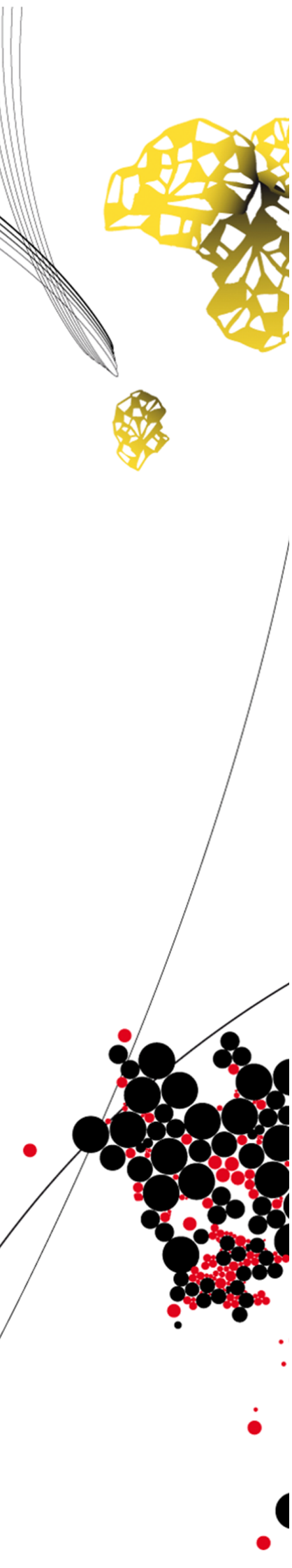


UNIVERSITY OF TWENTE.

Faculty of Behavioural, Management and Social sciences
Department of Industrial Engineering and
Business Information Systems

An optimization approach between service level and inventory via simulation: an example from the semiconductor industry

S.E. Lingelbach

M.Sc. Thesis
Munich, May 2017

Supervisors:
Dr. ir. A. Al Hanbali (University of Twente)
Dr. M.C. van der Heijden (University of Twente)
Dr. F. Federmann (Infineon Technologies AG)

University of Twente
in cooperation with
Infineon Technologies AG
am Campeon 1-12
85579 Neubiberg
Germany

Management summary

Company & Motivation
This graduation project is conducted as part of the Industrial Engineering and Management master program in cooperation with Infineon Technologies AG. Infineon is a German semiconductor manufacturer producing chips, sensors, and microcontrollers. To stay competitive and satisfy customer demand quickly Infineon places inventory at various stock points within their supply chain. However, the more products are stored, the higher the costs due to the binding capital effect of stock. Thus, Infineon has to balance the trade off between high stocks (characterized by a high α-service level) and high costs when examining its supply chain planning processes. In this thesis, we concentrate on the planning process of two products: chips for contactbased and contactless payment of the Chip Card & Security (CCS) department. The relevance lies in their high production volume and revenue share of more than 25% of CCS's total revenues.

Research objective
The graduation project aims to solve the below stated research objective:

Improve the supply chain planning process according to the service level and respective costs at CCS for two particular products considering the stocking strategies as well as the approach of quantifying the amount of wafers to be released to production.

The stocking strategy concerns the decision where to place inventory and which amount to be stored. The production release approach examines the question how to quantify the amount of wafers (release quantity) to be started in production in advance. Usually, the production of wafers, which are thin slices of semiconductor material and serve as basis for many products, is started on forecast due to long processing times. This enables faster response to customer demand. A clever chosen approach of estimating the needed quantity helps cutting costs as stock levels can be reduced at the same α-service level.

Methodology
The existing simulation model (discrete event simulation) of Infineon's flexibility & econometrics team is used to study various scenarios. These combine different stock strategies and production release approaches among others the current practice. Before conducting the simulation study we require to parametrize the simulation model to the needs of the two exemplary products. This involves to ensure that the generated demand by the simulation model is similar to the observed demand of the products such that results are valid. An iterative approach is performed consisting of setting the parameters in the demand generation method and subsequently assessing the fit between the generated demand and observed demand. The fit is assessed by a modified Kolmogorov-Smirnov approach where we compare

the total area between the cumulative distributions of the two demand series. The iterations are stopped when the total area is $< 10\%$ of the area below the cumulative distribution of the observed demand. To check the consistency of the chosen demand series we further apply the Chi-Square test.

Analysis of current situation
Infineons supply chain has three stock points (up- to downstream): the master storage, die bank, and distribution centre. The amount of products stored at these stock points is determined by the target reach. The target reach is defined as the safety stock in number of weeks. Currently, CCS has a target reach of 13 weeks at the master storage, and no stocks at the die bank nor the distribution centre since the customer order decoupling point (CODP) lies at the master storage and thus products become customer specific in the downstream manufacturing steps. Storing at the master storage employs the risk pooling effect. The stocks are managed periodically (per week). The production up to the master storage is done on forecast by using a four month moving average (MA) over the historical data. The remaining manufacturing steps are continued when a customer order arrives. The overall performance can be given by the α-service level. The α-service level becomes either 100% when all orders are satisfied by the on-hand inventory during period t, or it becomes 0% when demand is not satisfied completely from stock. Currently, the α-service level is 98%.

Conclusion

- Both new production release approaches: a simple MA over five weeks as well as single exponential smoothing (SES) outperform the current approach that uses a simple four months MA since they allow faster reaction in production as fluctuations are not as smoothed out as with a large time horizon of four months.
- Applying either of these new approaches costs can be cut by 40% since the target reach can be reduced from 13 to five weeks while keeping an α-service level of 98%.
- Comparing the simple MA over five weeks and SES, the moving average performs slightly better. In addition, as it is easy to understand and to apply, we recommend to use the simple MA with a five week time window. That is, reducing the current time window from 16 to five weeks.
- When keeping the current production release approach, the target reach at the master storage can be reduced from 13 to about eight weeks while having only a marginal drop by around 0.5% in the α-service level of currently 98%.

Recommendations

- Enhance the demand generation method of the simulation model such that it is able to create intermittent and autocorrelated demand which is currently not supported by the simulation model. Note, the products we are considering do not show autocorrelation nor are classified as intermittent, however there are autocorrelated and intermittent products at Infineon.
- In addition, implement machine capacity and idle costs as currently capacity is unlimited and costs are solely evaluate according to the WIP and stock levels. However, in reality capacity is restricted and idle costs play an important role as machines are very expensive.

- Last, more elaborated approaches of quantifying the amount of wafers to be started in production in advance such as advanced exponential smoothing techniques, Holt Winter procedure, or ARIMA models may be examined when the demand shows a trend, seasonality, or autocorrelation.

Preface

I hereby proudly present you my master's thesis. This marks the end of a wonderful, exciting, and instructive chapter of my life - my studies in Industrial Engineering at the University of Twente - but it also marks the beginning of a new chapter - starting my first real job at Infineon and eventually being a grown up. Staying at this point I want to thank a few people who guided me during this research and supported me throughout my studies.

First of all, I want to thank Frank Federmann, my company supervisor, for the enormous support he gave me. He patiently guided me through this research and always found time to discuss issues that came up, helped me with problems I struggled with, and provided me with valuable ideas. Also, I am very grateful for the support of the whole scenario & econometrics team and my initiation into their team. Not only the digestive walks after lunch refreshed my mind, but I also enjoyed our after work activities like go karting and 'escaping the room'. I am sure this will continue.

My special thanks goes also to Ahmad Al Hanbali, my university supervisor, without whom I would not have ended up at Infineon as he provided me with the contact when I told him that I want to go to the southern part of Germany, closer to the Alps. He contributed to this research by giving very useful advice and detailed reviewed the thesis. Thanks also to Matthieu v.d. Heijden for providing me with feedback and remarks.

Last but not least I want to say 'thank you' from the bottom of my heart to my beloved family (my mum Jutta & her partner Werner, my brother Yannick, and my sister Lara), relatives, and friends (sorry for not mentioning you by name, but that list would be quite long) who paved my way throughout my studies and went along this sometimes easy and fun but also sometimes rough and steep path. Without my mum, Jutta, I would not be where I am now as she always found the right words to encourage and motivate me during my whole studies and research whenever I felt lost. Also, she took care of the financial resources that are necessary when enjoying the student life, thanks also to my grandparents, Helga and Walter, as well as my godparents, Heidi and Eckhard. In addition, I am so grateful to have my twinsister, Lara. We do not need many words or emojis to understand each other and I can always rely on her. Together, we spent our weekends in the library cheering up one another and supporting each other. Finally, even though our path partly split up, I want to thank Julian who supported my decision to do my master studies in the Netherlands and came along. He was never tired of cheering me up and motivating me to another triathlon training session.

I certainly could fill some more pages with friends who I would like to thank, so to everyone I did not mentioned explicitly, but who I studied, did sport, lived, and worked with: 'THANK YOU VERY MUCH'!

Sophie Lingelbach

Contents

Acronyms

ADI average inter-demand interval. 36

asp average selling price. 33

ASSY assembly. 33

ATV Automotive. 1, 17

BE back end. 28, 33

CCS Chip Card & Security. 1, 8, 15, 52

CI confidence interval. 70

CODP customer order decoupling point. 61, 62, 64

CT cycle time. 14, 15, 28, 33, *Glossary:* cycle time

CV coefficient of variation. 20

CV^2 squared coefficient of variation. 36

DB die bank. 28

DC distribution centre. 14, 28

DES discrete event simulation. 1

DMOP data mart order processing. 19

DR delivery reliability. 16, 17

DTW dynamic time warping. 42

FAB fabrication. 13, 33

FE front end. 28, 33

FF freeze fence. 30, *Glossary:* freeze fence

Infineon Infineon Technologies AG. 1

IPC Industrial Power Control. 1

KPI key performance indicator. 16, 32

MA moving average. 37, 38

MAE mean absolute error. 39

MS master storage. 28

MSE mean squared error. 38

PMM Power Management & Multimarket. 2

POD proof of delivery. 16

RMSE root mean squared error. 39

SCOR Supply Chain Operations Reference Model. 11

SES single exponential smoothing. 38, 62

SP sales product. 20, *Glossary:* sales product

TC total costs. 32–34

wacc weighted average cost of capital. 33, 34

WIP work in process. 8, 33

WS wafer start. 28

Glossary

basic type Basis product that receives customer specific information in the sort, thereby splitting up into a variety of sales products. 13, 15

customer order Order by customers. They contain the required quantity and delivery date for the needed sales products, which is binding. 4

freeze fence The number of periods from now onwards into the future where demand does not get modified. That is, if the freeze fence is three weeks, we know the orders for sure that arrive in the following three weeks. 30

marketing forecast Forecast made by marketing. They use a four month moving average over the observed demand to determine the needed quantity for the two basic types. 4, 29

production release approach It is the approach of quantifying the amount of wafers to be released to production in advance. 3, 61

release quantity The number of wafers (thin slices of semiconductor material which are the basis for producing microchips) or pre-processed products to be started in production. For the front end production this is usually done on forecast. 27, 28, 37

sales product Customer specific product. 13, 15

simulation forecast Forecast created in the simulation model. Demand is generated for a period of 26 weeks, where demand for the period from the freeze fence to the end of the 26 weeks horizon is determined as forecast. This forecast is subject to changes. 29, 30, 50

stocking strategy It considers the two decisions at which stock points to place inventory as well as setting the inventory level. 3, 61

target reach The safety stock in number of weeks. 12, 16

List of Figures

List of Tables

Chapter 1

Introduction

Semiconductors are part of everyone's daily life. When it comes to electronics such as smartphones, power tools, medical systems, automobiles, robots and many more, semiconductor devices are an indispensable component of it. And still, the number of applications for microchips is continuously growing since the transistor was invented in 1948. The competition among semiconductor manufacturers goes hand in hand with the increasing demand. Making it necessary for companies to not only offer their products at favourable prices but also to deliver on time [29]. Thus, companies strive for a competitive advantage through their supply chain management.

Infineon Technologies AG (Infineon), a German semiconductor manufacturer, uses discrete event simulation (DES) to continuously improve its supply chain and to remain competitive. Simulation depicts a system in a software based model with the purpose of understanding its behaviour or evaluating different strategies [56]. Hence, one aim is to reveal bottlenecks and deficiencies. By altering input parameters one tries then to remedy these weaknesses. As a result improved system settings are proposed.

A crucial factor of a simulation is its validity. Meaning that the simulation model has to reflect reality appropriately. This includes that the input data to the simulation model is accurate. Commonly, one generates input data which reflects observed values. This research supports Infineon to find a method that assesses the fit between observed and generated data such that the simulation input can be verified to reflect reality sufficiently. After evaluating the accuracy of the input data we further conduct a simulation study to improve the stocking strategies as well as the method of controlling the production start for two exemplary products of the Chip Card & Security division.

We start with briefly introducing Infineon in section 1.1 and continue by motivating the research topic in section 1.2. Then, in section 1.3 and section 1.4, we define the core problem and formulate the research questions which contribute to solving the problem. Last, section 1.5 describes the scope and limitations of our research and section 1.6 concludes the chapter with the Plan of Approach.

1.1 Company Introduction

Infineon is a semiconductor company, that manufactures devices such as diodes, transistors, and integrated circuits known as 'chips' or 'microchips'. It is positioned in four main markets: Chip Card & Security (CCS), Automotive (ATV), Industrial Power Control (IPC),

and Power Management & Multimarket (PMM), where it holds leading positions. It strives for excellence by making life easier, safer and greener.

Main applications for Infineon's products in CCS are microcontrollers for payment systems, governmental identification documents, and sim cards to Gemalto, Oberthur, and G&D. ATV offers among others driver assistance and security systems such as airbags and ABS as well as general electronics like lighting and windowlifts. Customers include Bosch, Continental, and Tesla. IPC focuses on electric engines, renewable energy as well as energy transmission and conversion for machines, locomotives, wind turbines, and solar collectors sold among others to Siemens and ABB. The PMM division provides chips for consumer goods for instance mobile devices, televisions, and computers, and sells these to large OEM's and various large semiconductor distributors.

Infineon's microelectronic revenues are about $6.5 billion with around 35,400 employees worldwide in the fiscal year 2016. This is allotted to 41% of Automotive, 17% of Industrial Power Control, 32% of Power Management & Multimarket, and 10% of Chip Card & Security [36].

The master's thesis is conducted with the scenario & econometrics team of the corporate supply chain department at Infineon in Neubiberg. The team contributes to the success of Infineon by providing analyses and support services to the business divisions. This includes analysing trends and innovations over all four main markets and proposing interventions in the supply chain planning.

1.2 Research motivation

The key challenges faced by semiconductor supply chain management such as of Infineon include product variability (also refered to as product mix), demand fluctuations, long lead times, and a 24x7 production. These challenging issues influence the manufacturing efficiency, delivery performance, and volume elasticity considerably [9].

Product variability emerges due to the mere fact that product details are often customer specific. These may solely be slight changes or enhanced versions, however they alter the product noticeably. A reason for the rapid development of products is Moore's law, which states that the number of transistors on an integrated circuit is doubling every two years [45]. Also, the wide spectrum of applications for semiconductors leads to a variety of products. Applications range from chips for smart cards, over microcontrollers for automobiles, to large power supplies in industry. Hence, semiconductor companies manufacture an immense range of products simultaneously. This product variability issue is aggravated by unpredictable demand, long lead times, and a 24x7 production at Infineon. Semiconductor companies are plagued by demand fluctuations due to their upstream position in the end-to-end supply chain. Before a semiconductor device reaches the end customer it moves along the supply chain, in our case from Infineon to distributor to customer. This fosters the Bullwhip effect that distorts demand information as it is transmitted up the chain. More precisly, demand variability increases when travelling upstream [43]. Due to the Bullwhip effect, firms in upstream positions cope with high demand fluctuations that impair controlling inventory, forecasting demand and scheduling production. Long lead times and a 24x7 production add up to this difficulty. The manufacturing process of integrated circuits takes up to three to four month. The process from the silicon raw material to the finished good comprises four main processes: Wafer fabrication, sort, assembly or packaging, and final test, some of these

including several hundred steps [29]. However, many costumers do not order three month in advance. Thus, to hedge against uncertainties semiconductor companies need to hold comparably large inventories [9]. These problems are strengthened by a 24x7 production at Infineon. A 24x7 production does not allow for volume flexibility, meaning that a high demand cannot be fulfilled by increasing production hours since production runs already continuously. Hence, incorrect volume planning cannot be remedied by working extra hours, but instead leads to delayed deliveries, which in turn reflect a poor delivery performance.

To hedge against manufacturing inefficiency, demand uncertainty, and missing volume elasticity Infineon places stocks at various stocking points in its supply chain. To help the divisions improve their stocking strategies Infineon's flexibility & econometrics team uses discrete event simulation. Usually, the different strategies are evaluated according to the trade off between the α-service level and the costs.

1.3 Problem definition

In a former project Infineon's flexibility & econometrics team performed a simulation study for a group of products of CCS which showed that stocks can be reduced drastically. In this project we continue the successful collaboration with the CCS division. They are interested in analysing various stocking strategies and production release approaches for two particular products. That is, we try to answer the following questions:

stocking strategy:

1. At which stocking locations to place inventory in the supply chain?
2. How high should the inventory be at these locations?

production release approach:

3. How to quantify the amount of wafers to be released to production in advance?

The two products are of relevance for CCS due to their high production volume and revenue share of more than 25% of CCS's total revenue. To provide CCS with answers to these questions, we use the existing simulation model. It was built by the scenario & econometrics team and further enhanced as part of a master's project such that it allows for flexible product structures [13]. It is described in more detail in chapter 3. With the simulation model we are able to run various experiments where we alter the stocking strategy and the production release approach.

A key part of a simulation study is to have accurate input data. That is, the generated data should be similar to the observed data. Otherwise, results are misleading and proposed solutions do not show the same behaviour in reality as they do in simulation. Currently, there is no established method at Infineon to validate the generated input data according to the observed input data. As part of this project, we require to find a method that assesses the fit between the generated and observed data.

