based EBS lane assignment**

In Table 39 the flush-backs to the EBS are denoted. The first column describes the total amount of flush-backs of a bag, while the second column indicates the amount of baggage items that returned this amount of times. In total 430 bags return 589 times, relating to an overall flush-back percentage of approximately 30%.

| Max. EBS Flush-Backs | Frequency |
|---|---|
| 1 | 327 |
| 2 | 58 |
| 3 | 40 |
| 4 | 3 |
| 5 | 0 |
| 6 | 0 |
| 7 | 2 |
| Total Bags Flushed Back | 430 |
| Total EBS Flush-Backs | 589 |
| Overall % Flush-Backs | 0,3022 |

**Table 39: System based EBS flush-back rates**

Additional flush-back information is shown in Table 40, where the first flush-back of all returning bags is analyzed in more detail. The information is constrained to the first flush-back due to data complexity. It is assumed that the first flush-back is representative for the small amount of additional flush-backs. From the table it can be observed that a significant difference exists between the left and the right EBS. Because it appears that the left EBS has hardly been used, the alternative flush-back rates will be based upon the right EBS. The flush-back rate is calculated by combining the frequencies with the information of Table 38. Since both lanes of the right EBS have a flush-back rate of 30 %, this percentage will be used within the alternative simulation model configuration.

| | EBS Lane 1 | | EBS Lane 2 | |
|---|---|---|---|---|
| Flush-Back: | Amount | Rate | Amount | Rate |
| Left EBS | 27 | 0,0563 | 42 | 0,1474 |
| Right EBS | 205 | 0,3069 | 156 | 0,3023 |
| Total | 232 | 0,2021 | 198 | 0,2472 |

**Table 40: Detailed lane flush-back, based upon a bag's first return**

In Table 41 information can be observed about the percentage of bags changing EBS lane when recirculating. As for the EBS flush-back rate, this information is based upon partial data, the baggage items that return only once, due to data complexities. Furthermore, no distinction is made between the left and the right EBS. It is assumed that they both behave similarly.

Within the standard simulation model, lane 1 was specified as a relatively short term storage unit (0 % switches to lane 2), while lane 2 was specified as a relatively long term storage unit (75 % switches to lane 1). Within the real system it appears that the lanes are configured somewhat differently; it seems that they both have a more similar application area. The percentages found within the data will be applied as such within the alternative simulation configuration.

| | EBS lane 1 to 2 | EBS lane 2 to 1 | Total |
|---|---|---|---|
| Amount | 201 | 125 | 326 |
| Percentage | 0,3413 | 0,2122 | 0,5535 |

**Table 41: System based EBS lane switching**

Also the EBS time-to-flush is based upon the first flush-back of all recirculating bags (Figure 44) (note that the amounts of bags per lane resemble the amounts in Table 40). In the figure the time difference can be observed between the registration time of a bag entering an EBS lane, and the time that it is scanned again when entering the sorter (thus somewhat more than the time a bag stays in an EBS). This information is plotted for succeeding bags arriving at the EBS (per lane). The time that a bag stays in the EBS declines until the lane is flushed. Therefore the time a bag stays in an EBS just after such minimum is an indication for the time-to-flush of an EBS lane.

In the simulation model the time-to-flush is modeled as a constant time. Though, the small amount of high peaks in Figure 44 may indicate that this time is dependent on the circumstances. Because no exact information about the configuration of the EBS can be found in the functional specification (Vanderlande Industries, 2009b), the best approximation based on a constant time-to-flush will be used. Taking the figure into account, a time-to-flush of 30 minutes seems appropriate.

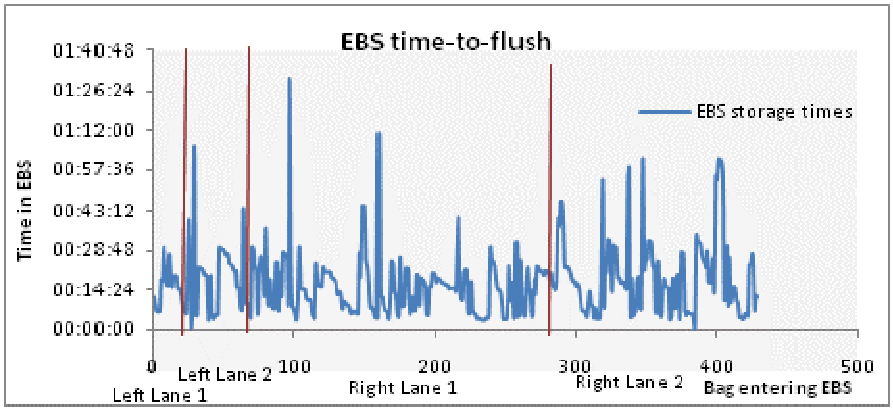



Figure 44: Time that bags are stored in an EBS, before getting flushed

Another factor is related to recurrence to manual coding after usage of the EBS. More specifically, when a license plate code is unreadable a bag needs to travel to the manual encoding station. A bag in the EBS is out of tracking. Therefore it has to be identified again when leaving the EBS, which happens at the sorter. Combining this information leads to the statement that a bag that requires to be stored in the EBS, with an LPC that has turned out to be unreadable, has a high probability of requiring manual encoding for identification again when leaving the EBS before returning its course. This logic has been depicted (based upon discussions with simulation engineers) as the reasoning behind Table 42, which indicates that recurrence to manual encoding takes place (in 10 % of the cases). Faulty manual encoding has been turned down as a possible cause. Recurrence to manual encoding is not taken into account in the standard simulation model. Within the alternative simulation model it is assumed that all bags going to the EBS and have been codec manually, require manual coding again when leaving the EBS.

| # MES Recurrence | Frequency |
|---|---|
| 1 | 20 |
| 2 | 4 |
| 3 | 0 |
| 4 | 1 |
| Total MES Recurrences | 32 |
| Recurrence Rate | 0,1039 |

Table 42: System based MES recurrence rates

In Table 43 the standard and the alternative configuration can be found of model components. More specifically, factor information is given about velocities and window lengths. In practice component velocities are typically set somewhat higher than pre-specified. Therefore alternative velocities are set 0,1 m/s higher. In order to determine an alternative level for the window length, it was decided to apply the inverse effect of the velocity increment on capacity. Thus, using both alternative levels for velocity and window length, the capacity remains the same. This may be useful since capacity levels are specified very accurate, whereas exact velocity levels and window length are uncertain.

| Component | Standard Configuration | | | Alternative Configuration | | |
|---|---|---|---|---|---|---|
| | Cap. (b/h) | V (m/s) | $L_{Window}$ (m) | Cap. (b/h) | V (m/s) | $L_{Window}$ (m) |
| Belt Floorveyor | 1500 | 1 | 2,4 | 1500 | 1,1 | 2,64 |
| Belt In Tracking | 1200 | 1 | 3 | 1200 | 1,1 | 3,3 |
| Speed Reduction 1 | 1200 | 0,75 | 2,25 | 1200 | 0,85 | 2,55 |
| Speed Reduction 2 | 1200 | 0,5 | 1,5 | 1200 | 0,6 | 1,8 |
| L1 Screener | 1200 | 0,33 | 1 | 1200 | 0,43 | 1,29 |
| L3 Screener | 300 | 0,25 | 3 | 300 | 0,35 | 4,2 |
| Vertisorter | 1800 | 1 | 2 | 1800 | 1,1 | 2,2 |
| Flat Triplanar Sorter | 1500 | 1 | 2,4 | 1500 | 1,1 | 2,64 |
| Divert Parallel Pushers | >1500 | - | - | >1500 | - | - |
| EBS Lane | >100 bags | 1 | 1 ,2 (V=0) | >60 bags | 1,1 | 2 (V=0) |

**Table 43: Standard and alternative component characteristics**

### *B.5.3 Defining factors and factor groups*

For performing a sensitivity analysis all parameters discussed for the configuration of the model may be considered (appendix B.1.1), as well as the input distributions evaluated earlier. However, this would result in a very large number of runs to be performed. Furthermore, AutoStat is only capable of handling up to 11 factors in a design of experiments. Consequently, factor grouping has to be applied. Probably the most self-evident factors to combine are the velocities and window lengths of the various section types. It reduces the amount of factors greatly, while the direction of the effects is expected to be identical for all parameters within the groups.