The input data to the simulation model is the demand of the produced products. We distinguish the demand into a forecasted and an actual demand arrival process. By forecast we refer to the estimated quantity customers buy. Marketing creates this forecast by using a four month moving average of the historical orders which is then validated by the supply chain planner. By demand we refer to the actual orders customers place. Those two arrival processes

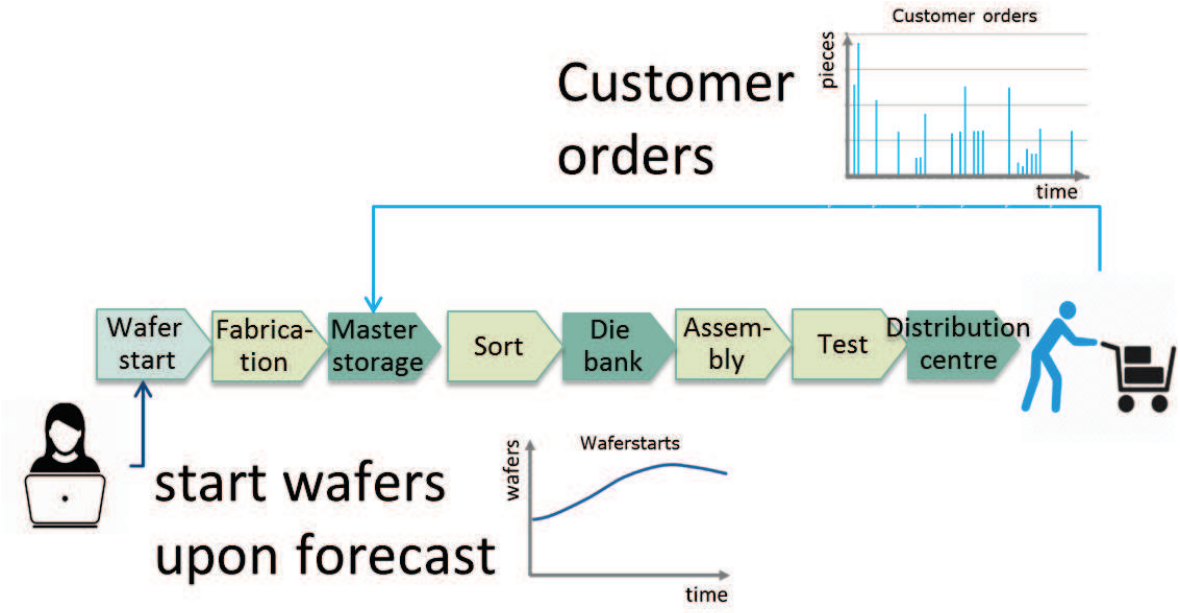


Figure 1.1: Production start according to forecast and further processing on basis of customer orders

differ since the forecast represents a moving average and the demand refers to the actual orders customers place. Figure 1.1 presents where in the supply chain the marketing forecast and the customer orders are employed. For the products we are considering, the marketing forecast is used to start the production of unprocessed wafers up to the diversification point (master storage). Hence, the current production release approach is based on the marketing forecast. Whereas the customer orders are used to start the production of the pre-processed wafers from the diversification point onwards. The chips become customer specific and are waiting at the distribution centre for delivery to the customers.

The simulation model generates the customer orders. To receive valid simulation results, we require that the generated data resembles actual customer order data such that their characteristics are similar. Figure 1.2 shows an example of actual and generated demand data. There exists various techniques in the literature to compare two time series and assess their fit. For example, we can use forecast accuracy measures such as the MAE (Mean Absolute Error), the MAPE (Mean Absolute Percentage Error) and the SMAPE (Symmetric Mean Absolute Percentage Error). These measures compare a forecasted value at time t with the observed value at time t. The difficulty with these measures is that they compare two points with one another and do not consider the overall behaviour of the time series. However, we are rather interested in a statistical equivalent behaviour than in the exact values. The advantage of having a statistical equivalent behaviour is that we can generate various realisations of this demand behaviour and use it for several simulation runs. This ensures that the output is not only based on one realisation but on many and thus reduces the effect of outliers. Hence, we want to find a method with which we can assess the fit between two time series regarding their statistical behaviour.

Our procedure is as follows. We parametrize the arrival processes in the given simulation model such that the generated data represents observed data according to our defined method.

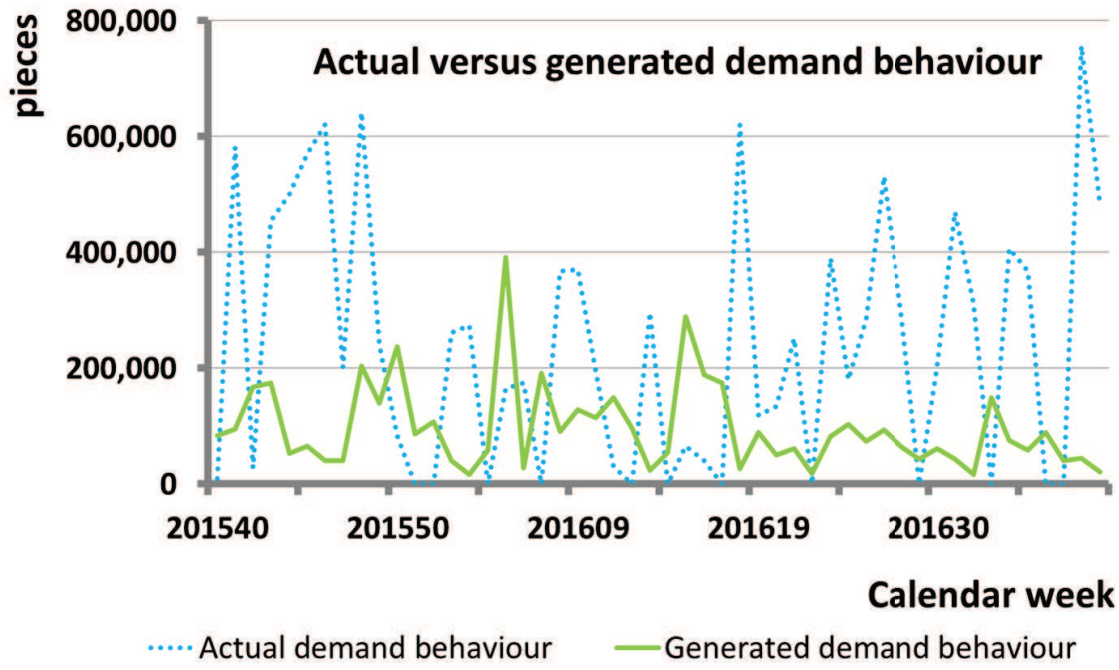


Figure 1.2: Example of actual versus generated demand data

Next, we use the adjusted simulation model to study the two products of CCS. This regards various stocking strategies and production release approaches which are evaluated according to the service level as well as respective costs. We aim to find the set up that improves the supply chain planning of CCS.

In summary, we can formulate the core problem in the following statement:

> ***Improve the supply chain planning process according to the service level and respective costs at CCS for two particular products considering the stocking strategies as well as the approach of quantifying the amount of wafers to be released to production.***

To be able to solve this core problem we need to solve the two subproblems below which are strongly interconnected with the core problem and thus are highly emphasized. These are:

> ***Parametrize the order arrival process in the existing simulation model such that the generated demand data correctly describes the observed data to make the simulation results more representative.***
>
> ***Define a method to assess the fit between generated and observed values according to their statistical behaviour.***

To solve the core problem, we proceed with the research cycle which provides a framework to generate lacking knowledge [30]. Our knowledge questions are formulated as research questions in the next section.

1.4 Research problem

1.4.1 Research goal

Currently Infineon does not have an established method that assesses the fit between the historical demand and generated demand data. In order to receive valid simulation results the input to the simulation model has to reflect observed values correctly. A key prerequisite of the method is its ease of use. Consequently, the goal of this research is, firstly, to parametrize the arrival process such that it models the true behaviour of representative products, secondly, to construct a method which assesses the fit between generated and observed data, and thirdly, to conduct a simulation study for two exemplary products of CCS. This simulation study aims to consider various stocking strategies as well as approaches to quantify the amount of wafers to start in production in order to give alternatives to the current practice.

1.4.2 Problem statement

The problem statement is formulated to generate the needed knowledge.

> ***'How can Infineon assess the fit between the generated and observed demand data for representative products of the CCS division and parametrize the simulation's arrival process to obtain valid results?'***

We want to exploit existing techniques to evaluate the fit between two time series. Values at time t of the generated data do not need to match values of the observed data at time t exactly, but we aim to assess whether the overall behaviour of the series is statistically similar.

1.4.3 Question formulation

To tackle the problem statement, we formulate several research questions. Each research question including its sub questions corresponds to a chapter of this thesis. These research questions will be answered by interviews with employees of Infineon, reviewing available literature, performing an elaborative data analysis, developing a method to evaluate the behaviour qualitatively and conducting a simulation study.

Current situation.

First, we obtain in-depth knowledge of the current situation. For this purpose we look at two domains, the broader context and the data of the considered products. The context involves gathering information about how supply chain planners at CCS define production volume, which data sources are used and how orders influence the production start. In addition, we look closer at the data of the representative products and conduct an analysis to identify patterns.

1.1) How is the supply chain planning carried out?
   a) How is the supply chain set up?
   b) Which products are representative and appropriate to consider?
   c) Which data sources are used for storing the demand data at CCS?
   d) How do orders and forecasts influence production start?

1.2) How does the demand data of the representative products from Chip Card & Security behave?

a) What patterns can be identified in the data?
b) Which statistical measures are important to consider?

Simulation model.
We use discrete event simulation to analyse various system settings. Thus, we explain the methods, inputs, and outputs of the existing simulation model.

2) How is the simulation model set up?
   a) What is the purpose of the simulation model?
   b) What is the structure of the model, e.g. logic, input and output parameters?
   c) How does the simulation model work?
   d) What are the input and output parameters of the simulation model?

Literature review.
We continue with a literature review to study existing approaches concerning how data series can be compared. Several approaches exits in literature which concern among others forecast accuracy measures, time series similarity measures, and hypothesis tests. This lays a foundation to assess our arrival process which should model observed demand behaviour appropriately.

3) What solution approaches exist in literature to assess the fit between generated and observed demand data?
   a) How can two time series be compared?
   b) What are the advantages and disadvantages of these measures and methods?

Parametrization and fit between time series.
The next step is to parametrize the arrival process of the simulation model to generate demand data which represents the behaviour of the observed data. Moreover, we apply a suitable method to assess the fit between two time series based on the findings of the literature review.

4) How do we need to parametrize the simulation model to create accurate demand data?
   a) What input parameters are relevant?
   b) How accurately does the generated data fit to the observed data?
   c) How can we improve the fit between the historical data and the generated data?

Simulation study and Evaluation.
The simulation study is performed in order to assess several stocking strategies and production release approaches for the analysed products of CCS. To compare the various approaches we need to define key performance indicators (KPIs). The results of the simulation study serve as an indication how CCS can improve their supply chain planning process. We conclude the thesis with recommendations and an outlook.

5) How can the planning process of CCS be improved?
   a) Which strategies can be used to start production?
   b) Which stocking locations should be used to store items?
   c) How high should the stocks be at the various stocking points?
   d) What are the improvements of the proposed set ups?

1.5 Research Scope and Limitations

Due to time constraints of this research project and some limiting factors we narrow down the scope and mention simplifying assumptions.

As introduced earlier, Infineon is structured in four divisions, all of which provide a wide range of products. Since time constraints do not allow considering all products, we will focus on two exemplary products from Chip Card & Security (CCS). We restrict our selection to products of CCS for the reason that these products show a volatile behaviour, whereas products from for example ATV are rather stable in their demand patterns. Furthermore, the existing simulation model is built on the supply chain specifics of products from CCS. Hence, major modifications of the simulation model will not be required.

We can omit the validation of the simulation model since it was validated by a previous master's thesis that enhanced the used model [13]. By valid we mean that the physical supply chain is mapped well enough in the simulation. The focus solely lies on adjusting the existing model with an accurate parametrized demand and forecast signal, however, we do not focus on the process steps in the simulation to accurately represent the supply chain.

In addition, we differentiate this master's project from a previous project which was also done with Infineon's flexibility & econometrics team in cooperation with the Faculty of Behavioural, Management and Social scienes of the University of Twente [1]. The previous project considered the trade off between the utilisation of machines and the resulting costs of storing inventory. A higher utilisation of machines leads to a higher cycle time due to a higher work in process (WIP). The focus lied on improving the accuracy of the simulation model. In contrary, in this project we concentrate on the trade off between the service level (associated with high stocks) and the respective costs without considering machine capacity. Moreover, we focus on adapting the simulation model according to two particular products to improve their planning process.

1.6 Plan of Approach

A well-known approach of structuring and solving research is the Managerial Problem Solving Method (MPSM). The method intends to solve an action problem as identified in section 1.3, which states to improve the supply chain planning process of Infineon and thereby modelling demand behaviour according to observed values. The MPSM is composed of various phases.

1. Identifying the problem
2. Planning the problem-solving process
3. Analysing the problem
4. Generating alternative solutions
5. Choosing a solution
6. Implementing the solution
7. Evaluating the solution

Figure 1.3 below maps these phases to the chapters of the thesis to give a brief overview of the structure and determine the activities for each step.

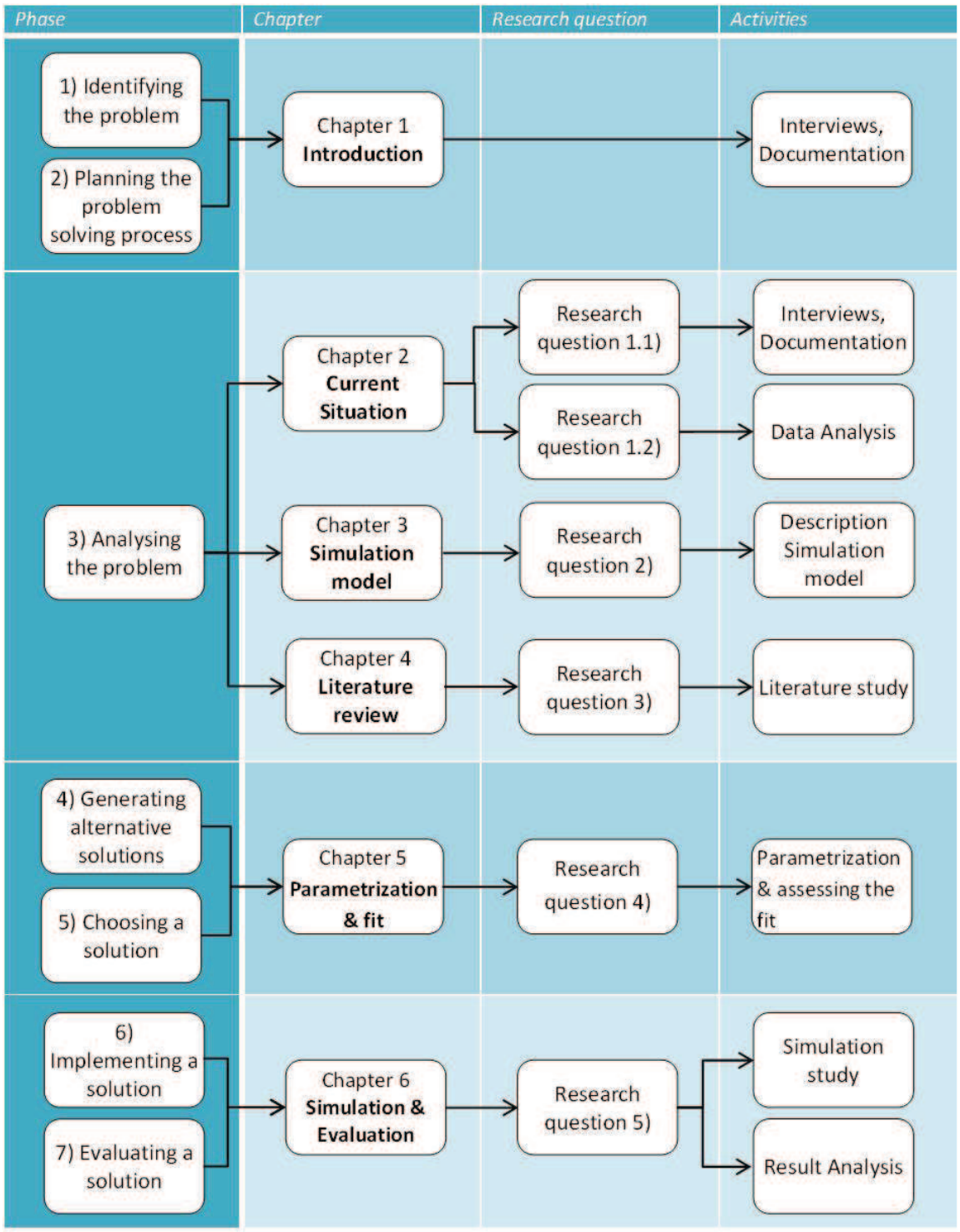


Figure 1.3: Plan of Approach

Chapter 2

Current situation

To start with, section 2.1 introduces the supply chain process of Infineon and goes into more detail regarding the planning of CCS and its products which we simulate later in this project to improve their stocking levels. In section 2.2, we continue with a data analysis that serves as a basis for modelling the arrival process of the demand and forecast in the simulation. The focus of the data analysis lies on two high runner products.

2.1 Current Situation

2.1.1 General description of Infineon's supply chain and its planning

In order to manage its supply chain processes Infineon implemented the Supply Chain Operations Reference Model (SCOR), which is a management tool recommended by the APICS Supply-Chain Council. It describes the business activities associated with five distinct phases to satisfy customer demand [19]. The phases are: Plan, Source, Make, Deliver, and Return. Figure 2.1 shows how these five phases of the SCOR model link to Infineon's supply chain.

We focus on the activities of the *plan* and *make* process which are relevant for our simulation study and only give a brief description of the other three phases. The discussion in the remainder of this section refers to the internal documents [19,55] of Infineon. The *plan* process is responsible for balancing the available resources with the given requirements. The *source* process takes care of the deliveries from internal and external suppliers, including purchasing activities as well as sourcing logistics. The *make* process includes the main production steps of the supply chain, namely fabrication, sort, assembly and final test. The *deliver* process concerns all sorts of deliveries to internal and external customers and thereby taking care of order management, and invoicing customers. Last, the *return* process deals with products that are either returned by Infineon to its suppliers or by customers to Infineon [19].

The Plan process.

Figure 2.2 shows the plan process which is further divided into five subprocesses. We concentrate on the demand planning at an operational level concerning a time horizon of about six months. The subprocesses are responsible for different tasks: 1) The capacity planning aggregates machine resources that are available, 2) the demand planning is responsible for machine requirements needed to produce specific products and 3) the supply planning balances the capacity with the demand, thereby creating a production request for production

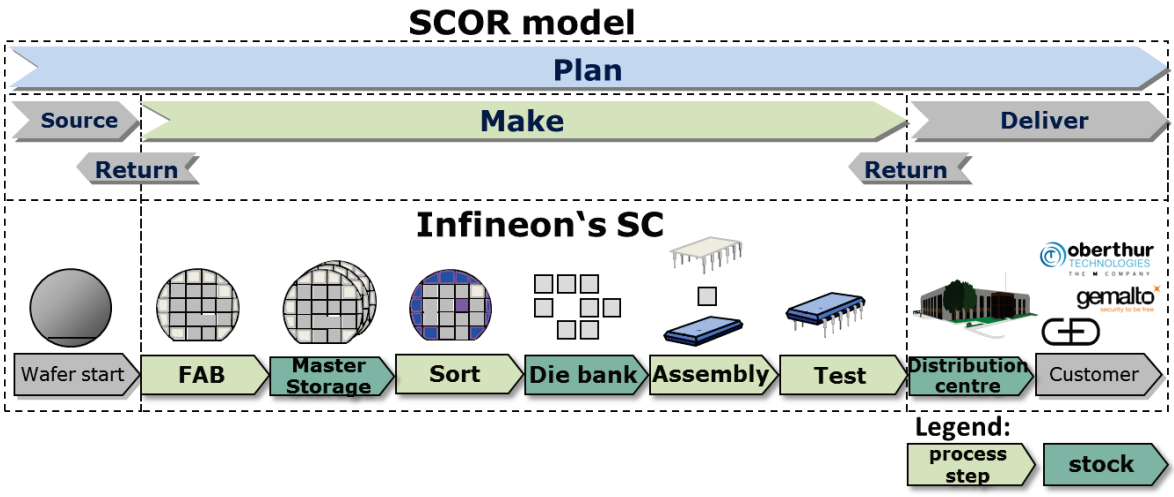


Figure 2.1: SCOR model linked to Infineon's supply chain [19]

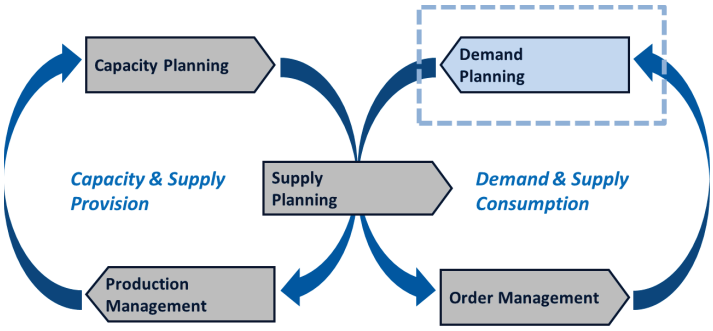


Figure 2.2: Plan processes at Infineon [55]

management and a constrained forecasts, that is, the amount Infineon can sell considering available machine resources, for order management. Last, 4) the production management as well as 5) the order management establish and communicate plans for production and customer deliveries, respectively.

There are two major activities in the demand planning process on the operational level regarding our simulation study. First, the generation of a forecast of what Infineon could sell into the market in number of pieces per week. We aim to model this weekly forecast data in the simulation model to increase the model's validity. Second major activity is the definition of the target reach for the stocking points. The target reach is defined as the safety stock in number of weeks. That is, the supply chain planner determines how much to store at the various stocking locations for each product. Stocks are needed for three main reasons:

1. Uncertainty in demand and production
2. Long cycle times
3. Strategic decisions

Demand uncertainty occurs due to varying orders of customers and production uncertainty occurs due to machine downtime which varies the cycle time. Stocks are built to hedge against

these uncertainties. In addition, cycle times are quite long due to a complex manufacturing process. To be able to respond quickly to customer demand, stocks are needed to reduce the lead time. Moreover, stocks are necessary when production gets transferred to another production location. E.g., production may be transferred from location 1 to location 2, however, customers may require to further receive their products from location 1 since they solely certified location 1. Thus, we need to build up stocks for these customers with products of location 1.

We intend to improve stocking levels and the approach of quantifying the amount to start in production by conducting the simulation study because it is important that stocks are neither too low nor too high. If stocks are too low, master storage and diebank products are missing and customer orders cannot be confirmed. As a result Infineon loses revenue and dissatisfies its customers who may move to competitors. On the other hand, if stocks are too high, Infineon invests in unnecessary products and hence raises the bind capital. Moreover, it increases the risk of scrapping master storage products, die bank goods and finished products [55]. This situation can be described by the trade off between the service level and costs. A high service level indicates comparably high stocks and thus also high costs, whereas low stocks are associated with a lower service level and also lower costs. The aim is to balance this trade off.

The Make process.

The main result of the make process is the finished product, namely the silicon chip or microchip. Making silicon chips is one of the most complex manufacturing tasks. It is grouped into front end and back end, taking between 40 and 100 days (6-15 weeks), and up to 20 days (3 weeks), respectively. In the front end chips are produced onto the silicon wafers. In the back end wafers are diced into single chips. These single chips are equipped with an outer package containing pins or a conductive surface to connect with other electronic components. Both processes are separated by the die bank. Figure 2.3 gives a schematic overview of the process and possible stocking points, similarly to the existing simulation model.

Wafers are produced from raw silicon, which builds the basis for microchips. Silicon is used due to its properties as a semiconductor. Depending on the treatment it either conducts or blocks the flow of electricity making it ideal to function as a transistor.

Figure 2.3 illustrates the description of the process steps: To start with, the silicon wafers are treated in the fabrication (FAB), where the developed chip design is coded onto the wafer. To fit several millions of transistors onto a single chip it is build up in three dimensions consisting of various layers. Steps in the process involve lithography, furnace, implanting, deposition as well as etching and these are repeated multiple times until the integrated circuit is completely built in the wafer. The master storage serves as a stocking point for processed wafers containing several hundreds of chips each. Subsequently, in the sort, wafers are tested for their functionality and marked accordingly. Also, one should note that the products we are considering are receiving customer specific information during this step. The rather general products are made to stock up to the master storage. We call them basic types. Whereas the customer specific products out of sort are produced upon order requests up to the die bank or distribution centre. We call them sales products. Hence, the sort is a diversification point in our supply chain, which indicates that the customer order decoupling point lies rather upstream in the supply chain. This is called make-to-order. Later, at the die bank customer specific products are stored temporarily waiting for disposition and further production at the assembly. There is no defined target reach at the die bank, however to fill up machine

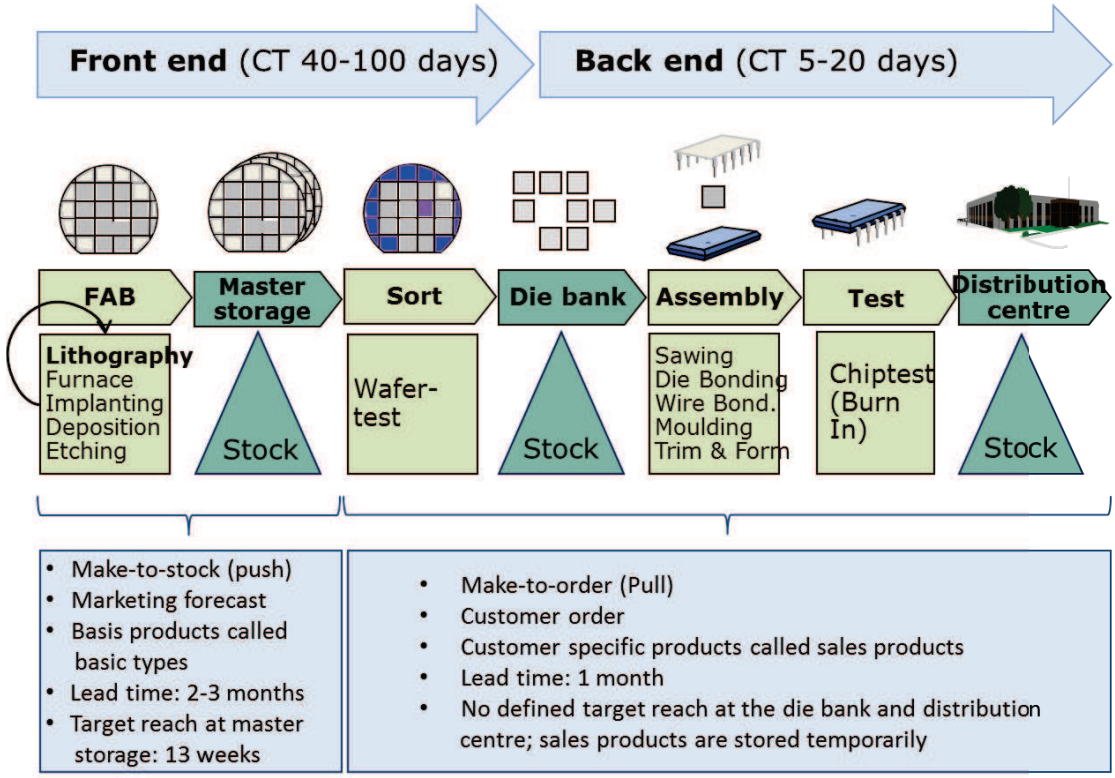


Figure 2.3: Make process at Infineon [55]

capacities more products than requested may be produced. At the assembly wafers are cut into individual chips and in the die bonding a package is attached to the chips. It is followed by the wire bonding, where interconnections between the integrated circuit and its package are made. Finally, the chips are moulded, trimmed and formed. Now they are ready for the last quality check, the Burn-in testing. It stresses the component under supervision to detect defective chips. Chips which fail this test are sorted out. The last part of Infineon's supply chain forms the distribution centre (DC). At this stocking point finished products are stored before they are transferred to the customers [19]. Typically, there is no stock at the distribution centre for make-to-order products. However, customers may request to have finished products at the distribution centre or products may be stored temporarily before delivery [37].

The cycle time (CT), defined as the length of time spent by a product unit in the system from the release of the wafer into the fabrication until finishing the last step in the test takes up to four month without considering storage time in the master storage or die bank. Two to three months are allotted to the front end and roughly one month is allotted to the back end. These long cycle times especially in the front end indicate that it is necessary to use forecasts up to the master storage and die bank such that customer orders can be fulfilled quickly in order to stay competitive.

Infineon's supply chain process is spanned over a global network, meaning that there are various production and stocking locations spread all over the world. Front end facilities are among others in Dresden (Germany), Regensburg (Germany), Villach (Austria) and Kulim (Malaysia) and external suppliers include the Taiwan Semiconductor Manufacturing Company

Table 2.1: Plan cycle time for production steps of BT1 and its sales products

Basic type BT1	**Front End**		**BackEnd**	
Productionstep	*FAB*	*SORT*	*ASSEMBLY*	*TEST*
Facility: CT in day	Dresden: 91	Dresden: 18	Regensburg: 7	Regensburg: 0
	TSMC: 91	ADT: 7	Wuxi: 7	Wuxi: 0

Table 2.2: Plan cycle time for production steps of BT2 and its sales products

Basic type BT2	**Front End**		**BackEnd**	
Productionstep	*FAB*	*SORT*	*ASSEMBLY*	*TEST*
Facility: CT in day	Dresden: 70	Dresden: 14	Amkor: 9	Amkor: 7
	TSMC: 70	ADT: 4.5	Regensburg: 7	Regensburg: 0
			Wuxi: 7	Wuxi: 0

(TSMC, Taiwan) and Ardentec (ADT, Taiwan). Back end facilities are located among others in Regensburg (Germany), Warstein (Germany), Malacca (Malaysia) and Wuxi (China). External partner is for instance Amkor Technology (United States of America). The die bank locations are either based at the front end or the back end facilities [19].

2.1.2 CCS's high runner products

Chip Card & Security (CCS) focuses on products in three main areas: payment systems, governmental identification documents and mobile communication [36]. The two products we are considering, BT1 and BT2, belong both to payment systems. Product BT1 is a chip for contactbased payment integrated in credit and debit cards and BT2 is a chip for contactless payment also integrated in credit and debit cards. Other payment systems are mobile payment and NFC-based contactless payment. Products BT1 and BT2 are of main interest, since they contribute to CCS's yearly total revenue by >25% and have a high production volume. [31].

As we introduced Infineon's supply chain in the previous section, we give here some further information of the two products. Table 2.1 and Table 2.2 summarize the specific CTs in days per production step and also depict the facility locations where the treatment takes place. When production is started at the fabrication we speak of basic types. BT1 and bp are both basic types. In the sort step these two basic types receive customer specific information and are then identified as sales products. A basic type can serve as a basis for several hundreds sales products. In our case, 95 different sales products are made from basic type BT1 and about 180 sales products are made from basic type BT2. Later, we solely consider the largest sales products of each basic type which account for $\geq$ 85% of the total volume. These are six sales products for BT1 and ten sales products for BT2. Customers order on sales product level [37].

The front end production steps, fabrication and sort, of basic type BT1 take place at three locations depending on the workload. These are Dresden, TSCM as well as ADT. That is, they produce at the facility with the lower utilization in order to be able to respond quicker to demand. The rather large time difference of eleven days in the sort between Dresden (18 days) and ADT (7days) is due to transportation. E.g. from fabrication at TSMC to sort in Dresden [31]. Assembling and testing is done in Regensburg and Wuxi. The cycle time for the test itself is negligible and therefore indicated with zero in the table. The total cycle

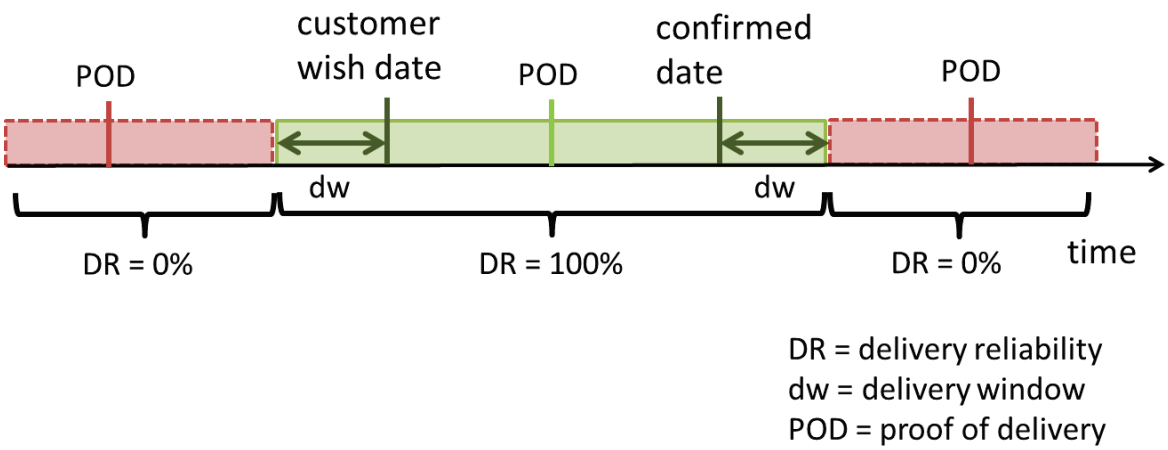


Figure 2.4: Delivery reliability at Infineon

time sums up to roughly four months, where a large part of about three and a half months is allotted to the front end and only a small part of about two weeks is allotted to the back end. Hence, it is significant to plan the right amount of stocks in the front end stocking points, master storage and die bank, to decrease the lead time and fulfil demand quickly. Otherwise, if production is started when orders arrive, lead times are too long for a highly competitive market. Basic type BT2 is processed at the same locations in front end and back end with similar cycle times. Next to the back end locations, there is also a facility of Amkor with a slightly higher cycle time which is due to their internal production process.

Clearly, the amount of produced products is dependent on the customer demand. As mentioned earlier, we distinguish between actual customer orders and marketing forecast. A customer order is a request by a customer for a certain amount of one or more customer specific products containing a delivery wish date. Customer orders are produced up to the distribution centre. A marketing forecast, on the other hand, is an approximation of what and how much a customer may order in the future and is done by marketing using a four month moving average over the historical orders. Marketing forecasts are produced up to master storage. This is done in order to decrease the lead time as the cycle time at the frontend can be omitted when orders are produced from master storage. There are no products produced on forecast to the die bank or the distribution centre, since die bank and DC products are customer specific and the risk of scrapping products is too high [31,37]. Currently, the amount of wafers at the master storage cover a target reach of 13 weeks.

The performance of the current approach, which defines the release quantity in front end by a four months moving average with a target reach at the master storage of 13 weeks is measured at Infineon by the delivery reliability (DR). The DR is an internal key performance indicator (KPI) which is calculated for each product. A delivery is considered to be reliable if the proof of delivery (POD) is a date between the customer's wish date minus some delivery window and the first confirmed delivery date by the supply chain planner plus some delivery window as shown in Figure 2.4.