Nonetheless, this reduction is not sufficient. Therefore, various parameters concerning EBS control have been evaluated as a single factor as well: lane assignment, time between flushes, probability to return to the EBS after a flush (flush-back), and the probability to change lane in case a bag flushes back. Combining these factors reduces the amount of parameters of appendix B.1.1 to 11. A resolution V design with 11 factors implies that 128 runs should be performed per replication.

### *B.5.4 Additional Sensitivity Analysis Results*

The main factors affecting the cycle time of bags requiring EBS usage are the EBS rate, EBS control, the distribution and velocity (Figure 45). Also the interaction between distribution and the EBS rate, and the interaction between the EBS rate and velocity are of relatively high importance. Other effects appear to be practically insignificant, due to their size, and in many cases statistically insignificant.

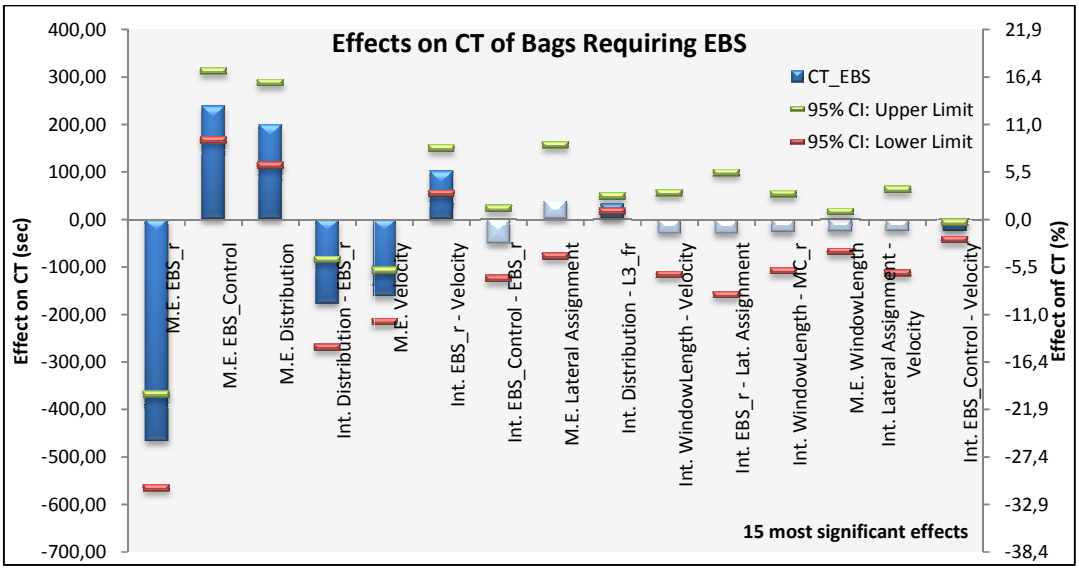



Figure 45: Factor effects on the cycle time of bags that require EBS usage

From Figure 46 it can be observed that velocity, service time, EBS rate, and distribution have a relatively high effect on the cycle time of bags that leave the system at an inspection location. Though, the effects of EBS rate and Distribution appear to be barely statistically significant. It is noticeably that only a small proportion of factor effects is statistically significant.

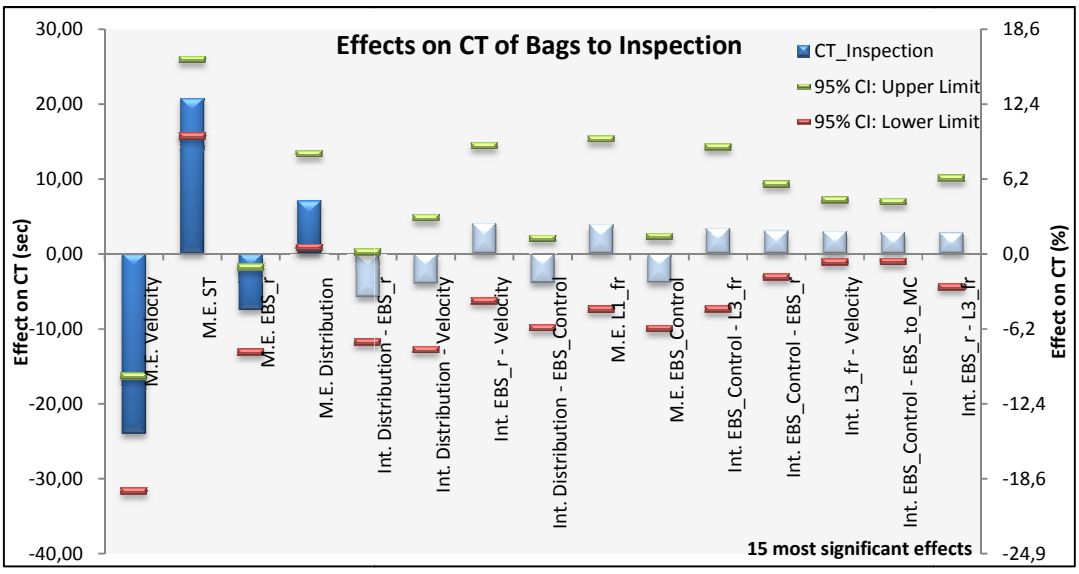



Figure 46: Factor effects on the cycle time of bags that leave the system at the inspection locations

Figure 47 presents factor effects on the work in process of the baggage handling system. The layout of the figure is identical to the layout of the previous figures addressing cycle times. The EBS rate, velocity, distribution, lateral assignment, EBS control, and some of their interactions mainly influence the work in process.

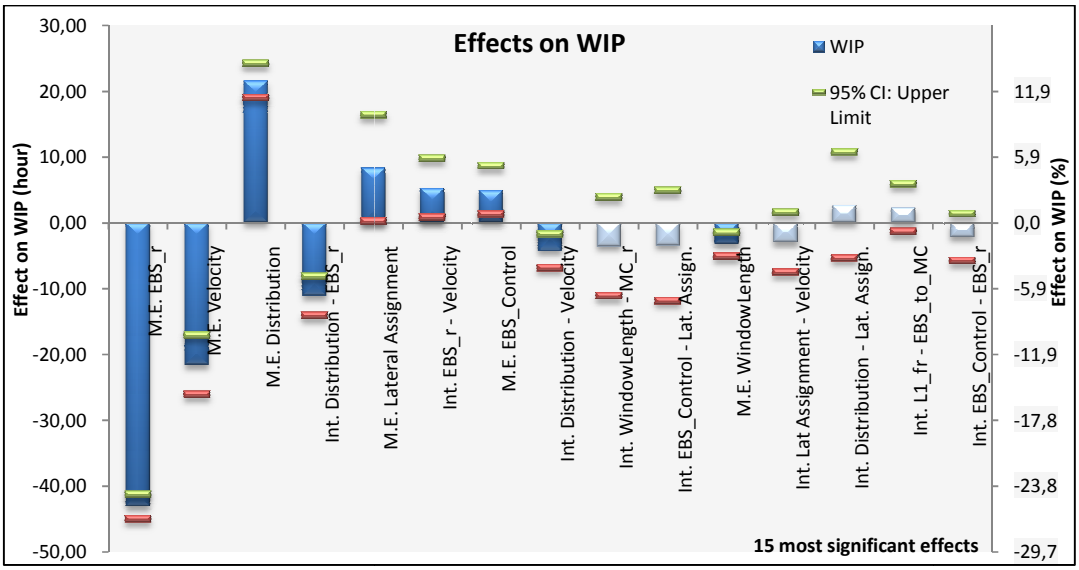



Figure 47: Factor effects on the work in process of the baggage handling system

### *B.5.5 Detailed Results of the Sensitivity Analysis*

In the tables that follow, the exact estimates of factor effects can be found, as well as their standard deviations.