The current delivery reliability for the products BT1 and BT2 is 93%. Note that, the current simulation model does not include a method which captures the interaction between the supply chain planner and the customers such that the DR can be measured since it would

introduce a higher level of complexity and may reduce the runtime of the simulation model. Instead, in order to prevent unnecessary high complexity the α-service level is implemented in the simulation model. It is chosen since it captures the idea of the DR without introducing further complexity. Similarly to the DR, the α-service level becomes either 0% when not all demand is met by on-hand inventory or it becomes 100% when all demand is met by on-hand inventory explained in detail in section 3.3.2. Note that, backorders are not taken into account. Since we have given the delivery reliability but not the α-service level, we need to find the α-service level that corresponds to the DR of 93%. This is done by using the current simulation model. The simulation model was verified by [13], thus we can determine the current α-service level by running the simulation for both basic types with a target reach of 13 weeks at the master storage and a four months moving average over historical data to determine the release quantity. This results in an α-service level of 98%.

2.1.3 Stocking policy approaches at Infineon

Infineon uses various approaches to plan stocking levels at the master storage, die bank, and distribution centre ranging from basic approximations to advanced simulation-based methods thereby increasing the quality of the proposed solution along with the effort. The following discussion based on [21] shows how we classify our project.

A basic approach on a high aggregation level is using a rule of thumb. The supply chain planner estimates the target reach according to his experience and uses the estimated value for all products. Thus, there is no differentiation between products nor fluctuations over time are considered. Nevertheless, it is easy to apply.

To add more detail, ATV introduced the 'Enhanced Inventory Planning' for some of their products. At this level of detail, products are considered separately and the target reach is calculated for the various stocking points by using an echelon stock policy. An echelon stock policy is characterized by central control and the visbility of customer demand in the entire network. An installation stock policy, on the other hand, is characterized by local control and the demand at each stocking point is based on the demand from the downstream stockpoints [6]. For the calculation general inventory models such as the (R, S) policy are used, where every R periods (weekly) an order is placed to rise the inventory position to the order-up-to level S. In order to apply these inventory models, usually the assumptions of a normal distributed demand as well as the independence of succeeding time periods are made, where demand of one period has no influence on demand of a subsequent period [6]. A normal distributed demand facilitates computations and gives a good approximation when demand is high enough [61]. Note that, there exist extensions in the literature in case of non-normal demand and dependent time periods. Fortuin examines five different probability density functions for the demand (Gaussian, logistic, gamma, log-normal, Weibull) [25]. He finds that except for the logistic distribution expressions are much more complex. Burgin further investigates on the Gamma distribution and devotes considerable effort to develop approximations [10]. In addition, other distributions such as Poisson [57], Laplace [51], and Negative binomial [20] have been studied. For a further listing and according references we refer to [57]. A short discussion is also provided in section 5.3. The described approaches are analytical methods to define the order-up-to level S for the stockpoints in multi-echelon inventory systems. It is advisable to use analytical methods when computations are comparably easy and assumptions such as a stable average demand rate are met [41, 57].

In contrast, simulation-based approaches are preferred over analytical methods when com-

plex relationships and detailed structures are modelled as well as when time depending events occur. Simulation allows to explicitly model the relation between products, machines, and operators. That is, different products may have different processing times and different routes through the system. These may further be influenced by various operators. Simulation also allows to include variability in processing times due to machine downtime. In addition, the product structure (basis product splits up into several specific products) can be included in a simulation model with the according demand for the specific sales products. Moreover, even though discrete event simulation implies with its name that events occur at discrete time steps, we can easily vary at which time steps to execute an action, e.g. every time step, every second time step or make it dependent on some conditions. Furthermore, another advantage is that actions can be triggered depending on certain conditions, which may be varying itself. Last, a practical upside of simulation is that processes and changes over time can be shown in graphs and moving figures. This facilitates the understanding of the system behaviour and the communication with management. Thus, using simulation allows for higher flexibility than analytical methods. However, it also requires a higher amount of effort and detail.

Since the manufacturing process of Infineon is highly complex with interactions between various processing steps, variability in the demand and processing times as well as a complex product structure and on the other hand, there is already an existing simulation model for the process of products from CCS, we decide to use discrete event simulation. Specifying the approach of simulation, our aim is to parametrize the demand generation method in the existing simulation model such that it reproduces the behaviour of observed demand precisely. Hence, we require to assess the fit between the generated and observed demand by a suitable method. With the enhanced simulation model we aim to improve the target reach, that is the stocking strategies as well as the production release approach, which is currently based on a four months moving average. The used key performance indicators (KPI) to measure the performance of the strategies are the service level and the costs, which we explain in section 3.3.2.

2.2 Data Analysis

We choose simulation to improve the supply chain process of CCS. To receive valid simulation results, the input to the simulation model has to reflect reality well enough. Thus, we conduct a comprehensive data analysis using Excel and the statistical software R to learn about the data's behaviour. R is used in addition to Excel since it has the advantage of various build in functions such as statistical tests and is able of coping with large data sets. To start with, we gather several data sources and select one based on its completeness and validity. Next, we give a numerical and graphical summary of the data. In addition, we attempt to decompose the time series into a trend, seasonality, and error term. As this fails to recognize a suitable trend or seasonality, we further consider the autocorrelation of the time series to detect whether there are succeeding periods of increasing/decreasing demand. Having several periods with an continuously high demand rises the probability of stock outs. Thus, we conclude that autocorrelated data behaves differently to non autocorrelated data. That is, if the generated data is autocorrelated, but the historical data is not or vice versa, wrong conclusions from the simulation results may be drawn.

2.2.1 Demand patterns

We gather and consolidate several data sources in order to have a complete and valid representation of customer orders. The below mentioned requirements ensure that the data source is representative:

1. Data should be on a weekly basis.
2. Data should contain data points over a minimum of 52 weeks.
3. Data should contain the requested quantities per sales product.
4. Requested quantities by customers should be represented in pieces.
5. Data should contain the due week of the order.

Since planning of customer shipments occurs on a weekly basis we require the level of data to be weekly as well. Further, we opt for a time span of minimally 52 weeks to have a data set that has enough data points to draw conclusions and represents patterns sufficient. Next, customers order the sales products in number of pieces. Last, we require the due week of the orders that customers request such that we can represent how demand occurs over time and to identify dependencies between time periods.

For the validation of the data sources we compare the revenue figures in € with the annual report for the fourth quarter of the fiscal year 2015 and the first three quarters of 2016. We assume that comparing the revenue figures of the available data sources with the annual report is suitable to determine whether the data is complete and contains all sales. This comparison shows that one out of three potential data sources is sufficient for further analysis as it differs only by 2%, whereas the other data sources differed by more than 15% due to missing and incomplete data. The sufficient data source is called data mart order processing (DMOP).

The maximum difference between the data of DMOP and the annual report on a quarterly basis is 3% and the minimum is roughly 0%. Aggregating the numbers on a yearly basis results in a difference of about 2%. The difference may be due to returned orders or when actual payments fall into another quarter. We assume that this represents the demand well enough. Regarding the above defined requirements, the DMOP data fulfils conditions 1 to 4. However, it does not fulfil condition 5, that is, the due week as requested by the customer is not contained in the data. Nevertheless, it contains the week the order was delivered at the customer site. We assume that this is sufficient to represent the demand for a certain week and hence neglect the case that orders are delivered deviating from the due week.

Demand behaviour on the example of basic type BT1.

On the example of basic type BT1 we present the results of our data analysis which aims to provide us with a better understanding of the data and its behaviour. The data analysis was done using Excel as well as the statistical software R. The graphical summary of basic type BT1 is plotted in Figure 2.5. It represents the orders in pieces per week over two years, from January 2014 to December 2015. One can see that the deliveries increase over time and that they are fluctuating. The increase over time is due to the product life cycle which suggests that demand grows until it matures and eventually levels off [31].

Table 2.3 summarizes several statistics for basic type BT1 and its three largest sales products accounting for roughly 72% of the total volume to draw some first conclusions from the data set. Note, later in the simulation study we consider the largest six sales products,

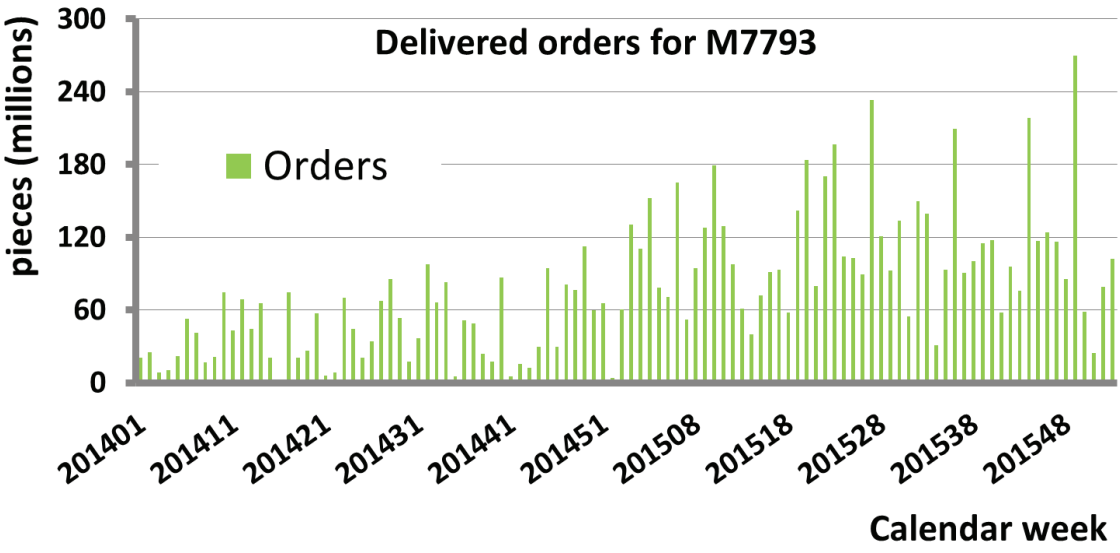


Figure 2.5: Delivered orders of BT1 in pieces (millions) per week from January 2014 to December 2015

which make up $\geq$ 85% of the total volume. The mean demand of the basic type over the two years is about 10 million pieces per week with a standard deviation of around 6 million pieces. When calculating the coefficient of variation (CV), which indicates the dispersion of data points, we receive a value of 0.61. It means that the deliveries are fluctuating since the CV is > 0, however these fluctuations are not very large. Moreover, the median lies with about 8 million pieces comparable close to the mean demand and therefore suggests that the distribution of the values is not skewed to the right or to the left but rather symmetric. Last, we classify the demand pattern by the scheme of Syntetos & Boylan [59] described in subsection 4.1.1. We choose this categorization since it can be applied independent of the empirical data set. The basic type is classified as smooth meaning that it has moderate fluctuations and constantly occurring demand. This implies for the planning that the production release in front end can be rather stable.

Looking closer at the data, we consider the largest sales products of the basic type BT1. In total 95 sales product (SP) are manufactured on basis of this basic type. The largest three sales products, SP1, SP2 and SP3, account for 39%, 24% and 9% over the two years, respectively. When looking at the data of 2015 only, the amount of the three sales products even increases to a total of 91%. SP1 has a mean of about 4 million pieces per week over the two years per week , SP2 has a mean of about 2 million pieces and SP3 has a mean of about 1 million pieces. The median for SP1 lies close to the mean suggesting that there are no big outliers and that the distribution is rather symmetric. However, the median of SP2 and SP3 is zero, meaning that over the two years in 50% of the weeks there are no orders delivered. This also implies that the distribution is skewed to the left. Furthermore, the coefficient of variation for all three products shows that SP1 has a smaller relative variability compared to SP2 and SP3, where SP3 has the largest relative variability. However, the variability of all three sales products is still noticeably. This is in accordance with the classification in erratic for SP1, and lumpy for SP2 as well as SP3. The erratic and lumpy demand patterns are both

Table 2.3: Summary statistics of delivered orders per week for basic type BT1 and its three largest sales products

	BT1	***SP1***	***SP2***	***SP3***
Mean	9,916,641	3,876,597	2,388,385	897,302
Median	8,501,889	3,791,629	0	0
Standard Deviation	6,090,168	3,362,978	3,747,042	1,513,485
Minimum	1,020,868	0	0	0
Maximum	28,130,769	11,338,889	14,961,713	9,795,001
CV	0.61	0.87	1.57	1.69
Classification of Syntetos & Boylan	smooth	erratic	lumpy	lumpy

characterised by fluctuating demand, while demand occurs frequently in the case of an erratic classification but rather seldom in case of a lumpy classification. Erratic demand patterns imply for the planning that forecasts of demand should not be based solely on one demand point as this may lead to large over- and underestimates. On the other hand, for the planning of lumpy demand Croston [14] suggests to analyse the volume of the non-zero demand and the interval between successive non-zero orders separately. Thus, two forecasts are made, one for the volume of the order and one for when the next order will occur.

Last, we consider the correlation among the sales products to check whether there are any dependencies among them. The Pearson correlation coefficient for two samples X and Y is defined by [64]

$$r_{X,Y} = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2}\sqrt{\sum_{t=1}^{n}(y_t - \bar{y})^2}} \tag{2.1}$$

The analysis shows that the sales products are not correlated among each other. Results can be found in Appendix A.

Times series decomposition for BT1.

Time series decomposition is a classical method of time series analysis. It tries to discover patterns in the historical data and extrapolates these into the future. In contrast, regression analysis aims to reveal an explanatory relationship between one or more independent variable and the output [22]. We focus on time series decomposition as we are interested in the patterns and not in an explanatory relationship. Time series decomposition breaks down a time series into subpatterns that identify separate components. This gives the analyst a better understanding of the series, and improves accuracy in forecasting since suited forecasting models can be chosen on the obtained information. Thereby, it assumes that the data is a function of a trend-cycle, seasonality and an error term [22, 33]:

$$\text{Data} = f(\text{trend-cycle, seasonality, error}) \tag{2.2}$$

There are two approaches of the classical decomposition [22, 33]: 1) Additive models and 2) multiplicative models. Additive models assume that seasonal fluctuations stay the same, whereas multiplicative models assume that seasonal fluctuations decrease or increase over time. Our data shows no significant increase or decrease in the variance over time. Hence, an additive approach is applied which comprises the four steps below :

1. Estimating the trend-cycle.
2. Removing the trend-cycle component.
3. Estimating the seasonal component.
4. Estimating the error term.

Time series data is often described by a non-stationary process. For a non-stationary process the mean and/or the variance change over time whereas a stationary process has the property that the mean and variance are constant and do not change over time [8]. Clearly, if there is a trend in the data the series is non-stationary as the mean changes over time. However, many models assume stationarity. Stabilizing the mean can be achieved by either differencing or de-trending the series, stabilizing the variance can be achieved by log-transformation of the data [33]. To detect non-stationarity various statistical tests such as the Dicker-Fuller tests exist in literature.

As we found that the variance does not change significantly over time, we do not need to stabilize it. A stabilization of the mean beforehand is also not necessary as it is part of the decomposition to remove the trend of the data. The trend-cycle is estimated by calculating an appropriate moving average. For instance, when data is monthly a 12^{th} order centred moving average can be used to represent how the data develops over a year. Similarly with daily data, we calculate a 7^{th} order moving average to show the trend over a week. Since our data is weekly we would assume to calculate a 4^{th} order centred moving average to aggregate the data over a month. However, this results in a random pattern that cannot be described as a trend. Therefore, we increase the order of the moving average such that more data points are aggregated and thereby smoothing the trendline. This results in a 26^{th} order moving average of our weekly data meaning that we aggregate data points over six months in order to detect a smoothed trend. It is reasonable to use a 26^{th} order moving average since the planning horizon is also 26 weeks. Next, we detrend the data by subtracting the trend from the time series. This leaves us with the seasonality and an error term. The seasonal component is estimated by calculating a seasonal index per week over the two years since the data is weekly. From the plot shown in Appendix B, one can see that there is no seasonality recognizable. Hence, we can conclude to detect a trend the data needs to be smoothed by aggregating it over at least six months otherwise a random pattern is drawn and there is no recognizable seasonality in the data.

2.2.2 Autocorrelated demand data

Erkip et al. [23] find that autocorrelation is important to detect consecutive periods of increased demand. In case of positive autocorrelated demand safety stock needs to be higher to attain the same stock out probability than in case of non autocorrelated demand. Thus, we are interested in investigating whether the observed demand shows autocorrelation.

Autocorrelation of basic type BT1.

The autocorrelation measures the internal dependency of a time series between two time periods [8]. Similarly to the correlation coefficient, which measures the dependency between two variables, the autocorrelation gives a value between -1 and 1 for highly negative and highly positive correlated values, respectively. In contrast, values close to 0 indicate that there is no correlation, that is, time periods are independent.

The autocorrelation for the time series Y at period t and lag k is defined in Equation 2.3. That is, it is the correlation between period t and the kth lagged period [64].

$$r_k = \frac{E[(Y_t - \mu)(Y_{t-k} - \mu)]}{\sigma^2} = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \tag{2.3}$$

When applying the stated definition of autocorrelation, also called serial correlation, a constant mean and variance over time are assumed.

For the time series of the basic type BT1 the variance stays constant over time and we neglect small changes as we compared the variance for the first six lags and only detected small variations. However, as we saw from the time series decomposition in the previous section, there is a smoothed trend recognizable. Hence, we stabilize the mean by de-trending the series before calculating the autocorrelation. De-trending the data is done by calculating a 26 weeks centred moving average and subtracting it from the time series. Next, we calculate the autocorrelation for lags one to eight. This shows no significant autocorrelation between these lags as all values fall below the threshold value of about 0.2. This threshold value is approximated by a hypothesis test, which is further described in Appendix C.1.

We can conclude from the investigation of the autocorrelation for the basic type BT1 that there are no dependencies between time periods and hence they behave independently.

We also assess the generated demand data according to autocorrelation. Note that we elaborate on the procedure of generating the demand in chapter 5. The evaluation of the generated demand data shows that it is neither autocorrelated. Thus, we do not face any problems due to autocorrelation within the data series. Nevertheless, we looked at other products of Infineon which do show autocorrelation, e.g. from ATV and PMM shown in Appendix C.2. These products are Body Power, Powertrain and Radarchips. Hence, when demand is generated the method should be able to create autocorrelated demand.

Stock outs due to autocorrelation.

Showing autocorrelation in solely one of the demand series, either the historical or the generated one, but not reflecting it in the other one can cause problems when defining stock levels. Stock levels may then be either too low leading to an increased stock out probability or too high resulting in larger holding costs.

The problems caused by correlated demands have been studied among others by Erkip et al. [23], Charnes et al. [11], as well as Lau and Wang [40]. Their research differs in the approach how autocorrelation is created. Erkip et al. analyse the effect of autocorrelation for real-world demand of consumer products. Charnes et al. employ an autocovariance function whereas Lau and Wang use an ARMA process to create autocorrelated demand. The research shows that if positive autocorrelation in demand goes undetected and stock levels are set on basis of uncorrelated demand, the actual probability of stock outs will be higher than expected. This can be explained by the continuous increasing demand over multiple periods. In contrast, non autocorrelated demand does not show high demand occurrences over multiple succeeding periods. Similar results have also been found by Fotopoulos et al. [26].

We construct an example ourselves to demonstrate the difference in stock out probabilities for positive autocorrelated and non autocorrelated demand, where we apply autocorrelation in the first four legs. The analysis compares the number of stock outs for both series. A stock out occurs when the demand at time t cannot be satisfied by the production output of time t plus the stock at the end of the previous time period $t-1$. We determine the

production output at time t as the ten weeks moving average of the demand for periods $t-20$ up to $t-10$. We use a period of ten weeks since it approximately reflects the cycle time from the production start in the fabrication to the master storage from which the customer demand is fulfilled. Moreover, we do not consider backlogs, that is, in case of a stock out the stock level is set to 0, but does not become negative. This implies that there are less stock outs than in the case were backlogs are considered. When we consider backlogs the demand at time t includes not only the orders at time t, but also the backlog up to time t. Thus, the overall demand is higher and hence the probability of a further stock out increases. The generated time series having positive autocorrelation is constructed using a sine function and some random error term, more precisely the values are computed by: $demand_t = (1 + factor * sinus(t/4) + random) * 100$, where $0 \leq factor \leq 1$, t are time periods $1, 2, ..., n$ and the random number is drawn from a uniform distribution of range [0,1]. The whole term is scaled by 100 in order to make demand more realistic. One should note, that the higher the factor, that is, the closer it is to one, the higher is the autocorrelation. We choose a factor of 0.3 such that the autocorrelation in the first four lags varies between 0.2 and 0.3 which we consider as realistic. The time series without autocorrelation is constructed by drawing a random number between the minimum and maximum value of the autocorrelated data series such that both series are comparable, that is, they have a similar range and mean. The computation in the Excel spreadsheet can be found in Appendix C.3.

The comparison shows that the number of stock outs for positive autocorrelated demand is approximately 55.48% higher than uncorrelated demand when the stock level is chosen under the assumption of uncorrelated demand. The reason for a higher stock out probability in case of positive autocorrelated demand can be explained by the different behaviour of the data. Positive autocorrelated demand has a high demand over multiple consecutive periods. This stresses the system since it has to provide a high amount of products over consecutive time periods. In contrast, uncorrelated demand behaves randomly such that a week of high demand is usually followed by a week of lower demand. This allows the system to recover.

The drawn results indicate that it is necessary to detect autocorrelation in the demand. Hence, for the simulation study we have to make sure that the generated and historical data behave the same, otherwise we will have distorted results regarding the stocking levels. In the case that the generated data contains positive autocorrelation but the historical data does not, stocks will tend to be too high. That is, the results will suggest higher stocking levels to hedge against a continuing increase of demand over multiple time periods. However, as this continuing increase does not occur stocks will accumulate. In the case that the generated data does not contain positive autocorrelation but the historical data does, stocking levels will tend to be too low. This leads to a shortage of products and a drop in the service level since customer will face longer lead times.

2.3 Conclusion

In this research we consider two exemplary products of the division Chip Card & Security, where one product is a contactbased and the other one a contactless chip. Due to their high volume which accounts for $\geq 25\%$ of CCS's yearly revenue, they are highly relevant. The study about the plan and make process at Infineon in subsection 2.1.1 shows that the production of chips takes up to four month, where approximately three months are allotted to the front end and roughly one month to the back end. That is, in order to stay competitive

and respond quickly to customer demand products have to be produced on forecast. Since the basic types we are considering become customer specific sales products in the sort, production upon forecast is only done up to the master storage where they are still generic. This practice is done to prevent producing highly specific products on forecast. The forecasts, on which the production is based, are made by calculating a four months moving average over the historical data. The further processing from the master storage to the distribution centre is then initiated by customer orders. The current target reach at the master storage is 13 weeks for both products which results in an α-service level of 98%.

From the data analysis based on the deliveries of the two basic types BT1 and BT2 for the years 2014 and 2015 we receive several findings. The time series decomposition shows that there is a slightly upward trend for both basic types, however there is no seasonality recognizable. The upward trend is due to the product lifecycle. These two payment products require a security certificate [31]. The certificate is valid for four years and hence the product lifecycle spans over four years. Renewing the certificate is possible, however it is not very common. That is, the deliveries are likely to increase furthermore, but eventually will level off as the certificate expires by the end of 2018 [31]. In addition, analysis reveals that there is no significant autocorrelation for both basic types. That is, the periods are independent of one another. However, we consider the autocorrelation as an important statistical measure since it indicates whether succeeding time periods show a rising demand. This in turn increases the probability of stock outs as we examine in subsection 2.2.2 and advice to employ another stocking strategy. Indeed it does not apply to our considered products, but there are products of Infineon having autocorrelation. Last, we investigate the correlation among sales products for each basic type. It showed that there is no significant correlation. Hence, the deliveries of the sales products do not influence each another.

Chapter 3

Simulation model

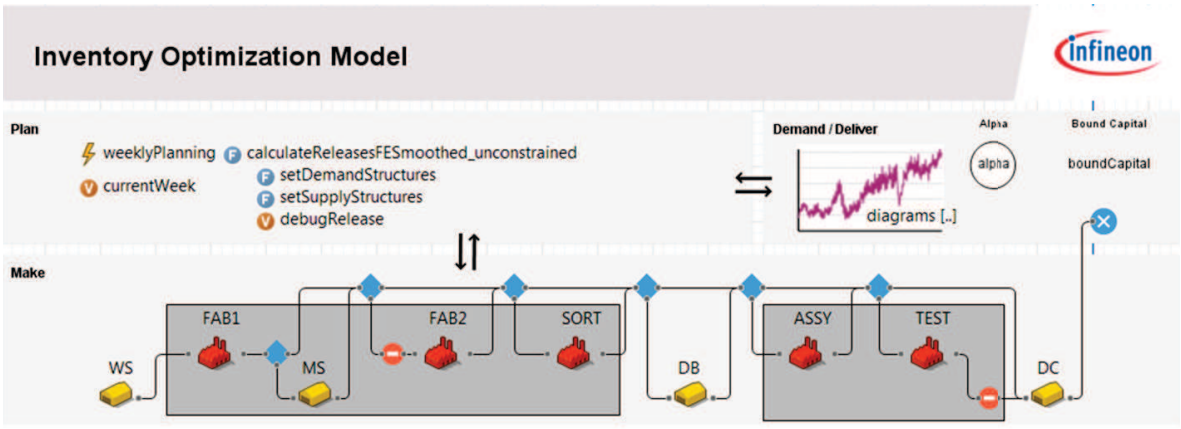
Simulation is a powerful tool when studying various stocking strategies since the impact of various systems settings can be examined in the simulation model first, before applying changes to the real world as these may be too expensive, too dangerous or just impossible. Thus, it offers the opportunity to understand, evaluate, and improve a complex system without taking high risks.

The existing simulation model is used to define the best stocking strategies for the two basic types and to consider various approaches to determine the release quantity. The release quantity is defined as the amount of wafers to be released in production. The model resembles the production process as described in subsection 2.1.1. It was developed by Infineon's flexibility & econometrics team and further enhanced by a master thesis [13]. As part of the thesis it also got validated. In our project we implement minor modifications such as various production release approaches. The model is built with the Anylogic software version 7.3.5 professional. Anylogic is a java-based simulation tool which enables discrete event simulation, system dynamics, and agend based modelling. The model under consideration is a discrete event simulation. It is called *discrete* since the state, which are the variables that describe the system, changes at separate points in time. At these points in time an *event* is triggered. Events are instantaneous occurences that may result in a changed state of the system. In contrast, the state of a continuous system changes constantly with respect to time, for example the state such as the position of a flying airplane is changing with respect to time [41].

At Infineon we address four simulation levels. These levels are distinguished by their scope of the considered supply chain steps. At a high level the end-to-end supply chain is modelled whereas at a low level single workcentres (machines) are described. The four levels are (top-down):

- Level 4: The end-to-end supply chain from the supplier to the customer.
- Level 3: The internal supply chain of Infineon. It is part of the end-to-end supply chain.
- Level 2: A single factory or single site of the internal supply chain.
- Level 1: A single workcentre of a factory.

We model the internal supply process of two basic types illustrated by Figure 3.1. Thus, we can class our model with level 3. It can be divided into three main parts: Plan, Make, and Data (input and output of simulation) which correspond to section 3.1, section 3.2, and section 3.3.



storage locations		processing steps	
WS	wafer start	FAB 1(2)	fabrication 1(2)
MS	master storage	SORT	sort
DB	die bank	ASSY	assembly
DC	distribution centre	TEST	test

Figure 3.1: Snippet of the graphical user interface of the simulation model built with anylogic

3.1 Plan functions

3.1.1 Release quantity

The planning covers the decision of how many products to be started into the production, at the wafer start (WS), the master storage (MS), the die bank (DB), and the distribution centre (DC). It matches the demand with the supply and releases the according quantity. Figure 3.2 illustrates the planning at the stock points. The release quantity in each period t at the stock points is determined as follows:

$$\begin{aligned}
\text{release}_{WS} &= \text{FC}_{CT_{WS}} + (\text{target reach}_{MS} - \text{stock}_{MS}) + (\text{target reach}_{DB} - \text{stock}_{DB}) \\
&\quad + (\text{target reach}_{DC} - \text{stock}_{DC}) + \text{backlog} - \text{WIP}_{FE} - \text{WIP}_{BE} \\
\text{release}_{MS} &= \text{Order}_{CT_{MS}} + (\text{target reach}_{DB} - \text{stock}_{DB}) \\
&\quad + (\text{target reach}_{DC} - \text{stock}_{DC}) - \text{WIP}_{FE} + \text{backlog} - \text{WIP}_{BE} \\
\text{release}_{DB} &= \text{Order}_{CT_{DB}} + (\text{target reach}_{DC} - \text{stock}_{DC}) + \text{backlog} - \text{WIP}_{BE} \\
\text{release}_{DC} &= \text{Order}_{t} + \text{backlog}
\end{aligned}$$

The amount of wafers to be started into front end ($release_{WS}$) is determined by the forecasted demand during the cycle time ($\text{FC}_{CT_{WS}}$) from start in production to the finished product, the target reach at each of the downstream stock points netted by the actual stocks as well as the current backlog and the work in process in front end (FE) and back end (BE). Similarly, the release quantity at the master storage is dependent on the incoming orders during the cycle time from the master storage to the customer (CT_{MS}), the netted target stocks at the downstream stock points, the backlog as well as the downstream work in

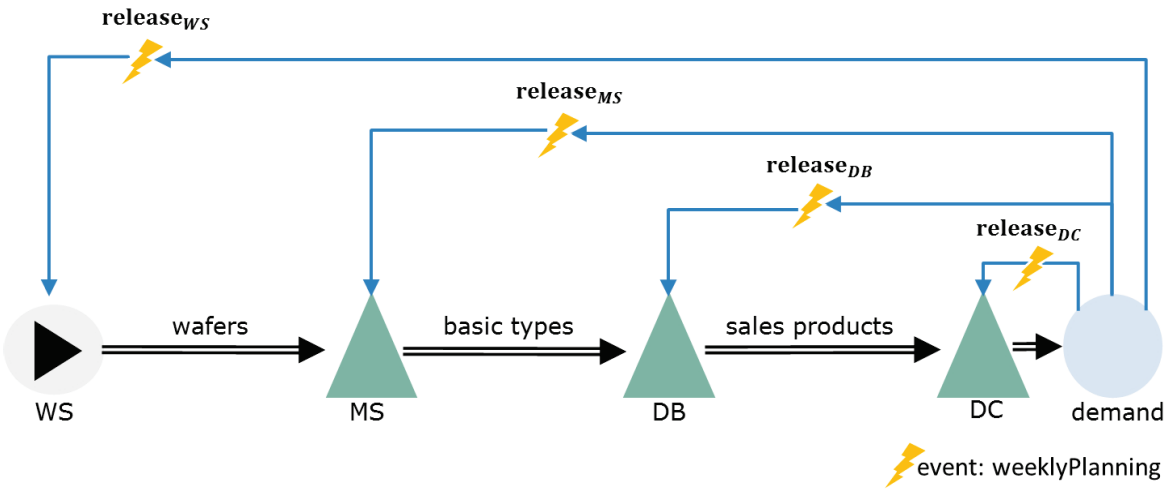


Figure 3.2: Weekly planning of release quantities at the stock points in the simulation model

process. The release quantity at the die bank considers the incoming orders during the cycle time from the die bank to the customers (CT_{DB}), the netted stocks at the distribution centre, the backlog and the work in process in back end. Last but not least, at the distribution centre the orders of period t and the backlog are released.

3.1.2 Demand generation

In the following it is explained how the demand generation method works in general. We discuss in section 5.2 how the parameters are determined such that it applies to the two basic types. To clarify, the parameters are entered manually in an internal database file of the simulation before the run is started. Further note, the simulation model creates a forecast, which we call simulation forecast, however this forecast should not be confused with the marketing forecast. It is important to note, that when we simulate the current practices in chapter 6, we do not use the simulation forecast to predict the demand of the two basic types, but instead use a moving average over the historical data (marketing forecast). Nevertheless, we describe it since it is part of the demand generation method. We give a definition in the description below.

In Figure 3.3 we schematically show the logic of the demand generation. In week one the method creates demand for each defined sales product over the next 26 weeks by a truncated normal distribution. A horizon of 26 weeks is chosen, since the short term planning at Infineon covers a time frame of 26 weeks. The calculations is defined below:

$$\begin{aligned} value &= roundToInt[normal(min, max, \mu, \sigma, random)] \\ \text{with } min &= 0 \\ max &= 2 * average\ demand \\ \mu &= average\ demand \\ \sigma &= cv * average\ demand \end{aligned}$$

A value from a truncated normal distribution is drawn with minimum 0 since the demand cannot become negative, and a maximum of two times the average demand of the sales

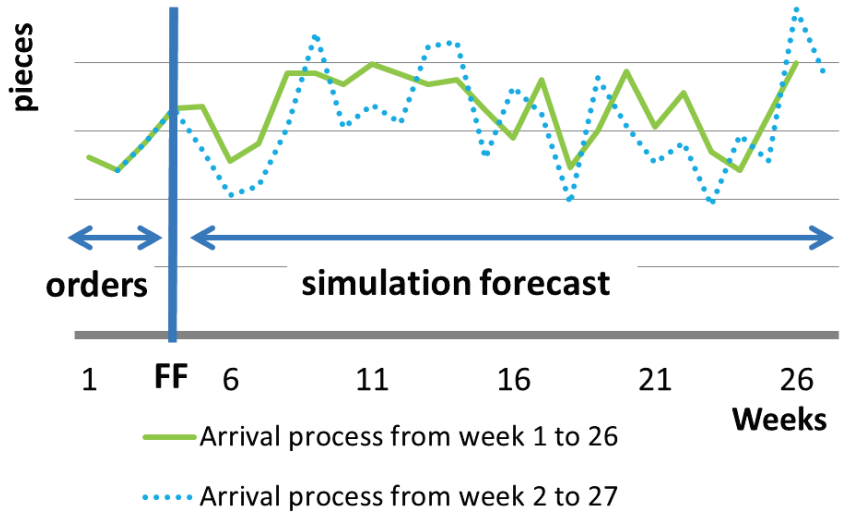


Figure 3.3: Illustration of the demand generation in the simulation model

product, a mean μ equalling the average demand as well as a standard deviation σ. In week two all values in the period between the freeze fence (FF) and the defined horizon are modified according to some method and a new demand point for week 27 is created. We call the period from the freeze fence to the defined horizon simulation forecast. Demand during this period is changing each week. On the other hand, the demand from the current week to the freeze fence are known incoming orders which do not change. Hence, the freeze fence sets the time horizon for which the demand stays constant. That is, for the two basic types we set the freeze fence according to the cycle time of the master storage (CT_{MS}) to imitate that orders are known at the master storage.

The calculations for the demand adjustment given below are such that demand points lying further afar in the future are varying more than demand points closer in the future. This reflects the increasing uncertainty illustrated in Figure 3.4. The adjustment for each forecast point $\in [FF; t+26]$ in week t is calculated as follows:

$$biasSigma = (AAMax - AAMin) * \frac{(t-1)^{AAN}}{(26-1)^{AAN}} + AAMin$$

$$value = roundToInt[normal(\sigma\prime * biasSigma, \mu\prime, random)] * previousForecastPoint$$

$$if\,(previousForecastPoint + value < 0)$$

$$value = -previousForecastPoint$$

where the parameters $AAMax$ and $AAMin$ describe the changes in the furthest and closest horizon, respectively and AAN controls whether $biasSigma$ is linear ($AAN = 1$), concave ($AAN < 1$), or convex ($AAN > 1$). Furthermore, $\mu\prime$ and $\sigma\prime$ describe the change in the mean and standard deviation. In section 5.1 we further examine these parameters.