| | | Distribution | WindowLength | EBS_Control | EBS_r | L1_fr | L3_fr | Lateral Assignment | MC_r | ST | Velocity | EBS_to_MC | Int. Distribution - WindowLength | Int. Distribution - EBS_Control | Int. Distribution - EBS_r | Int. Distribution - L1_fr | Int. Distribution - L3_fr | Int. Distr. - Lateral Assignment | Int. Distribution - MC_r | Int. Distribution - ST | Int. Distribution - Velocity | Int. Distribution - EBS_to_MC | Int. WindowLength - EBS_Control |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CT_EBS | E[X] | 200.2 | -26.2 | 239.2 | -467.5 | 21.1 | -4.6 | 38.9 | -8.2 | -7.2 | -162.0 | 13.8 | -6.0 | -8.3 | -177.5 | 10.7 | 32.7 | 16.2 | -7.4 | -0.7 | -18.5 | 1.9 | 13.5 |
| | σ[X] | 35.1 | 16.9 | 29.2 | 39.6 | 29.2 | 17.3 | 47.2 | 35.4 | 4.8 | 21.7 | 6.7 | 22.3 | 26.7 | 37.3 | 38.7 | 6.4 | 58.6 | 43.5 | 26.5 | 4.6 | 7.0 | 12.9 |
| CT_Inspection | E[X] | 7.1 | -2.2 | -3.9 | -7.5 | 3.9 | 0.1 | 1.4 | -1.2 | 20.8 | -24.0 | -1.5 | -2.3 | -3.9 | -5.8 | 2.4 | -1.8 | 1.2 | -0.9 | -0.3 | -4.0 | -2.4 | 2.1 |
| | σ[X] | 2.5 | 0.8 | 2.5 | 2.3 | 4.6 | 4.2 | 3.3 | 3.9 | 2.1 | 3.1 | 2.5 | 0.8 | 2.4 | 2.4 | 3.8 | 3.8 | 3.1 | 4.6 | 2.5 | 3.5 | 2.2 | 0.6 |
| CT_Standard | E[X] | 12.0 | -4.4 | -1.9 | -4.2 | 1.2 | -0.6 | 11.1 | -1.4 | -0.6 | -27.5 | -0.8 | -3.1 | -1.5 | -3.0 | -0.4 | -0.2 | 2.9 | -0.4 | -0.6 | -5.4 | -0.8 | 0.8 |
| | σ[X] | 0.5 | 0.5 | 0.5 | 0.6 | 0.9 | 0.9 | 1.0 | 1.0 | 0.3 | 0.9 | 0.6 | 0.4 | 0.5 | 0.6 | 1.0 | 0.9 | 0.9 | 1.0 | 0.3 | 0.9 | 0.7 | 0.4 |
| CT_Total | E[X] | 32.0 | -5.9 | 5.4 | -75.4 | 4.7 | -0.4 | 15.9 | -1.2 | -4.1 | -35.6 | 0.0 | -3.5 | -5.6 | -22.3 | 3.1 | 2.5 | 6.5 | -0.6 | -4.4 | -6.9 | -0.9 | 3.2 |
| | σ[X] | 3.3 | 4.5 | 3.7 | 1.7 | 9.2 | 3.9 | 7.5 | 4.9 | 0.6 | 5.4 | 5.7 | 5.5 | 2.7 | 3.3 | 9.8 | 3.5 | 7.5 | 5.4 | 0.9 | 4.2 | 5.4 | 5.7 |
| TH_Induct1_1 | E[X] | 53.9 | 0.0 | 0.2 | -0.2 | -10.2 | 0.0 | 0.1 | -0.3 | 0.1 | -0.2 | 0.1 | -0.1 | 0.0 | 0.0 | -0.7 | 0.1 | 0.0 | 0.1 | -0.1 | -0.2 | 0.2 | 0.2 |
| | σ[X] | 0.1 | 0.3 | 0.2 | 0.3 | 0.2 | 0.2 | 0.1 | 0.2 | 0.1 | 0.4 | 0.3 | 0.2 | 0.1 | 0.4 | 0.2 | 0.2 | 0.1 | 0.2 | 0.3 | 0.2 | 0.3 | 0.3 |
| TH_Induct1_2 | E[X] | -189.4 | -0.1 | 0.1 | 0.0 | -7.1 | -0.1 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | -0.4 | -0.1 | 2.9 | 0.1 | -0.2 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 |
| | σ[X] | 0.1 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | 0.2 | 0.2 | 0.0 | 0.1 | 0.3 | 0.1 | 0.2 | 0.0 | 0.1 | 0.1 | 0.1 | 0.3 | 0.1 | 0.4 | 0.1 | 0.3 |
| TH_Induct1_3 | E[X] | 90.5 | 0.0 | 0.1 | 0.1 | -10.9 | -0.2 | 0.2 | -0.1 | 0.0 | -0.2 | -0.1 | -0.1 | 0.1 | 0.0 | -1.1 | -0.1 | 0.1 | -0.1 | 0.1 | -0.2 | -0.2 | 0.2 |
| | σ[X] | 0.3 | 0.2 | 0.2 | 0.2 | 0.3 | 0.2 | 0.4 | 0.2 | 0.1 | 0.1 | 0.2 | 0.6 | 0.4 | 0.2 | 0.2 | 0.4 | 0.2 | 0.2 | 0.1 | 0.0 | 0.2 | 0.3 |
| TH_Induct1_41 | E[X] | -2.7 | 0.0 | -0.4 | 0.2 | 26.9 | -4.0 | -0.4 | 0.4 | -0.2 | 0.2 | -0.1 | 0.2 | 0.2 | 0.0 | -0.8 | 0.1 | 0.1 | 0.0 | -0.1 | 0.4 | 0.0 | -0.4 |
| | σ[X] | 0.3 | 0.4 | 0.3 | 0.3 | 0.4 | 0.1 | 0.4 | 0.1 | 0.2 | 0.4 | 0.2 | 0.7 | 0.7 | 0.5 | 0.4 | 0.2 | 0.1 | 0.3 | 0.4 | 0.6 | 0.2 | 0.1 |
| TH_Induct1_42 | E[X] | -2.5 | 0.0 | -27.4 | -40.5 | 0.1 | -0.2 | -0.3 | -0.2 | 0.2 | 0.5 | -0.3 | 0.5 | 0.7 | 1.1 | 0.2 | -0.6 | -0.5 | -0.2 | -0.3 | -0.2 | -0.1 | 0.2 |
| | σ[X] | 0.3 | 0.2 | 0.3 | 0.3 | 0.6 | 0.9 | 0.4 | 0.2 | 0.1 | 0.5 | 0.3 | 0.8 | 0.0 | 0.1 | 0.5 | 0.3 | 0.7 | 0.3 | 0.8 | 0.4 | 1.1 | 0.3 |
| TH_Induct2_1 | E[X] | 29.8 | 0.8 | 3.0 | 6.2 | -12.5 | -0.6 | -2.0 | -0.6 | 2.0 | 0.6 | 0.5 | 0.7 | 2.9 | 6.2 | -2.7 | -0.7 | -2.2 | -0.5 | 1.8 | 0.5 | 0.5 | -1.2 |
| | σ[X] | 4.1 | 4.3 | 2.4 | 4.0 | 5.3 | 2.5 | 2.8 | 4.5 | 0.5 | 3.0 | 3.1 | 4.1 | 2.2 | 4.1 | 5.3 | 2.3 | 2.6 | 4.3 | 0.2 | 2.8 | 3.0 | 5.1 |
| TH_Induct2_2 | E[X] | 21.8 | 1.3 | 4.6 | 9.0 | -12.6 | 0.4 | -3.2 | -0.2 | 2.4 | 1.8 | 0.6 | 1.5 | 4.4 | 8.9 | -2.5 | 0.4 | -3.5 | -0.3 | 2.5 | 2.1 | 0.7 | -2.0 |