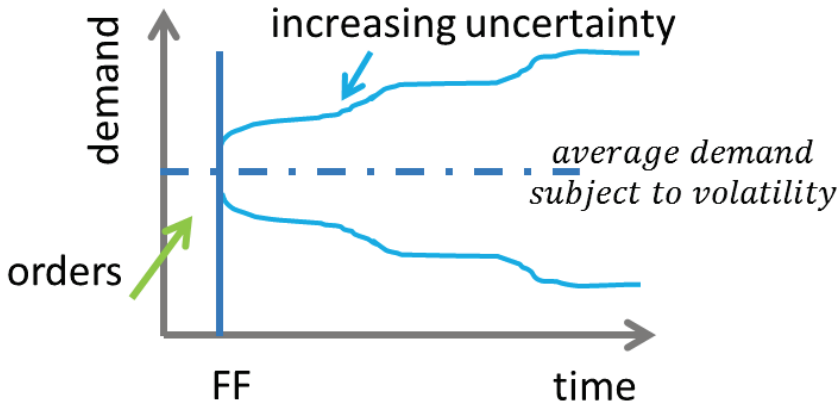


Figure 3.4: Illustration of the increasing uncertainty range over the simulation forecast

3.2 Make functions

The make part of the simulation corresponds to the make process described in section 2.1.1. The steps are represented by the yellow and red icons in Figure 3.1. The yellow icons represent stocking positions and the red icons represent processing steps. We have four stocking positions, namely the wafer start (WS), the master storage (MS), the die bank (DB), and the distribution centre (DC). However note, the wafer start represents the source to the process. The stocking positions are modelled by queue elements where the products remain until they are processed by the next facility. Moreover, we have five facilities where the processing of the products takes place. These are fabrication 1 (FAB1), fabrication 2 (FAB2), sort, assembly (ASSY), and test. They are modelled by delay elements representing the processing time at each facility. Currently, there are no capacity restrictions for the processing steps. The fabrication is split up into fabrication 1 and fabrication 2 in case that the master storage lies inbetween the steps of the fabrication. This can be set up by splitting the total cycle time of the fabrication between fabrication 1 and fabrication 2 accordingly. In our case, the cycle time of the fabrication is completely allotted to fabrication 1 as the master storage lies at the end of this process. Defining the cycle times for the facilities belongs to the data part of the simulation.

3.3 Input and output data

The data of the simulation model comprises both, the input and output data. The input data allows to set up various system settings which are stored in internal database files. These are entered manually before the simulation run. The output data gives the α-, β-, and γ-service level as well as the total costs defined below in Equation 3.2, Equation D.1, Equation D.5, and Equation 3.4. They are not only shown in diagrams during the simulation run but can also be exported automatically to an Access database at the end of the simulation in order to elaborate on the data analyses.

3.3.1 Experimental and system settings

The input parameters stored in internal database files are distinguished in system and experimental settings. System settings are basic type specifics and experimental settings are factors we alter for each experiment to study the system behaviour. Relevant system settings are:

- Freeze fence: The number of periods from now onwards into the future where demand does not get modified.
- Average sales price. It is used to calculate the costs.
- Profit margin. It is also used to calculate the costs.
- Front end cost share: The split of costs allotted to front end and back end.
- Weighted cost of capital: It represents the costs of investing capital in products stored at master storage, die bank or distribution centre.
- Cycle time: The number of days the processing takes at the fabrication, sort, assembly, and test.
- Product structure. It specifies the diversification point.

The experimental settings concern the target reach at the master storage, die bank, and distribution centre. They define the number of weeks that stock should be available at each storage location. We vary these three parameters in the simulation study in chapter 6 to evaluate various stocking strategies.

3.3.2 Key Performance Indicators

The evaluation of the stocking strategies and production release approaches in section 6.3 is done according to the α-service levels as well as the total costs (TC). The service level indicates how well we satisfy customer orders. That is, whether we deliver on time and can meet customer demand. The KPIs are evaluated over several weeks ($t = 1...T$) and various simulation runs ($r = 1...R$). The service levels are measured at the end of the supply chain, that is, each week t the incoming orders and outgoing deliveries are compared. Notice, in the simulation model week t corresponds to time period t.

α-service level.
The α-service level gives the probability that the incoming demand during period t is completely met by on-hand inventory. It either becomes 0% when demand is not met completely or it becomes 100% when demand is met completely. Note, that backorders of previous weeks are not considered as part of the incoming demand for period t [60]. The α-service level is defined by [60]:

$$\alpha\text{-sevice level} = P\{\text{demand during time period t} \leq \text{on-hand inventory} \tag{3.1}$$

$$\text{at beginning of time period t}\} \tag{3.2}$$

It is implemented in the simulation model by aggregating over all sales products ($p = 1...P$), weeks ($t = 1...T$), and replications ($r = 1...R$):

$$\alpha\text{-sevice level} = \frac{\sum_{r=1}^{R}\sum_{p=1}^{P}\sum_{t=1}^{T}\alpha_{rpt}}{R * P * T} \tag{3.3}$$

Next to the α-service level, the simulation model also provides the β- and γ-service level. For an detailed explanation we refer to Appendix D since we do not use them to evaluate the approaches. Nevertheless, they may be used in addition to the α-service level.

Costs.
Next to achieving a high service level, we aim to keep the total costs considerable low. The TC are the fixed capital costs bind in products. It is calculated by the weighted average cost of capital (wacc) times the value of the WIP and the stock locations:

$$TC = \text{wacc} * (\text{value of WIP} + \text{value of stock points}) \tag{3.4}$$

In the simulation model the calculation of the costs is divided into two task: the distribution of the total bound capital along the facilities and the weekly calculation of the total value in the supply chain [13].

The distribution of the costs along the facilities is done top down. First, costs are split proportionally between front end ($cost^{FE}$) and back end ($cost^{BE}$) according to the 'front end cost share' and then further divided between the facilities. The costs are based on the cost per unit ($cost^{unit}$) which is defined by the difference between the average selling price (asp) and its profit margin:

$$cost^{unit} = (1 - margin)asp \tag{3.5}$$
$$cost^{FE} = cost^{unit} * costshare^{FE} \tag{3.6}$$
$$cost^{BE} = cost^{unit} * (1 - costshare^{FE}) \tag{3.7}$$
$$\tag{3.8}$$

The costs at the facilities FAB and sort in front end as well as at assembly (ASSY) and test in back end are assigned according to the cycle time (CT) of each processing step denoted by $CT^{\text{processing step}}$. Moreover, the costs of the fabrication are split between fabrication 1 (FAB1), incurred before the product enters the master storage, and fabrication 2 (FAB2), incurred after the master storage and before the sort step:

Costs in front end:

$$cost^{FAB} = cost^{FE}\left(CT^{FAB}/CT^{FE}\right) \tag{3.9}$$
$$cost^{FAB1} = cost^{FAB}\left(CT^{FAB1}/CT^{FAB}\right) \tag{3.10}$$
$$cost^{FAB2} = cost^{FAB}\left(CT^{FAB2}/CT^{FAB}\right) \tag{3.11}$$
$$cost^{SORT} = cost^{FE}\left(CT^{SORT}/CT^{FE}\right) \tag{3.12}$$

Costs in back end:

$$cost^{ASSY} = cost^{BE}\left(CT^{ASSY}/CT^{BE}\right) \tag{3.13}$$
$$cost^{TEST} = cost^{BE}\left(CT^{TEST}/CT^{BE}\right) \tag{3.14}$$

The second tasks considers the weekly calculation of the value in the supply chain. It is calculated by the average value of WIP and inventory at all facilities and stock points denoted by $value^{SC}$. The WIP value at each facility is calculated as the WIP units times the average cost of each unit at the respective facility. The same procedure is done for the inventory value:

the total stock units times the cost of the units at the respective stocking point. The costs at the stock points are the accumulated costs of the previous facilities. The total costs (TC) are the value of goods in the supply chain times the the weighted average cost of capital (wacc), which describes the rate of return that could have been earned when investing elsewhere. We define the TC over all replications and weeks as follows:

$$TC = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T} wacc * value_{RT}^{SC}}{R * T} \tag{3.15}$$

Note, that we do not consider idle costs as input of the total costs since we concentrate on the trade off between keeping enough products in stock at the various locations and the costs of storing these products.

3.4 Conclusion

To improve the supply chain planning process we use discrete event simulation since it is able to capture complex relations between processes. The existing model represents the supply chain of Infineon with its plan and make process. The plan functions are responsible for starting production whereas the make functions concern the processing steps. This model can be used to study various stocking strategies and production release approaches. The evaluation of the various procedures is done by considering the respective service level and costs. However, before we use the model for our simulation study we need to parametrize it according to the situation of the two exemplary basic types. This includes to generate demand similar to the observed one. In chapter 4 we study how we can assess the fit between the generated and observed demand.

Chapter 4

Literature review

The aim of this thesis is to improve stocking levels of two basic types as well as determining the release quantity in front end by using discrete event simulation. In order to receive precise results, the demand arrival process has to generate demand data that captures the stochastic behaviour of the observed demand data. The closer the generated demand data is to the observed data the more accurate are the simulation results. Thus, we are interested in a measure that assesses the fit between the generated and observed data. In the literature, there exist various measures to compare time series. One can compare two time series using forecast accuracy measures explained in section 4.2. Forecast accuracy measures evaluate the residual, also called error term, between forecasted and observed values. Another way of evaluating the similarity between two time series are time series similarity measures, which we elaborate in section 4.3. Similarities can be based on the distance, on features of the two series, or on the shape of the series. Last, in section 4.4 we consider hypothesis tests or also called 'goodness of fit' tests to assess the fit between two data series. One can distinguish these tests between one-sample and two-sample tests. One-sample tests evaluate whether an empirical distribution is drawn from some known distribution whereas two-sample tests evaluate whether two empirical distributions are drawn from the same unknown underlying distribution. In this case, we speak of consistent data. We are interested in the two-sample tests to compare the generated and observed demand and evaluate whether they are consistent. Before we discuss the several methods, we give some theoretical background in section 4.1 on how demand data can be classified, we provide definitions for time series as well as stochastic processes, and introduce two basic concepts of forecasting demand.

4.1 Introducing common terms and concepts

4.1.1 Categorization of demand patterns

To start with, we give a well-known classification of demand patterns according to Syntetos & Boylan [59], which we refer to in the remainder of this thesis. Other classification approaches exist, for example Williams [63] partitions the demand according to its variance. Eaves [18] revises this classification scheme and proposes a categorization based on the demand size variability as well as lead time variability. However, the cut off values of these approaches apply only to the particular empirical situation under study. This is not the case for the classification of Syntetos & Boylan. They show that their proposed cut off values can be used for any empirical demand data, which they classify into four categories: *smooth*, *intermittent*,

erratic, and *lumpy*. The categorization is based on the squared coefficient of variation (CV^2) as well as the average inter-demand interval (ADI):

- Smooth demand is quite stable with rather low fluctuations and demand occurring in most of the periods: $CV^2 \leq 0.49$ and $ADI \leq 1.32$
- Intermittent demand is also quite stable, but demand occurs rarely: $CV^2 \leq 0.49$ and $ADI \geq 1.32$
- Erratic demand is characterized by large fluctuations in demand and regular demand occurrences: $CV^2 \geq 0.49$ and $ADI \leq 1.32$
- Lumpy demand has also large fluctuations in demand, but rather few demand occurrences: $CV^2 \geq 0.49$ and $ADI \geq 1.32$

Figure 4.1 summarizes the cut of values and according categories of the demand classification.

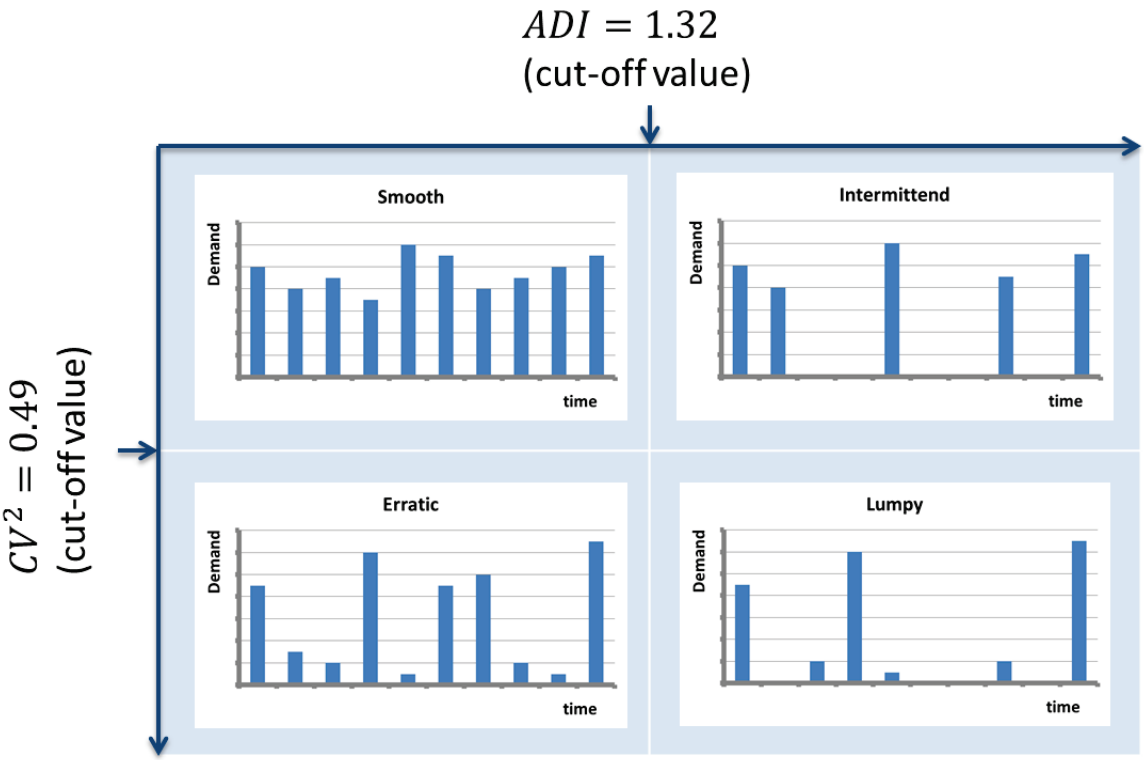


Figure 4.1: Categorization of demand patterns according to Syntetos&Boylan [59]

4.1.2 Time series and stochastic processes

A time series is a collection of observations made sequentially in time [8,22]. If the observations are made at fixed points in time $t_1, t_2, ..., t_N$, the series is said to be discrete, otherwise, if the observations are made continuously over time, the time series is said to be continuous. The time series we are analysing are discrete since we consider the demand at fixed points in time.

Furthermore, a time series can be described either as deterministic or as stochastic. A time series is deterministic if its future values can be determined exactly by some mathematical function. That is, future values are known and no randomness exists. For example they can be calculated by $z_t = 2 * sin(\pi t)$. On the other hand, a time series is stochastic if its future values are described by a probability distribution, that is future values are not known in advance and are said to be non-deterministic [8]. Obviously, our data is stochastic.

A stochastic process is a collection of random variables which describe a random phenomenon evolving over time. Thus, a stochastic process can be described by a set of random variables $(Z_1, Z_2, ..., Z_N)$ and their corresponding probability distribution $p(Z_1, Z_2, ..., Z_N)$ for representing the different values at different times. Then, a non-deterministic time series can be viewed as a realization of a stochastic process [8]. We distinguish between stationary stochastic processes and non-stationary stochastic processes. As mentioned earlier in section 2.2.1, a stationary process is based on the assumption that the joint probability distribution of the observations does not change when shifting all the observations forward or backward in time [8]. This is, that the mean μ_{z_t} and variance σ_{z_t} for a stationary process is the same for all random variables $Z_1, Z_2, ..., Z_N$. The assumption is not met by non-stationary processes. That is, the mean and the variance of non-stationary processes change over time. In case there is an upward (or downward) growing trend in the data, the process is non-stationary since the mean increases (or decreases) over time. Stationarity of the data is for example important when we consider autocorrelation. For the autocorrelation we need the data to be stationary otherwise it shows the increasing (or decreasing) trend but one does not know whether the series itself is autocorrelated. Therefore, the trend needs to be removed such that the series becomes stationary.

4.1.3 Basic forecasting techniques

The production in front end is started in advance. Thus, we need a forecast to determine the release quantity. There exists a variety of methods to create a forecast ranging from basic to advanced techniques. We are concentrating on basic techniques since we will use these to determine the production release quantity in subsection 6.1.1. Forecasting methods based on smoothing can be distinguished into averaging and exponential smoothing methods [22, 46]. Averaging methods rely on equally weighted observation. In contrast, exponential smoothing methods rely on an unequal set of weights, where more weight is given to observations lying closer in the series and less weight is given to observations lying farther away in the series. These weights are exponential decreasing as the observation gets older [22].

Averaging methods.
The mean and the simple moving average are two basic techniques to forecast the next value where each observation is weighted equally. The mean takes the average over all observations n and the simple MA takes the average over the last k observations. Using a simple MA, we can write the forecast for the next value F_{t+1}^{MA} in period $t+1$, which relies on the previous observations Y_i, as follows [22]:

$$F_{t+1}^{MA} = \frac{1}{k} \sum_{i=t-k+1}^{t} Y_i \tag{4.1}$$

A new forecast point is calculated each time a new observation becomes available. Increasing the order k the forecast becomes more smooth. Note, the simple moving average of order

$k = n$ is just the mean and the MA of order $k = 1$ is the naive forecast [22].

Exponential smoothing methods.
On the other hand, in exponential smoothing methods there are one or more smoothing parameters which determine the weights assigned to the observations. Methods are among others SES, Holt's Linear Exponential Smoothing, and Holt's Trend and Seasonality method [22, 46]. SES is the simplest of the exponentially smoothing methods which we will apply in subsection 6.1.1. Using SES, we can write the forecast F_{t+1}^{SES} for period $t + 1$ in terms of the observed point Y_t and its forecasted value F_t for period t where we assign the weights $\alpha \in [0, 1]$ and $(1 - \alpha)$ [22, 33, 46]:

$$F_{t+1}^{SES} = \alpha Y_t + (1 - \alpha)F_t \tag{4.2}$$

This can be written as:

$$F_{t+1}^{SES} = F_t + \alpha(e_t) \tag{4.3}$$

where $e_t = Y_t - F_t$ is the forecast error in period t. Thus, the forecast represents a weighted moving average of the previous observed value Y_t adjusted by the error of the last forecast e_t. When α has a value close to 1, the new forecast includes a substantial adjustment for the error in the previous forecast. Usually, α is chosen in such a way that the applied forecast accuracy measure is minimized [22].

4.2 Forecast accuracy measures

In order to evaluate a forecast one commonly uses a suitable forecast accuracy measure to compare the difference between the forecast and observed data. In our case, we do not assess the forecast data, but the generated data by the simulation model. We describe how one usually proceeds when using forecasting methods to create data and assess its accuracy to the observed data.

Forecast accuracy measures generally assess how well a forecasting method predicts actual data. Usually, the observed data is partitioned into a training set, called 'in-sample-data', and a test set, called 'out-of-sample-data'. The training set is used as input for the forecasting method such as a MA and SES in order to create forecasting data, afterwards the generated data is compared to the test set according to some accuracy measures [22, 33]. These measures can be distinguished in scale-dependent and scale-independent measures. As the name suggests, the scale-dependent measures rely on the scale of the data. That is, if the data is for example given in million of pieces then the measure is also given in million of pieces. On the other hand, the scale-independent measures do not rely on a scale. They are either described in percentage or given as relative measure. Relative measures compare a method with some benchmarking method. First, we explain scale-dependent measures and continue later in this section with scale-independent measures.

4.2.1 Scale-dependent measures

Let y_i be the ith observation and $\hat{y}_i$ denote a forecast of y_i. The forecast error is simply $e_i = y_i - \hat{y}_i$, which is scale-dependent as it is on the same scale as the data [33]. Commonly used and intuitive scale-dependent forecast accuracy measures are the mean squared error

(MSE), the mean absolute error (MAE), and the root mean squared error (RMSE) [22,33,57] defined below:

$$MSE = mean({e_i}^2), \qquad i = 1, ..., N \tag{4.4}$$

$$MAE = mean(|e_i|), \qquad i = 1, ..., N \tag{4.5}$$

$$RMSE = \sqrt{mean({e_i}^2)}, \qquad i = 1, ..., N \tag{4.6}$$

The MSE computes the mean over the squared difference between the observed and forecasted data. The MAE computes the mean over the absolute difference between the observed and forecasted data and the RMSE takes the root of the mean over the squared differences. These measures have in common that they are easy to understand and to compute. The advantage of using absolute or squared values is that negative and positive values do not offset each other [32]. They are suitable when one has a feeling for the magnitude of the data since these measures are given on the same scale as the data or for the MSE the scale is even larger since one computes the squared difference. Therefore, the RMSE is often preferred to the MSE since it takes the root and hence is on the same scale as the data [34]. However, as these measures are scale-dependent one cannot compare them among data sets with different scales. For example, if we have two data sets with MSE_1 and MSE_2, respectively and we compute $MSE_1 > MSE_2$, one would usually conclude that the forecast for the second data set is more accurate than the forecast for the first data set since the MSE_2 is smaller than the MSE_1. But one has to pay attention since the first data set may be in million of pieces and the second data set may be in thousands of pieces. In that case, one cannot conclude that the forecast for the second data set is better than the forecast for the first data set as the MSE are on different scales and hence difficult to compare. Thus, when comparing measures of data with different scale, one should use scale-independent accuracy measures.

4.2.2 Scale-independent measures

In order to compare forecast accuracy measures of data sets with different scales one can use percentage errors which are scale-independent. The percentage error is given by $p_i = \frac{100e_i}{y_i}$. Common percentage errors are the mean absolute percentage error (MAPE), and the symmetric mean absolute percentage error (SMAPE) [33,34] as defined below:

$$MAPE = mean(|p_i|), \qquad i = 1, ..., N \tag{4.7}$$

$$SMAPE = mean(200 * \frac{|y_i - \hat{y_i}|}{y_i + \hat{y_i}}), \qquad i = 1, ..., N \tag{4.8}$$

The MAPE computes the mean over the absolute percentage error and the SMAPE computes the mean over the absolute difference between the observed and forecasted value divided by the sum of the observed and forecasted values and multiplies this fraction by 200. The advantage of these percentage errors are their scale-independence. Hence, the forecast accuracy can be compared among data sets with different scales. But, they have the disadvantage that if values are very low a small deviation between the observed and forecasted data leads to a high percentage error. In case that the observed value is one unit and the forecasted value are two units the MAPE is 100% [57]. In some cases a deviation by one is significant, in other cases a deviation by one is not significant. When we think of a product that has low production cost, then producing one or two of these products is not a high risk. On the other

hand, when the production costs are very expensive we risk a lot of capital when producing two products where only one is needed. A second disadvantage is that the percentage error is not defined when demand is zero [33, 57]. Thus, in case of lumpy demand where there are many periods with zero demand percentage errors are not suitable [32].

Furthermore, Makridakis [44] argued that for the same error value e_i MAPE puts a heavier penalty on positive errors, that is if values are overforecasted, than on negative errors, that is if values are underforecasted, due to the different value y_i in the denominator. For example, in the case of overforecasted values with observed value $y_i = 100$ and forecasted value $\hat{y}_i = 150$, we have an error of $e_i = 50$ and we receive a MAPE of $MAPE = 100 * \frac{50}{150} = 50\%$. In the case of underforecasted values with an observed value of $y_i = 150$ and a forecasted value of $\hat{y}_i = 100$, we still have an error term of $e_i = 50$, but the MAPE results in $MAPE = 100 * \frac{50}{100} = 33.33\%$ since y_i differs. Therefore, he proposed the 'Symmetric' MAPE [44]. However, SMAPE is not as symmetric as its name suggests according to Goodwin and Lawton [28]. For the same value of the observed data y_i, SMAPE puts a heavier penalty on lower forecasts than on higher forecasts. For example, if we have an observed value $y_i = 100$, a lower forecast $\hat{y_l} = 50$ and a higher forecast $\hat{y_h} = 150$, then the SMAPE of the lower forecast is $SMAPE_l = 200 * \frac{50}{150} = 66.67\%$ and of the higher forecast it is $SMAPE_h = 200 * \frac{50}{250} = 40\%$. Further, SMAPE is difficult to interpret since on the one hand it has no lower bound, that is it can become negative in case that forecasted values can take on negative values even though it is an 'absolute percentage error' and on the other hand its upper bound is 200% making comparisons with alternative percentage errors ambiguous [32, 34]. As we generate demand data, we do not have negative values as demand is always ≥ 0, and hence in our case the SMAPE is bounded from below by 0. Nevertheless, the upper bound is still 200% and thus makes comparison difficult.

Infineon uses a modification of SMAPE, also called SMAPE 3, which we shortly discuss here. A definition is given in Equation 4.9, where y_i is the observation and $\hat{y}_i$ is the forecast point in week i:

$$SMAPE3 = \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i + \hat{y}_i} \tag{4.9}$$

It does not multiply the fraction by 200 and thus, its results range from 0% to 100% making it easier to interpret and compare with other percentage errors. Further, even though it cannot completely eliminate the upward bias, it alleviates the problem when forecasts are low compared to when forecasts are high by summing the denominator and nominator separately [35]. At Infineon, the problem of an undefined error when demand is zero is avoided by not considering these periods. However, this distorts the information since the SMAPE is not updated accordingly. In case there was an under- or overforecast the SMAPE would decrease, in case of a correct forecast the SMAPE would improve.

Further scale-independent measures include measures based on relative errors and relative measures [34]. These accuracy measures use a benchmark method, most often the naive forecasting method, which takes the observed value of the previous period as forecast for the succeeding period, to compare various approaches. Measures based on relative errors divide the error e_i obtained from the examined forecasting technique by the error e_i^* obtained from the benchmark method. Thus, the relative error is denoted as $r_i = e_i/e_i^*$. Then we can define

the accuracy measures as [32, 34]:

$$\text{Mean Relative Absolute Error}(MRAE) = mean(|r_i|)), \quad i = 1, ..., N \tag{4.10}$$

$$\text{Geometric Mean Relative Absolute Error}(GMRAE) = gmean(|r_i|), \quad i = 1, ..., N \tag{4.11}$$

The MRAE computes the mean over the absolute relative error and the GMRAE computes the geometric mean over the absolute relative error. The geometric mean is defined as the nth root of the arithmetic product of the values $1, ..., n$, it can be written as $\sqrt[n]{r_1 * r_2 * ... * r_n}$. The disadvantage of these measures is that the error obtained from the benchmark method can be very small or zero as with intermittent or lumpy demand and in that case the error is not defined since it would involve a division by zero [32].

Instead of using measures based on relative errors, one can also use relative measures. A relative measure compares two forecasting methods by dividing some accuracy measure of the forecast method of interest by the same accuracy measure for the benchmark method. Let MAE denote the measure of the method of interest and MAE_b denote the measure of the benchmark method, then the relative measure is given by $RelMAE = MAE/MAE_b$. Similarly, one can define relative measures using MSE, RMSE or MAPE. When the RMSE is applied, we speak of Theil's U statistic. The interpretation of relative measures is straightforward. A value < 1 indicates that the method of interest performs better than the benchmarking method, and vice versa, a value > 1 indicates that the benchmarking method gives more accurate values [22, 34]. However, relative measures can only be applied when comparing forecasting methods, for example when one is interested whether forecasting method 1 or forecasting method 2 provides more accurate results. Thus, they are not applicable when the out-of-sample forecast accuracy is measured where forecasts are generated by applying only one method.

The introduced measures are all point estimates that is they compare the observed value at time t with the forecasted value at time t. Thus, one tries to predict the actual value exactly, but does not account for the overall behaviour of the data. However, as we are not interested in the exact values but rather want to capture the overall behaviour of the data such as short upward trends that may lead to stock outs, we further consider time series similarity measures to assess the fit between the generated and observed data as well as hypothesis tests that evaluate whether two samples come from the same underlying distribution.

4.3 Time series similarity measures

Time series similarity measures evaluate the (dis)similarity between time series. Thus, they try to recognize similar objects even though they are not mathematically identical [24]. Historically, the research for similarity measures has been done mainly by the pattern recognition and data mining commmunity, where similarity measures are used as criterion for indexing and clustering of time series in databases [27, 39].

Various categorizations of time series similarity measures exist in literature. In [17], Ding et al. differentiate the measures into four categories: Lock-step, elastic, treshold-based and pattern-based measures. Another categorization done by Esling and Agon in [24] proposes the four categories: shape based, edit based, feature based and structure based similarity measures. For our further discussion of various similarity measures we use the classification of Esling and Agon [24] which we consider as most suited. We do not further consider Lock-step measures since they are distance measures which compare the ith point of one time

series with the ith point of another time series [17] and thus, have the same disadvantages for our case as the considered forecast accuracy measures. That is, they compare specific values and do not account for the overall behaviour. However, as we consider the time series as realizations of a stochastic process we allow for inequalities in the ith point.

The most frequently used shape based similarity measure is dynamic time warping (DTW). It was introduced by Berndt and Clifford [7] and applied to speech recognition as well as data mining. Other than the lock step measures, DTW allows a time series to be 'streched' or 'compressed' in time to better match with some other time series. Hence, when comparing two time series where one of them is the same as the other one but shifted forward or backward in time, DTW will give a perfect fit for these two series. There exist a variety of extensions to the original DTW. Ding et al. [17] propose to constrain the warping window size which improves the accuracy for measuring time series similarities and speeds up computation.

Edit based similarity measures are based on the idea to find the minimum number of operations that transform one time series into the other one by applying insertion, substitution and deletion [24]. Historically, these measures were used to show the difference between strings. The Longest Common SubSequence (LCSS) distance using the model proposed of Andre-Joesson and Badal [4] is a commonly known edit based similarity measure. A treshold parameter ϵ is introduced. In case that the distance between two points from two time series is less than the threshold value ϵ they are considered to match. Extensions to this measure are made by Vlachos et al. [62]. He adds a warping treshold γ which constrains the matching of the points such that the temporal dimension is considered. Another commonly used edit distance measure is the Edit Distance on Real sequence [12], which also introduces a threshold value ϵ similar to the LCSS. Contrary to LCSS, they assign penalties according to the length of gaps between two matched pairs.

Feature based similarity measures compare selected features such as coefficients from discrete fourier transform for both time series. Vlachos et al. [62] extract period features in order to compare time series among each other. Thereby, they aim to detect and monitor structural periodic changes and determine similarities between time series by considering the periodicity. Therefore, a two-tier approach is developed which considers the information in both the autocorrelation and the periodogram [62]. Janacek et al. [38] propose a likelihood ratio statistic for discrete fourier transform coefficients to test the hypothesis of difference between series. Therefore, they take the fast fourier transforms and base the distance metric on the differences in the periodogram of the series.

Last, we shortly look at structure based similarity measures which are especially applicable for longer series. Compared to the other introduced measures they do not try to find local similarities between patterns but rather look at a higher scale to identify global similarities in series. Therefore, they incorporate prior knowledge about the data generated process [24]. That is, they determine the similarity by considering whether one series comes from the same underlying model such as ARMA than the other one.

Most of the introduced time series similarity measures are based in some way on fourier transforms, that is they take the periodicity of a series into account. These measures are not suitable as our data does not show periodicity. In addition, structure based measures consider the similarities of the underlying model. As we do not have any knowledge of the underlying model, they are not applicable. Others are based on the distance between two series either by transforming one series into the other one with a minimum number of operations, or by considering the exact shape of a series. Dynamic time warping considers the distance between two time series by comparing the ith point of one series with all points of the other series.

Hence, it measures the exact distance but allows for time shifts, however not for randomness in the data. On the other hand, edit based distance measures such as the LCSS and EDR allow for local time shifts and noise in the series by introducing a threshold value. Thus, they may be applicable for comparing the time series.

4.4 Hypothesis tests

We continue with a discussion on 'goodness of fit' tests, also called 'one sample' tests, and 'two sample' tests, where the former ones are used to evaluate the fit of observed values (one sample) with a theoretical distribution and the later ones, which are often a modification of the test statistics for the one sample case, are used to test whether two samples are drawn from the same distribution. Thereby, no assumption about an underlying theoretical distribution is made [16]. Since we want two compare two data series, we are interested in the methods that allow to evaluate the fit of two samples.

A goodness of fit test is a statistical hypothesis test that evaluates whether the observations $X_1, X_2, ..., X_i$ are an independent sample drawn from a particular distribution. The underlying null hypothesis can be stated in general by [16, 41]:

H_0: The $X_i's$ are IID random variables with distribution function $\hat{F}$.

The two sample test evaluates whether the samples X and Y of size n and m, respectively, both come from the same underlying distribution, where all $n + m$ random variables being mutually independent. The null hypothesis can be stated by [16]:

H_0: The $X_i's$ and $Y_i's$ come from the same underlying distribtion.

In order to test such a hypothesis one computes the value for an appropriate test statistic using the observed data and compares it with the critical value at a level of significance α. The null hypothesis is rejected if the test statistic exceeds the critical value.

These tests can be described more formally as parametric, non-parametric and distribution-free. A parametric hypothesis makes assumptions about the underlying distribution and concerns specific parameters. The number of parameters is finite. An example of such a hypothesis is 'that a normal distribution has a specified mean and variance' [58]. On the other hand, non-parametric tests still make assumptions on the underlying distribution but do not consider the parameters. A hypothesis would be 'that a distribution is of normal form with both mean and variance unspecified' [58]. Similarly to non-parametric hypotheses, distribution-free tests do not consider parameters and in addition do not make any assumptions on the underlying distribution [58], as formulated in the hypthesis for the two-sample tests.

Since we assess whether two samples are drawn from the same underlying distribution without knowledge of the true parameters and distribution, we are interested in distribution-free hypothesis tests. There exist a variety of distribution-free goodness of fit tests in literature, however, we will not discuss all of them, but restrict ourselves to the discussion of the most commonly known ones, the Chi-Square, the Kolmogorov-Smirnov and the Anderson-Darling test.

4.4.1 Chi-Square Test

The Chi-Square test, which is the oldest goodness of fit test and was introduced by Pearson in 1900 [48], is a more formal comparison of a histogram with a fitted density or mass function. The data under consideration is binned into several intervals. As binning involves a loss of

information and choosing the size of the intervals sometimes is considered to be arbitrary, it is recommended to avoid unnecessary binning of data and use the Chi-Square test when data is discrete only [50].