| | σ[X] | 3.0 | 3.2 | 3.9 | 3.0 | 6.5 | 2.7 | 4.6 | 3.9 | 0.0 | 2.7 | 2.7 | 3.6 | 4.1 | 2.9 | 6.3 | 2.4 | 4.5 | 3.9 | 0.5 | 2.4 | 3.0 | 3.6 |
| TH Induct2 3 | E[X] | 37.5 | 0.0 | 2.3 | 5.9 | -12.5 | -0.7 | -2.6 | -0.6 | 2.0 | 0.3 | -0.1 | 0.4 | 2.6 | 5.9 | -2.7 | -0.9 | -2.0 | -0.4 | 2.1 | 0.2 | 0.0 | -0.9 |
| | σ[X] | 4.8 | 4.1 | 2.2 | 4.5 | 5.5 | 2.3 | 2.9 | 4.4 | 0.3 | 2.8 | 3.7 | 4.0 | 2.5 | 4.1 | 5.4 | 2.4 | 2.9 | 4.5 | 0.8 | 2.8 | 3.4 | 4.8 |
| TH_Induct2_41 | E[X] | 5.2 | 0.4 | 0.5 | 1.8 | 28.5 | -4.2 | -0.5 | -0.5 | 0.4 | -0.1 | -0.2 | 0.2 | 0.6 | 2.0 | 0.1 | 0.2 | -0.7 | -0.4 | 0.4 | 0.0 | -0.2 | -0.2 |
| | σ[X] | 0.4 | 0.5 | 0.5 | 0.8 | 0.9 | 0.3 | 0.7 | 1.3 | 0.3 | 0.5 | 0.6 | 0.4 | 0.1 | 1.2 | 1.1 | 0.7 | 1.0 | 1.4 | 0.4 | 0.9 | 1.0 | 0.5 |
| TH_Induct2_42 | E[X] | 3.6 | 0.2 | -27.9 | -40.2 | -0.7 | 0.1 | -0.8 | -0.3 | 0.5 | 1.7 | 0.5 | 0.8 | -0.2 | 1.7 | -1.0 | 0.4 | -1.4 | -0.3 | 0.8 | 0.7 | 0.6 | -0.3 |
| | σ[X] | 0.6 | 0.9 | 1.3 | 1.0 | 1.7 | 1.5 | 1.5 | 1.3 | 1.0 | 0.4 | 0.2 | 0.7 | 1.4 | 0.8 | 2.2 | 0.5 | 1.1 | 2.3 | 0.3 | 0.5 | 1.0 | 1.2 |
| TH_Inpection | E[X] | 0.0 | 0.0 | 0.0 | 0.2 | 2.4 | 8.4 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | 0.0 | 0.0 | 0.2 | 0.0 | 0.1 | 0.0 | -0.1 | 0.1 | 0.0 | 0.0 | -0.1 |
| | σ[X] | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.0 | 0.2 | 0.1 | 0.2 | 0.1 | 0.1 | 0.1 | 0.1 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | 0.1 |
| TH_MC1 | E[X] | -0.3 | -0.1 | -0.3 | -0.1 | 0.0 | 0.1 | 0.1 | -6.6 | 0.0 | -0.1 | 0.5 | -0.1 | -0.1 | 0.0 | 0.1 | 0.0 | 0.1 | 0.1 | 0.0 | 0.1 | 0.0 | -0.1 |
| | σ[X] | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.2 | 0.1 | 0.1 | 0.0 | 0.1 | 0.0 | 0.1 | 0.1 | 0.2 | 0.0 | 0.2 | 0.1 | 0.0 | 0.2 |
| TH_MC2 | E[X] | 0.6 | -0.2 | 0.0 | -0.1 | 0.3 | 0.0 | 0.0 | -6.9 | 0.1 | 0.0 | 0.4 | -0.1 | 0.0 | 0.2 | 0.0 | -0.2 | -0.2 | -0.3 | 0.1 | 0.0 | 0.2 | 0.0 |
| | σ[X] | 0.0 | 0.2 | 0.1 | 0.1 | 0.1 | 0.1 | 0.2 | 0.2 | 0.2 | 0.1 | 0.1 | 0.2 | 0.3 | 0.0 | 0.0 | 0.2 | 0.2 | 0.1 | 0.1 | 0.1 | 0.4 | 0.3 |
| TH_Sorter1_1 | E[X] | 47.0 | -0.2 | -0.5 | -6.6 | 6.8 | -2.2 | 34.4 | -2.5 | -0.5 | -0.3 | 0.0 | -0.1 | -0.1 | 1.2 | -1.3 | -0.3 | 5.1 | 0.2 | 0.4 | 0.2 | 0.2 | -0.3 |
| | σ[X] | 0.5 | 0.5 | 0.1 | 0.4 | 0.4 | 0.2 | 0.3 | 0.5 | 0.2 | 0.4 | 0.5 | 1.2 | 0.5 | 0.6 | 0.4 | 0.2 | 0.5 | 0.1 | 0.9 | 0.2 | 0.9 | 0.8 |
| TH_Sorter1_2 | E[X] | -77.0 | -0.6 | -0.1 | -7.2 | -9.1 | -0.2 | 57.6 | -2.8 | 0.0 | -0.4 | 0.1 | 0.3 | 0.4 | 0.8 | 1.2 | 0.5 | -7.2 | 0.7 | 0.1 | 0.1 | -0.3 | 0.3 |
| | σ[X] | 0.7 | 0.2 | 0.2 | 0.6 | 0.2 | 1.0 | 0.5 | 0.5 | 0.4 | 0.2 | 0.3 | 0.3 | 0.5 | 0.8 | 0.5 | 0.2 | 0.7 | 0.2 | 0.4 | 0.6 | 0.5 | 0.3 |
| TH_Sorter2_1 | E[X] | 22.9 | 0.3 | 1.4 | -3.1 | 7.6 | -2.5 | 35.5 | -3.5 | 1.4 | 1.0 | 0.5 | 0.7 | 2.5 | 5.3 | -1.8 | -0.1 | 0.3 | -0.4 | 1.8 | 0.5 | -0.4 | -1.1 |
| | σ[X] | 3.3 | 2.8 | 2.5 | 3.4 | 4.7 | 1.3 | 2.1 | 3.2 | 0.7 | 2.6 | 2.5 | 2.7 | 1.5 | 2.7 | 3.5 | 2.0 | 3.3 | 3.0 | 0.4 | 2.5 | 2.6 | 4.1 |
| TH_Sorter2_2 | E[X] | 27.1 | 1.7 | 3.1 | -0.1 | -13.0 | -0.4 | -3.0 | -3.6 | 2.3 | 0.6 | 0.5 | 0.9 | 3.6 | 7.8 | -3.1 | 0.4 | -2.5 | -0.4 | 2.2 | 0.8 | 0.6 | -1.2 |
| | σ[X] | 4.1 | 3.6 | 2.5 | 3.6 | 5.6 | 1.9 | 3.1 | 5.3 | 0.6 | 2.7 | 2.8 | 3.5 | 3.9 | 4.2 | 6.0 | 2.7 | 3.3 | 5.2 | 0.7 | 2.7 | 3.6 | 4.0 |
| TH_System | E[X] | 48.5 | 2.4 | 10.7 | 20.4 | -7.8 | -0.8 | -7.7 | -1.8 | 6.9 | 0.9 | 1.0 | 2.6 | 10.2 | 22.4 | -7.8 | -0.8 | -8.2 | -1.9 | 6.8 | 2.4 | 1.1 | -4.1 |
| | σ[X] | 12.1 | 12.4 | 8.7 | 12.2 | 17.9 | 7.6 | 10.7 | 13.8 | 1.0 | 8.6 | 10.2 | 12.1 | 8.8 | 12.1 | 17.7 | 7.6 | 10.6 | 13.7 | 1.0 | 8.4 | 10.2 | 14.1 |
| WIP | E[X] | 21.6 | -3.3 | 5.0 | -43.1 | 2.0 | -0.4 | 8.3 | -1.0 | -1.2 | -21.6 | 0.2 | -1.9 | -1.7 | -11.1 | 1.0 | 1.3 | 2.7 | -0.6 | -1.4 | -4.3 | -0.4 | 1.4 |
| | σ[X] | 1.0 | 0.7 | 1.5 | 0.8 | 2.9 | 1.1 | 3.3 | 1.3 | 0.2 | 1.8 | 1.7 | 1.4 | 0.6 | 1.2 | 3.3 | 0.9 | 3.2 | 1.9 | 0.9 | 1.0 | 1.5 | 1.0 |

**Table 44: The average and standard deviation of factor effects on different responses (part 1)**

| | | Int. WindowLength - EBS_r | Int. WindowLength - L1_fr | Int. WindowLength - L3_fr | Int. WindowLength - Lateral Assignment | Int. WindowLength - MC_r | Int. WindowLength - ST | Int. WindowLength - Velocity | Int. WindowLength - EBS_to_MC | Int. EBS_Control - EBS_r | Int. EBS_Control - L1_fr | Int. EBS_Control - L3_fr | Int. EBS_Control - Lateral Assignment | Int. EBS_Control - MC_r | Int. EBS_Control - ST | Int. EBS_Control - Velocity | Int. EBS_Control - EBS_to_MC | Int. EBS_r - L1_fr | Int. EBS_r - L3_fr | Int. EBS_r - Lateral Assignment | Int. EBS_r - MC_r | Int. EBS_r - ST | Int. EBS_r - Velocity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CT_EBS | E[X] | 16.4 | 5.1 | 14.8 | 3.9 | -27.5 | 10.9 | -30.5 | 0.3 | -51.6 | -10.2 | 3.8 | -21.8 | 6.6 | 1.9 | -25.1 | 12.0 | -20.9 | -5.5 | -30.5 | -8.2 | 1.6 | 101.7 |
| | σ[X] | 15.9 | 36.6 | 20.1 | 40.1 | 32.6 | 2.3 | 34.6 | 22.1 | 29.7 | 13.9 | 8.2 | 33.2 | 10.5 | 34.8 | 7.3 | 10.3 | 24.8 | 12.5 | 51.8 | 22.2 | 9.8 | 19.2 |
| CT_Inspection | E[X] | 2.0 | 0.6 | 1.1 | 0.3 | -1.5 | 1.1 | 0.8 | 1.3 | 3.1 | -1.5 | 3.4 | -1.8 | -0.5 | 0.1 | 2.0 | 2.9 | -2.0 | 2.8 | -0.1 | 1.7 | 0.0 | 4.0 |
| | σ[X] | 1.3 | 3.5 | 1.9 | 5.7 | 3.8 | 4.1 | 2.2 | 0.3 | 2.5 | 3.5 | 4.4 | 3.3 | 5.0 | 2.9 | 3.2 | 1.6 | 4.6 | 2.9 | 3.6 | 4.6 | 1.9 | 4.2 |
| CT_Standard | E[X] | 1.2 | 0.6 | 0.1 | -1.6 | -1.1 | 0.8 | 2.5 | 0.5 | 1.4 | -0.5 | 1.3 | -1.1 | -1.4 | 0.1 | 0.3 | 0.8 | -0.6 | 0.3 | -1.0 | 0.4 | 0.5 | 1.8 |
| | σ[X] | 0.4 | 0.8 | 1.1 | 1.2 | 1.1 | 1.4 | 1.1 | 0.4 | 0.6 | 1.2 | 0.9 | 1.2 | 1.2 | 0.7 | 0.9 | 0.5 | 1.0 | 0.8 | 1.2 | 0.8 | 0.2 | 0.9 |
| CT_Total | E[X] | 2.5 | -4.1 | 6.2 | -3.2 | -10.3 | 2.1 | 1.9 | -4.4 | -1.0 | -1.7 | 2.6 | -10.1 | -2.7 | 4.7 | -2.7 | 3.2 | -4.4 | -0.2 | -5.9 | -0.5 | 4.1 | 9.1 |
| | σ[X] | 4.1 | 6.3 | 3.3 | 8.6 | 8.0 | 6.9 | 6.6 | 1.1 | 3.9 | 7.0 | 4.0 | 8.7 | 7.7 | 6.0 | 3.8 | 4.1 | 9.3 | 4.0 | 8.2 | 4.3 | 1.5 | 5.4 |
| TH_Induct1_1 | E[X] | 0.0 | 0.2 | -0.3 | 0.4 | -0.1 | -0.1 | 0.1 | -0.1 | -0.1 | 0.3 | 0.0 | 0.0 | 0.3 | -0.1 | 0.0 | 0.2 | -0.2 | 0.0 | 0.2 | 0.0 | 0.2 | -0.1 |
| | σ[X] | 0.2 | 0.3 | 0.1 | 0.0 | 0.3 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.4 | 0.0 | 0.1 | 0.1 | 0.1 | 0.5 | 0.2 | 0.2 | 0.2 | 0.3 | 0.2 |
| TH_Induct1_2 | E[X] | -0.1 | 0.0 | 0.1 | -0.2 | 0.0 | 0.2 | 0.0 | -0.1 | -0.1 | 0.1 | 0.0 | 0.1 | 0.2 | -0.1 | 0.2 | 0.2 | -0.2 | -0.3 | 0.1 | 0.1 | 0.0 | 0.0 |
| | σ[X] | 0.1 | 0.1 | 0.2 | 0.2 | 0.1 | 0.1 | 0.1 | 0.2 | 0.3 | 0.4 | 0.1 | 0.0 | 0.2 | 0.1 | 0.1 | 0.1 | 0.1 | 0.3 | 0.2 | 0.1 | 0.1 | 0.1 |
| TH_Induct1_3 | E[X] | 0.1 | -0.1 | -0.1 | 0.0 | 0.1 | 0.2 | 0.1 | 0.0 | 0.3 | 0.0 | -0.1 | -0.1 | 0.1 | 0.1 | 0.0 | 0.2 | 0.3 | 0.0 | 0.2 | -0.2 | 0.2 | -0.1 |
| | σ[X] | 0.4 | 0.0 | 0.3 | 0.2 | 0.0 | 0.1 | 0.1 | 0.1 | 0.2 | 0.3 | 0.2 | 0.2 | 0.3 | 0.4 | 0.3 | 0.3 | 0.4 | 0.2 | 0.1 | 0.1 | 0.3 | 0.5 |
| TH_Induct1_41 | E[X] | 0.0 | -0.1 | 0.2 | -0.2 | -0.1 | -0.2 | -0.2 | 0.2 | 0.0 | -0.4 | 0.1 | 0.1 | -0.5 | 0.0 | -0.2 | -0.5 | 0.2 | 0.2 | -0.5 | 0.2 | -0.3 | 0.1 |
| | σ[X] | 0.4 | 0.3 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 | 0.1 | 0.6 | 0.1 | 0.3 | 0.5 | 0.3 | 0.2 | 0.2 | 0.8 | 0.2 | 0.4 | 0.1 | 0.2 | 0.7 |
| TH_Induct1_42 | E[X] | 0.1 | -0.1 | 0.0 | 0.2 | 0.2 | -0.2 | -0.4 | 0.0 | 8.6 | -0.1 | 0.0 | 0.1 | -0.5 | -0.2 | -0.4 | -0.1 | 0.2 | 0.1 | 0.2 | -0.4 | -0.1 | 0.4 |
| | σ[X] | 0.4 | 0.3 | 0.5 | 0.4 | 0.2 | 1.1 | 0.0 | 0.3 | 0.2 | 0.7 | 1.1 | 0.2 | 0.2 | 0.2 | 0.4 | 0.1 | 1.0 | 0.2 | 0.5 | 0.2 | 0.8 | 1.0 |
| TH_Induct2_1 | E[X] | -0.8 | 3.8 | -3.5 | 1.4 | 5.2 | -1.0 | -0.5 | 3.4 | -2.8 | 1.0 | -0.5 | 4.9 | 1.4 | -4.0 | 0.6 | -1.9 | 2.5 | 0.4 | 2.0 | 0.4 | -2.1 | -0.4 |
| | σ[X] | 4.2 | 2.4 | 2.4 | 5.5 | 4.2 | 3.3 | 3.4 | 1.9 | 2.2 | 3.5 | 2.7 | 4.7 | 5.4 | 2.5 | 2.6 | 1.9 | 5.3 | 2.6 | 2.7 | 4.2 | 0.4 | 2.9 |
| TH_Induct2_2 | E[X] | -1.4 | 3.3 | -3.7 | 1.6 | 6.9 | -1.0 | -1.0 | 3.5 | -4.1 | 1.1 | -2.0 | 7.2 | 1.6 | -3.3 | -0.5 | -2.1 | 2.5 | -0.4 | 3.5 | 0.2 | -2.7 | -2.0 |
| | σ[X] | 3.0 | 3.1 | 2.6 | 4.5 | 5.0 | 5.6 | 4.1 | 1.3 | 3.8 | 5.4 | 2.6 | 4.8 | 4.4 | 3.2 | 2.4 | 2.0 | 6.3 | 2.6 | 4.5 | 3.6 | 0.4 | 2.5 |
| TH Induct2 3 | E[X] | -0.3 | 3.6 | -3.4 | 1.3 | 5.2 | -0.8 | -0.1 | 3.7 | -2.6 | 0.6 | 0.0 | 5.5 | 1.7 | -3.7 | 0.6 | -1.4 | 1.7 | 0.7 | 2.2 | 0.4 | -2.0 | -0.1 |
| | σ[X] | 4.2 | 2.9 | 3.1 | 5.6 | 4.1 | 3.6 | 2.9 | 1.5 | 2.4 | 3.8 | 3.2 | 4.2 | 5.9 | 2.7 | 2.4 | 2.8 | 5.4 | 2.3 | 2.8 | 4.8 | 0.3 | 2.9 |
| TH_Induct2_41 | E[X] | -0.2 | 0.6 | -0.5 | 1.0 | 1.2 | -0.4 | 0.1 | 0.7 | -0.8 | 0.5 | -0.3 | 1.0 | 0.6 | -0.4 | -0.1 | -0.3 | 1.2 | -0.2 | 0.5 | 0.5 | 0.0 | -0.3 |
| | σ[X] | 0.9 | 0.7 | 0.7 | 0.9 | 0.8 | 1.0 | 0.8 | 0.6 | 0.5 | 0.7 | 0.9 | 0.6 | 1.0 | 0.9 | 0.4 | 0.9 | 1.2 | 0.4 | 1.0 | 1.6 | 0.2 | 0.7 |
| TH_Induct2_42 | E[X] | -0.4 | 1.2 | -0.6 | 0.2 | 1.3 | 0.1 | -0.8 | 0.9 | 7.6 | 0.4 | -0.8 | 1.6 | 1.1 | -1.0 | -0.8 | -0.4 | 0.0 | 0.1 | 1.0 | -0.4 | -0.6 | -1.5 |
| | σ[X] | 1.4 | 0.8 | 0.5 | 1.8 | 0.9 | 1.5 | 0.7 | 0.6 | 0.6 | 1.5 | 0.7 | 1.6 | 1.7 | 1.1 | 0.3 | 1.0 | 0.9 | 0.7 | 0.7 | 1.2 | 0.6 | 1.1 |
| TH_Inpection | E[X] | 0.1 | 0.0 | 0.0 | 0.0 | 0.1 | -0.1 | -0.1 | 0.1 | -0.1 | 0.1 | 0.0 | 0.0 | -0.1 | 0.1 | 0.1 | -0.1 | 0.0 | 0.2 | 0.0 | 0.1 | -0.1 | 0.1 |
| | σ[X] | 0.1 | 0.1 | 0.2 | 0.1 | 0.2 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.2 | 0.1 | 0.2 | 0.0 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| TH_MC1 | E[X] | -0.2 | 0.1 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | -0.3 | 0.1 | 0.0 | 0.0 | 0.2 | -0.1 | -0.1 | 0.0 | 0.1 | 0.1 | 0.1 | 0.0 |
| | σ[X] | 0.2 | 0.2 | 0.2 | 0.1 | 0.1 | 0.1 | 0.2 | 0.1 | 0.2 | 0.2 | 0.0 | 0.1 | 0.1 | 0.2 | 0.0 | 0.2 | 0.0 | 0.0 | 0.3 | 0.1 | 0.1 | 0.1 |
| TH_MC2 | E[X] | 0.0 | -0.1 | -0.1 | 0.0 | 0.2 | 0.1 | 0.0 | 0.1 | 0.0 | 0.1 | 0.0 | 0.3 | 0.0 | -0.1 | 0.0 | -0.3 | 0.1 | 0.0 | 0.1 | 0.1 | 0.0 | -0.2 |
| | σ[X] | 0.0 | 0.1 | 0.1 | 0.2 | 0.1 | 0.2 | 0.1 | 0.0 | 0.1 | 0.0 | 0.0 | 0.1 | 0.2 | 0.1 | 0.1 | 0.1 | 0.1 | 0.2 | 0.2 | 0.3 | 0.0 | 0.2 |
| TH_Sorter1_1 | E[X] | 0.0 | -0.2 | 0.4 | 0.1 | -0.6 | -0.1 | 0.5 | 0.2 | -0.4 | -0.3 | -0.4 | 0.0 | -0.6 | 0.5 | -0.8 | 0.0 | 0.1 | -0.3 | 1.1 | -0.2 | -0.1 | 0.2 |
| | σ[X] | 0.9 | 0.5 | 0.2 | 0.1 | 0.4 | 0.3 | 1.3 | 0.8 | 0.2 | 0.3 | 0.3 | 0.4 | 0.5 | 0.6 | 0.5 | 0.6 | 0.9 | 0.7 | 0.7 | 0.2 | 0.1 | 0.6 |
| TH_Sorter1_2 | E[X] | 0.2 | -0.1 | 0.0 | 0.7 | 0.6 | 0.8 | 0.4 | -0.6 | 0.0 | -0.2 | -0.8 | -0.2 | 0.2 | 0.1 | -0.3 | 0.4 | -0.2 | -0.1 | 1.3 | 0.1 | -0.1 | -0.2 |
| | σ[X] | 0.8 | 0.4 | 0.6 | 0.2 | 0.7 | 0.7 | 1.0 | 0.9 | 0.5 | 0.4 | 0.8 | 0.6 | 0.2 | 1.3 | 0.6 | 0.5 | 1.0 | 0.7 | 0.5 | 0.2 | 0.1 | 1.0 |
| TH_Sorter2_1 | E[X] | -0.5 | 3.0 | -2.1 | 1.7 | 3.8 | -0.2 | -0.4 | 2.3 | -2.6 | 0.8 | -1.0 | 4.0 | 1.2 | -2.7 | 0.2 | -1.5 | 1.8 | 0.7 | 2.7 | 0.8 | -1.6 | -0.9 |
| | σ[X] | 3.3 | 1.7 | 2.5 | 3.6 | 2.5 | 3.7 | 2.2 | 1.2 | 2.2 | 2.6 | 2.3 | 3.5 | 4.2 | 2.5 | 2.2 | 1.8 | 4.3 | 1.1 | 1.9 | 4.4 | 1.0 | 1.9 |
| TH_Sorter2_2 | E[X] | -0.7 | 3.8 | -3.5 | 1.0 | 6.5 | -0.8 | -1.2 | 4.1 | -3.6 | 0.8 | -0.8 | 6.5 | 1.0 | -3.9 | 0.1 | -2.4 | 2.7 | -0.9 | 2.5 | 0.5 | -2.6 | -0.8 |
| | σ[X] | 3.9 | 2.6 | 1.7 | 5.0 | 5.4 | 4.2 | 4.5 | 2.0 | 3.4 | 4.3 | 3.4 | 4.8 | 5.7 | 2.7 | 2.7 | 1.2 | 6.0 | 2.6 | 2.9 | 4.6 | 0.5 | 3.0 |
| TH_System | E[X] | -2.6 | 11.3 | -11.0 | 5.0 | 18.3 | -3.2 | -1.6 | 11.0 | -10.2 | 3.2 | -2.8 | 18.3 | 5.1 | -11.3 | 0.9 | -5.5 | 7.8 | 0.7 | 8.2 | 1.6 | -6.8 | -2.5 |
| | σ[X] | 12.2 | 9.1 | 8.2 | 16.1 | 13.7 | 13.2 | 11.0 | 4.9 | 8.7 | 13.1 | 8.4 | 13.7 | 16.2 | 9.0 | 7.6 | 7.7 | 17.9 | 7.6 | 10.6 | 13.9 | 0.8 | 8.5 |
| WIP | E[X] | 1.2 | -0.6 | 2.0 | -1.4 | -3.6 | 0.9 | 0.6 | -1.0 | -2.2 | -0.7 | 1.4 | -3.5 | -1.0 | 1.0 | -1.5 | 1.3 | -1.8 | 0.0 | -2.2 | -0.1 | 1.2 | 5.3 |
| | σ[X] | 0.7 | 2.6 | 1.2 | 2.9 | 3.0 | 2.0 | 2.6 | 0.5 | 1.4 | 2.1 | 0.9 | 3.4 | 2.2 | 2.4 | 1.0 | 1.0 | 2.9 | 1.1 | 3.7 | 0.6 | 0.6 | 1.8 |

**Table 45: The average and standard deviation of factor effects on different responses (part 2)**

| | | Int. EBS_r - EBS_to_MC | Int. L1_fr - L3_fr | Int. L1_fr - Lateral Assignment | Int. L1_fr - MC_r | Int. L1_fr - ST | Int. L1_fr - Velocity | Int. L1_fr - EBS_to_MC | Int. L3_fr - Lateral Assignment | Int. L3_fr - MC_r | Int. L3_fr - ST | Int. L3_fr - Velocity | Int. L3_fr - EBS_to_MC | Int. Lateral Assignment - MC_r | Int. Lateral Assignment - ST | Int. Lateral Assignment - Velocity | Int. Lateral Assignment - EBS_to_MC | Int. MC_r - ST | Int. MC_r - Velocity | Interaction MC_r - EBS_to_MC | Int. ST - Velocity | Int. ST - EBS_to_MC | Int. Velocity - EBS_to_MC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CT_EBS | E[X] | -15.2 | 2.8 | 6.9 | -0.8 | -12.6 | -3.5 | -2.2 | 9.3 | -9.3 | 2.5 | 9.2 | 9.2 | -1.6 | 3.8 | -25.6 | -0.7 | -13.7 | -4.8 | -3.6 | 1.5 | 10.6 | -12.5 |
| | σ[X] | 5.1 | 35.7 | 10.1 | 46.5 | 7.2 | 15.8 | 12.2 | 20.4 | 12.3 | 8.4 | 27.0 | 57.1 | 45.1 | 28.8 | 35.4 | 14.4 | 34.0 | 8.5 | 33.0 | 27.4 | 10.9 | 29.2 |
| CT_Inspection | E[X] | 2.0 | -1.6 | -1.6 | -2.4 | -2.4 | -1.8 | -2.1 | 0.4 | 1.3 | 0.1 | 3.0 | 1.3 | 2.1 | 2.6 | -0.1 | -1.3 | 1.5 | 1.7 | 1.1 | 0.0 | 0.5 | 1.6 |
| | σ[X] | 3.0 | 4.7 | 2.3 | 4.0 | 0.8 | 3.8 | 4.8 | 2.8 | 4.4 | 1.5 | 1.7 | 3.2 | 1.6 | 3.4 | 1.2 | 3.3 | 2.5 | 3.5 | 4.8 | 2.0 | 2.0 | 2.1 |
| CT_Standard | E[X] | 0.5 | -1.0 | -0.7 | -0.2 | -0.6 | 0.3 | 1.1 | -0.3 | -0.2 | -0.4 | 1.4 | 0.0 | 0.7 | -0.1 | -2.9 | -0.5 | 0.1 | 0.8 | 0.7 | 0.5 | -0.4 | 0.2 |
| | σ[X] | 0.6 | 1.6 | 1.2 | 1.2 | 0.5 | 1.5 | 0.9 | 0.7 | 1.0 | 1.0 | 0.7 | 1.3 | 0.4 | 1.2 | 0.4 | 1.0 | 1.4 | 0.8 | 1.6 | 0.6 | 0.6 | 0.5 |
| CT_Total | E[X] | -0.6 | -5.0 | -0.3 | 4.9 | -2.8 | -2.7 | 6.3 | 0.2 | -5.4 | -1.2 | 5.4 | -4.8 | 1.7 | -4.9 | -4.7 | -1.6 | -0.9 | 0.6 | 4.8 | 1.4 | -0.5 | -4.2 |
| | σ[X] | 5.8 | 3.6 | 7.1 | 9.1 | 5.7 | 6.3 | 4.0 | 5.0 | 3.7 | 5.8 | 2.9 | 9.2 | 5.9 | 7.5 | 4.5 | 5.4 | 8.7 | 5.7 | 3.5 | 4.1 | 4.7 | 2.7 |
| TH_Induct1_1 | E[X] | 0.1 | 0.0 | 0.1 | -0.1 | 0.0 | -0.1 | 0.0 | 0.2 | 0.1 | 0.0 | 0.1 | 0.0 | -0.1 | 0.1 | -0.1 | 0.1 | 0.0 | 0.0 | 0.2 | 0.1 | -0.1 | -0.2 |
| | σ[X] | 0.1 | 0.3 | 0.1 | 0.3 | 0.1 | 0.1 | 0.3 | 0.6 | 0.1 | 0.2 | 0.3 | 0.2 | 0.3 | 0.3 | 0.0 | 0.4 | 0.1 | 0.2 | 0.1 | 0.4 | 0.5 | 0.3 |
| TH_Induct1_2 | E[X] | -0.1 | 0.1 | 0.2 | 0.1 | 0.0 | -0.1 | 0.0 | -0.1 | -0.3 | 0.1 | -0.1 | -0.1 | 0.2 | 0.1 | 0.0 | 0.0 | -0.1 | -0.1 | -0.1 | 0.0 | 0.3 | 0.0 |
| | σ[X] | 0.1 | 0.1 | 0.3 | 0.3 | 0.3 | 0.2 | 0.1 | 0.2 | 0.1 | 0.4 | 0.0 | 0.3 | 0.1 | 0.3 | 0.1 | 0.1 | 0.1 | 0.1 | 0.3 | 0.1 | 0.3 | 0.1 |
| TH_Induct1_3 | E[X] | -0.1 | 0.0 | 0.1 | 0.0 | 0.2 | 0.3 | 0.2 | 0.0 | 0.0 | 0.2 | 0.0 | -0.1 | 0.0 | 0.2 | -0.1 | 0.1 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | σ[X] | 0.2 | 0.1 | 0.4 | 0.4 | 0.1 | 0.2 | 0.3 | 0.3 | 0.3 | 0.1 | 0.0 | 0.3 | 0.3 | 0.4 | 0.0 | 0.1 | 0.2 | 0.1 | 0.1 | 0.3 | 0.2 | 0.1 |
| TH_Induct1_41 | E[X] | 0.0 | -0.9 | -0.2 | -0.1 | -0.2 | -0.1 | -0.1 | 0.0 | 0.1 | -0.3 | 0.0 | 0.4 | -0.1 | -0.4 | 0.2 | -0.2 | -0.1 | 0.2 | -0.2 | -0.1 | -0.2 | 0.1 |
| | σ[X] | 0.1 | 0.3 | 0.2 | 0.7 | 0.4 | 0.4 | 0.5 | 0.6 | 0.1 | 0.5 | 0.3 | 0.2 | 0.7 | 0.4 | 0.0 | 0.3 | 0.2 | 0.2 | 0.3 | 0.5 | 0.5 | 0.3 |
| TH_Induct1_42 | E[X] | -0.8 | 0.1 | 0.1 | 1.0 | 0.2 | 0.3 | -0.1 | 0.4 | 0.0 | -0.2 | -0.2 | -0.2 | -1.0 | -0.1 | 0.3 | 0.1 | -0.1 | -0.2 | 0.3 | -0.1 | 0.1 | -0.7 |
| | σ[X] | 0.7 | 0.4 | 0.6 | 0.3 | 0.4 | 0.7 | 0.4 | 0.4 | 0.4 | 0.4 | 0.1 | 0.3 | 0.5 | 0.3 | 0.5 | 0.2 | 0.4 | 1.0 | 0.5 | 0.3 | 1.1 | 0.2 |
| TH_Induct2_1 | E[X] | -0.4 | 3.0 | -0.6 | -3.5 | 1.0 | 2.0 | -2.7 | 0.3 | 2.7 | 0.5 | -2.9 | 3.4 | -1.0 | 3.5 | -0.4 | 0.5 | 0.6 | -0.8 | -3.1 | -0.4 | -0.2 | 2.5 |
| | σ[X] | 3.3 | 1.5 | 3.0 | 4.1 | 5.0 | 3.7 | 1.7 | 2.7 | 1.6 | 3.1 | 2.1 | 4.1 | 4.9 | 4.1 | 2.0 | 3.2 | 2.9 | 3.4 | 1.8 | 4.0 | 2.7 | 1.3 |
| TH_Induct2_2 | E[X] | -0.9 | 3.8 | 0.6 | -2.6 | 1.7 | 2.0 | -2.8 | 0.6 | 2.7 | 0.2 | -4.5 | 2.9 | -2.0 | 2.8 | 0.9 | 0.4 | -0.5 | -1.0 | -3.7 | -0.6 | 0.7 | 2.0 |
| | σ[X] | 2.7 | 1.7 | 4.7 | 5.1 | 3.6 | 4.1 | 2.2 | 2.0 | 2.3 | 3.3 | 3.9 | 5.1 | 3.3 | 5.3 | 2.0 | 3.9 | 4.8 | 4.1 | 1.5 | 3.7 | 1.8 | 1.8 |
| TH Induct2 3 | E[X] | 0.0 | 3.3 | -0.6 | -3.9 | 0.8 | 1.9 | -3.3 | 0.5 | 3.1 | 0.5 | -2.8 | 3.7 | -0.5 | 4.1 | 0.1 | 0.5 | 0.5 | -0.3 | -3.0 | -0.2 | 0.1 | 2.4 |
| | σ[X] | 3.2 | 2.2 | 3.3 | 3.9 | 4.7 | 3.4 | 2.1 | 3.1 | 2.4 | 3.1 | 2.3 | 4.0 | 4.9 | 3.6 | 2.7 | 3.1 | 3.2 | 3.7 | 1.9 | 4.1 | 3.3 | 0.7 |
| TH_Induct2_41 | E[X] | 0.2 | -0.2 | -0.1 | -0.6 | 0.6 | 0.6 | -0.4 | 0.0 | 0.5 | 0.6 | -0.1 | 0.6 | -0.7 | 0.4 | -0.2 | 0.4 | 0.0 | -0.4 | -1.1 | -0.4 | 0.6 | 0.6 |
| | σ[X] | 0.9 | 0.3 | 0.7 | 0.9 | 0.9 | 1.0 | 1.1 | 0.3 | 0.7 | 1.2 | 0.8 | 0.8 | 1.0 | 0.8 | 0.6 | 0.8 | 0.5 | 0.6 | 0.7 | 0.7 | 0.5 | 0.6 |
| TH_Induct2_42 | E[X] | -0.1 | 1.1 | -0.2 | -1.3 | 0.0 | 1.1 | -0.4 | 0.5 | 0.8 | 0.4 | -1.2 | 1.0 | -0.5 | 0.9 | -0.1 | -0.2 | 0.4 | 0.0 | -0.8 | 0.0 | -0.5 | 1.0 |
| | σ[X] | 1.2 | 0.4 | 1.6 | 1.4 | 1.2 | 1.5 | 0.3 | 0.8 | 0.4 | 1.0 | 0.9 | 1.0 | 0.9 | 1.4 | 1.3 | 1.2 | 1.5 | 1.1 | 0.7 | 0.9 | 1.3 | 1.3 |
| TH_Inpection | E[X] | -0.1 | 1.7 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | -0.1 | 0.0 | 0.1 | 0.0 | 0.1 | 0.0 | 0.0 | -0.1 | 0.1 | 0.0 | -0.2 | 0.0 | 0.0 | 0.1 | 0.1 |
| | σ[X] | 0.1 | 0.1 | 0.2 | 0.1 | 0.0 | 0.1 | 0.1 | 0.1 | 0.1 | 0.0 | 0.1 | 0.0 | 0.1 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.0 | 0.2 |
| TH_MC1 | E[X] | -0.1 | 0.1 | -0.1 | 0.0 | 0.0 | -0.1 | -0.1 | 0.1 | -0.1 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | -0.1 | -0.3 | 0.0 | 0.0 | 0.1 |
| | σ[X] | 0.2 | 0.2 | 0.1 | 0.1 | 0.1 | 0.2 | 0.1 | 0.2 | 0.1 | 0.1 | 0.1 | 0.2 | 0.3 | 0.1 | 0.0 | 0.1 | 0.1 | 0.2 | 0.2 | 0.1 | 0.2 | 0.1 |
| TH_MC2 | E[X] | -0.2 | 0.1 | -0.1 | -0.2 | 0.1 | 0.2 | 0.0 | 0.1 | 0.0 | -0.2 | 0.0 | 0.0 | -0.1 | 0.1 | -0.2 | 0.0 | 0.0 | -0.1 | -0.2 | 0.0 | 0.0 | 0.0 |
| | σ[X] | 0.1 | 0.1 | 0.2 | 0.1 | 0.1 | 0.1 | 0.2 | 0.1 | 0.2 | 0.1 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.1 | 0.2 | 0.3 | 0.2 | 0.2 |
| TH_Sorter1_1 | E[X] | 0.1 | -0.7 | -1.0 | 0.3 | -0.1 | 0.6 | -0.3 | -0.2 | 0.2 | 0.1 | 0.0 | -0.1 | -0.7 | -0.3 | -0.3 | 0.2 | 0.5 | 0.4 | -0.4 | -0.6 | 0.2 | 0.4 |
| | σ[X] | 0.6 | 0.0 | 0.6 | 0.4 | 1.4 | 0.4 | 0.5 | 1.0 | 0.5 | 0.2 | 0.1 | 0.6 | 0.6 | 0.9 | 0.8 | 0.4 | 0.6 | 0.3 | 0.7 | 0.2 | 0.5 | 0.5 |
| TH_Sorter1_2 | E[X] | -0.3 | -0.3 | -0.6 | -0.1 | 0.4 | -0.3 | 0.0 | 0.5 | -0.1 | -0.1 | -0.7 | -0.1 | -0.4 | -0.5 | -0.6 | 0.2 | 0.4 | 0.0 | -0.3 | 0.2 | -0.2 | -0.6 |
| | σ[X] | 0.5 | 0.6 | 1.2 | 1.0 | 0.3 | 0.9 | 0.3 | 0.9 | 0.3 | 0.8 | 0.9 | 0.4 | 0.2 | 0.5 | 1.0 | 0.5 | 0.6 | 0.8 | 0.7 | 0.6 | 0.8 | 0.1 |
| TH_Sorter2_1 | E[X] | -0.5 | 2.4 | -0.7 | -3.2 | 1.4 | 1.7 | -1.9 | 0.1 | 1.7 | 0.9 | -2.2 | 2.9 | -0.8 | 2.7 | 0.0 | 0.1 | 0.3 | -0.8 | -3.3 | -0.5 | 0.2 | 2.2 |
| | σ[X] | 2.8 | 1.4 | 2.9 | 1.8 | 3.9 | 2.8 | 2.4 | 2.3 | 1.6 | 2.3 | 2.2 | 2.8 | 4.1 | 2.4 | 2.1 | 2.5 | 2.0 | 2.5 | 1.4 | 2.5 | 1.7 | 0.8 |
| TH_Sorter2_2 | E[X] | -1.0 | 3.1 | -0.1 | -3.8 | 1.1 | 2.4 | -2.6 | 1.2 | 2.5 | 0.5 | -3.9 | 3.2 | -1.4 | 3.3 | 0.1 | 0.6 | -0.3 | -0.7 | -3.5 | -1.1 | 0.0 | 2.9 |
| | σ[X] | 2.9 | 2.1 | 3.9 | 4.3 | 3.9 | 3.8 | 1.8 | 3.1 | 2.0 | 2.6 | 2.8 | 5.3 | 5.2 | 5.9 | 2.7 | 4.3 | 4.1 | 4.3 | 1.6 | 4.8 | 2.9 | 1.9 |
| TH_System | E[X] | -1.0 | 10.7 | -0.7 | -10.9 | 4.1 | 6.3 | -9.1 | 1.4 | 9.0 | 1.9 | -10.0 | 10.8 | -4.1 | 10.8 | 0.2 | 2.0 | 0.5 | -2.4 | -10.7 | -1.7 | 1.4 | 7.3 |
| | σ[X] | 10.3 | 4.5 | 11.2 | 13.7 | 14.1 | 11.9 | 6.3 | 7.6 | 6.3 | 10.6 | 8.9 | 13.5 | 13.8 | 13.7 | 7.3 | 10.7 | 11.3 | 11.6 | 4.6 | 12.2 | 7.7 | 4.3 |
| WIP | E[X] | -0.5 | -1.4 | -0.3 | 1.2 | -1.2 | -0.5 | 2.3 | 0.3 | -1.7 | -0.4 | 1.7 | -1.1 | 0.5 | -1.2 | -3.0 | -0.7 | -0.4 | 0.5 | 1.2 | 0.8 | -0.2 | -1.2 |
| | σ[X] | 1.7 | 2.0 | 2.4 | 3.7 | 1.0 | 2.4 | 1.5 | 1.6 | 1.3 | 1.8 | 0.6 | 3.9 | 1.8 | 2.8 | 1.8 | 1.6 | 3.3 | 2.0 | 1.9 | 0.5 | 1.3 | 1.5 |

**Table 46: The average and standard deviation of factor effects on different responses (part**