In the one-sample case, where we compare an empirical distribution with a hypothesized probability distribution of interest, we start with binning the observed data into k adjacent intervals $[a_0, a_1), ...[a_{k-1}, a_k)$ and define N_i to be the number of observations in the ith bin, where $n = \sum_{i=1}^{k} N_i$ is the total number of observations. Next, we determine the expected proportion of observations p_i that would fall into the ith bin according to the hypothesized distribution. In the continuous case, we have $p_i = \int_{a_{i-1}}^{a_i} \hat{f}(x)$ and in the discrete case, we have $p_i = \sum_{a_{i-1}}^{a_i} \hat{p}(x_i)$. Note that, np_i gives the expected number of observations that would fall into the ith interval. The Chi-square test statistic is computed by [41, 50]:

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \tag{4.12}$$

If the difference in the number of observed values and the expected number of observations $(N_i - np_i)$ is large, we expect the fit to be poor. Thus, when χ^2 statistic is large we reject the null hypothesis that the empirical distribution comes from the hypothesized distribution. On the other hand, if the difference between the number of observed and expected observations is small, we expect the fit to be good and do not reject the null hypothesis [41].

Next, we consider the two-sample case, that is, whether two empirical distributions are consistent. Therefore, let R_i be the number of observations in bin i for the first data set and let S_i be the number of observations in bin i for the second data set. The Chi-Square test statistic is defined by [50]:

$$\chi^2 = \sum_{i=1}^{k} \frac{(R_i - S_i)^2}{R_i + S_i} \tag{4.13}$$

Considering both equations, Equation 4.12 and Equation 4.13, note that in the two-sample case the denominator is not the average, but the sum of R_i and S_i. This is due to the reason that each term in a Chi-square sum is required to model the square of a normally distributed quantity with unit variance and the variance of the difference of two normal quantities is the sum of their individual variances [50].

As mentioned earlier, the binning of data leads to a loss of information. Again, this is reflected in the power of the test. By power we refer to the probability that the correct decision is taken when the alternative hypothesis is true. That is, the power is the probability of rejecting the null hypothesis when the alternative hypothesis is true. The Chi-Square test has less power than the Kolmogorov-Smirnov or Anderson-Darling test. However, it is simple to apply to discrete data.

4.4.2 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test for the one-sample case compares an empirical cumulative distribution function F_n created on basis of observed values with the hypothesized cumulative distribution function $\hat{F}$. The empirical distribution is defined by $F_n(X_{(i)}) = i/n$ for $i = 1, 2, ..., n$. This is a step function which increases by $1/n$ at each ordered data point [41]. To evaluate the goodness of fit between the observed values and the hypothesized distribution function, the closeness between the empirical distribution F_n and the hypothesized

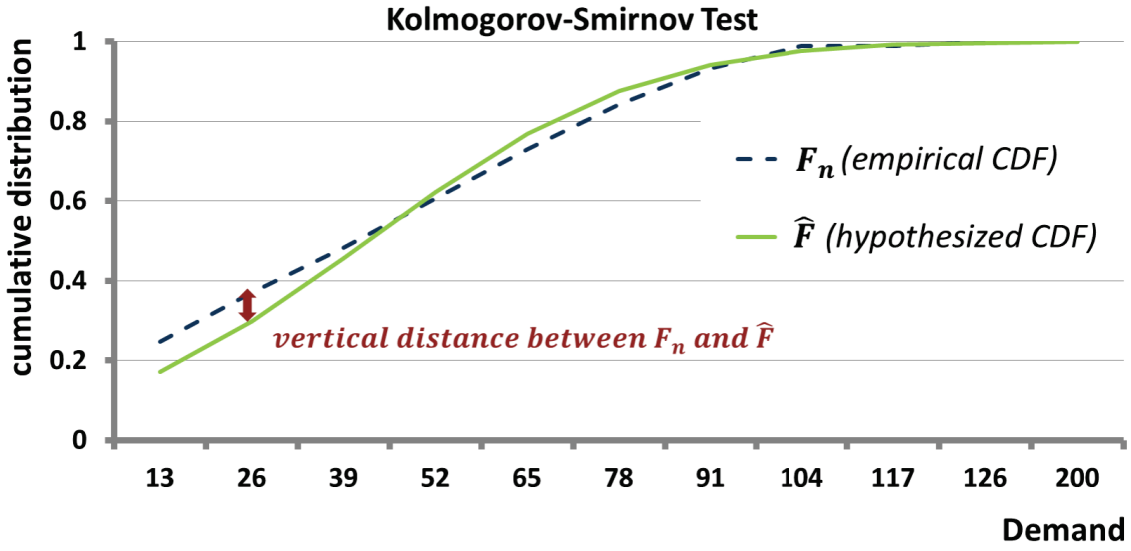


Figure 4.2: Kolmogorov-Smirnov test

distribution $\hat{F}$ is measured. The test statistic D_n describes the maximum (vertical) distance between F_n and $\hat{F}$ *for all values of* x [41], where each distance is weighted equally. That is, there are no distinctions made between the distances in the tails of the distributions or the distances in the head of the distribution as it is done in the Anderson-Darling test, which we explain in the next section, namely subsection 4.4.3. Figure 4.2 visualizes the idea of the Kolmogorovo-Smirnov test.

The test statistic is computed by:

$$D_n = max(D_n^+, D_n^-) \tag{4.14}$$

$$where\ D_n^+ = \max_{1\leq i\leq n}(\frac{i}{n} - \hat{F}(X_{(i)}))\ and\ D_n^- = \max_{1\leq i\leq n}(\hat{F}(X_{(i)}) - \frac{i-1}{n}) \tag{4.15}$$

The null hypothesis 'H_0: The $X_i's$ are IID random variables with distribution function $\hat{F}$' is rejected in favour of 'H_1: The $X_i's$ are not IID random variables with distribution function $\hat{F}$', when the test statistic D_n is greater than the critical value regarding a significance α.

The original form of the Kolmogorov-Smirnov test assumes that the hypothesized distribution is continuous and all its parameters are known. In that case, the distribution of D_n does not depend on the distribution of $\hat{F}$. That is, a single table of all values for $d_{n,1-\alpha}$ is sufficient for any hypothesized continuous distribution [41].

When modifying this original test one can evaluate whether two samples, F_n and G_m, are drawn from the same underlying distribution without any knowledge of the underlying distribution [46]. The null and alternative hypothesis are then stated as: H_0: The $X_i's$ and $Y_i's$ are from a common distribution function' against 'H_1: The $X_i's$ and $Y_i's$ are not from the same common distribution'. Similarly to the original form of the test, the test statistics compares the largest distance between the two empirical cumulative distribution functions

and is defined in the following [47]:

$$D_{mn} = \left(\frac{m * n}{m + n}\right)^{1/2} \sup_x |F_n(x) - G_m(x)| \tag{4.16}$$

The null hypothesis is rejected by a significance level α if $D_{mn} > c(\alpha)$, where $c(\alpha) = \sqrt{-\frac{1}{2} ln(\frac{\alpha}{2})}$ [54]. The Kolmogorov-Smirnov test for the one sample or two sample case is widely used in literature and industry. However, in our case it is difficult to apply since the test requires that the values are continuous. In case of discrete values the test statistic D_{mn} is not defined since the empirical distribution has jumps and thus we cannot take the supreme value for each value of x.

4.4.3 Anderson-Darling Test

Anderson and Darling also developed a distribution-free goodness of fit test. The advantage of the Anderson-Darling statistic over the Kolmogorov-Smirnov statistic is, that they do not assign the same weights to the difference $|F_n(x) - \hat{F}(x)|$ for each value of x, but define a weight function and thus detect discrepancies in the tails of the distributions where the main differences for many distributions of interest lie [3]. The weights are largest at the tails. It can be shown that the Anderson-Darling test statistic can be computed by [2]:

$$A_n^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1)\{ln[\hat{F}(X_{(i)}] + ln[1 - (\hat{F}(X_{(n+1-i)}))]\} \tag{4.17}$$

where $x_{(1)} < ... < x_{(n)}$ is the ordered sample of size n and $\hat{F}(X_{(i)})$ for $i = 1, 2, ..., n$ is the theoretical cumulative distribution to which we compare the sample. The null hypothesis is rejected at a significance level α if A_n^2 exceeds some critical value $a_{n,1-\alpha}$. D'Agostino and Stephens [15] provide among others tables for the critical values $a_{n,1-\alpha}$ for five continuous distributions.

Modifications of the original test statistic proposed by Pettitt [49], generalize the formula for the two-sample Anderson-Darling test:

$$A_n^2 = \frac{1}{mn} \sum_{i=1}^{N-1} \frac{(M_i N - ni)^2}{i(N - i)} \tag{4.18}$$

where $N = n + m$ is the ordered and combined sample of the X's and Y's with sample size n and m, respectively and M_i is defined as the number of X's that are less than or equal to the ith observation of the ordered and combined samples of X and Y. The null hypothesis which states that the two samples X and Y are drawn from the same distribution is rejected at a level of significance α if the two-sample test statistic exceeds the critical value.

4.5 Conclusion

In order to compare two time series we looked at different approaches, namely forecast accuracy measures, time series similarity measures, and hypothesis tests. Forecast accuracy measures compare the ith point of one time series with the ith point of the other time series.

The difference is called error term and is further evaluated. Hereby, we distinguish scale dependent and scale independent measures. Scale dependent measures such as the MSE, MAE, and RMSE have the drawback that one cannot compare the measures when scales are different. Though, scale independent measures overcome the drawback of having to scale data first, they are not defined when demand is zero. As most of the sales products are lumpy, that is there are many periods of zero demand, we cannot apply these measures. In addition, they do not account for the overall behaviour, but compare points. Thus, making it not applicable to our case.

Next, we considered time series similarity measures. These measures are widely used in pattern recognition and data mining research areas to search for similar objects. Most of the introduced time series similarity measures are either based on fourier transforms, that is they consider the periodicity of a time series, or on the distance between two series either by transforming one series into the other one, or by considering the exact shape of a series. The methods based on fourier transform are not suitable since our data does not show any periodicity. Though, edit based distance measures such as DTW, the LCSS and EDR allow for local time shifts and hence may be applicable, we decide due to reasons of practicability to further consider hypothesis tests.

Hypothesis tests are used to test whether two samples are drawn from the same underlying distribution if so, they are considered as consistent. Hence, they account for the overall behaviour of two time series. The Chi-Square test, Kolmogorov-Smirnov test, and the Anderson-Darling test are three widely used hypothesis tests in industry and research. All three of them are non-parametric and distribution-free tests. The Chi-Square test is applicable to both, continuous and discrete data, whereas the other two tests assume a continuous distribution. In our case, the data is discrete, however we can create a stepwise cumulative distribution for the data series to be able to compare the absolute differences as done in the Kolmogorov-Smirnov and Anderson-Darling test. The Chi-Square test has the drawback of binning the data and hence loosing information, which results in a lower power. By power we refer to the probability of rejecting a false null hypothesis. However, it can be applied to discrete data without modifying it. The Anderson-Darling test differs from the Kolmogorov-Smirnov test that it puts heavier weights on the tails of the distributions where many distributions differ and hence gives more precise results. Nevertheless, we decide to use the Kolmogorov-Smirnov test since we do not need the precision of the Anderson-Darling test. We modify the Kolmogorov-Smirnov test such that we compare the total area between two distributions. This allows to compare various generated distributions among each other, however it does not indicate which distribution gives the best fit to the observed one. Thus, we take the Chi-Square test to assess the general fit of the two series.

Chapter 5

Generating demand and checking the fit between data

Figure 5.1 schematically shows the interactions between the input, the output, and the simulation model. The simulation model was verified and validated by the scenario & econometrics team of Infineon [13]. Thus, we concentrate the validation solely on the input to the simulation model. Roughly, we distinguish the input parameters between system settings and demand generation parameters. With system settings, explained in detail in subsection 3.3.1, we refer to the parameters which describe the specifics of the basic types. These are parameters such as the cycle times of the processing steps, the costshare between front end and back end, the product structure as well as the freeze fence. They are set according to the knowledge of the two responsible supply chain planners. The generated demand, on the other hand, needs to be validated by checking the fit with the observed data. Therefore, we modify the Kolmogorov-Smirnov test to choose the best generated series out of many and assess its fit by applying the Chi-Square test, which is done using Excel.

The chapter is structured as follows, we analyse the existing demand generation method by an experimental study in section 5.1, where we aggregate the data using R and analyse the output with Excel. Next, section 5.2 describes the parametrization of the demand signal. We conclude the chapter with suggestions for improving the demand method in section 5.3.

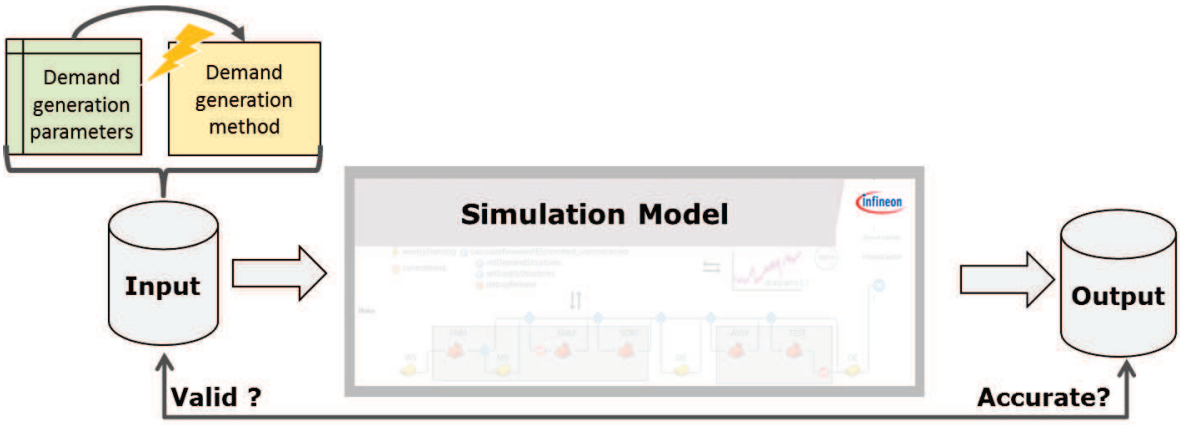


Figure 5.1: Interaction between input, simulation model, and output

5.1 Experimental study of demand generator

In order to parametrize the demand generating method accurately, we study it experimentally. We investigate various set ups of the input parameters. Thereby, we treat the method as a black box and solely analyse the outputted demand using R and Excel. We check whether the method is able to create smooth, intermittent, erratic, and lumpy as well as autocorrelated demand. The parameters are set manually in an internal database file of the simulation before the run is started. An detailed description how the demand generation method is implemented is given in subsection 3.1.2. The method creates demand for the next 26 weeks (six months) to emulate the short term planning of Infineon that spans over six months. To reflect that the unrealised demand of the future is changing, the demand points in the simulation forecast period are modified according to some adjustment term. The following input parameters regard the demand generation method and the adjustment:

- *Average demand* (μ): The average demand per sales product. We do not use actual demand values, but scale them down, since large values slow down considerably the simulation run time.
- *Coefficient of variation* (cv): The ratio between the standard deviation and average demand per sales product. Note, the standard deviation σ can be calculated by $cv * \mu$.
- *Change in standard deviation* ($\sigma\prime$): A constant value that is used in the adjustment term to reflect the changes in the standard deviation.
- *Change in average demand* ($\mu\prime$): A constant value that is used in the adjustment term to reflect the changes in the average demand.
- $AAMax$: Describes the changes in the furthest demand point.
- $AAMin$: Describes the changes in the closest demand point.
- AAN: Controls whether $biasSigma$ is linear ($AAN = 1$), concave ($AAN < 1$), or convex ($AAN > 1$) as shown in Figure 5.2, Figure 5.3, and Figure 5.4, respectively (where $AAMax = 2$, $AAMin = 1$).

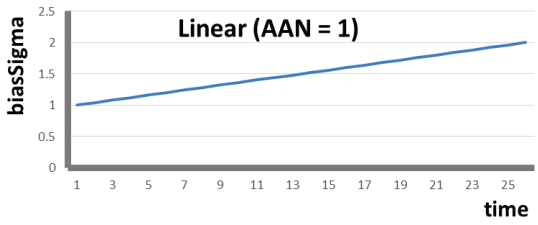


Figure 5.2: biasSigma linear

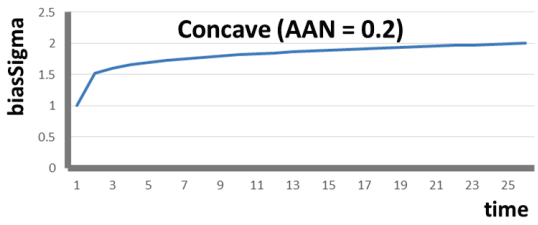


Figure 5.3: biasSigma concave

Table 5.1 shows the input parameters for the demand generation method. We varied each parameter between a minimum and maximum value by a stepsize given in the last column. The range of the values is chosen by considering former parametrizations of the demand method and taking common values into account. One may notice, that we omit to vary the mean as it would solely lead to another centre around which the demand fluctuates and thus makes comparison and conclusions difficult. Note, that when parametrizing the demand

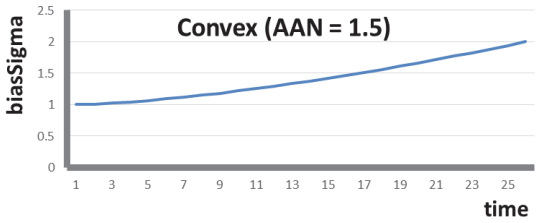


Figure 5.4: biasSigma convex

Table 5.1: Input parameters to demand generating function

	min	**max**	**step**
average demand (μ)	10	10	1
coefficient of variation (cv)	0.2	1.4	0.2
change in standard deviation ($\sigma\prime$)	0.1	0.3	0.05
change in mean ($\mu\prime$)	0	0.1	0.05
aamax	0.5	2	0.5
aamin	0	1.5	0.5
aan	0	1.5	0.5

method according to the considered sales products, we do set various mean values. Combining the parameters for each value results in 6720 experiments. We run the 6720 experiments for a period of 52 weeks, such that we receive 52 demand points for each experiment. Furthermore, we set the freeze fence to one week, that is, the demand is continuously modified in each week. A warmup period is not considered since we are interested in the demand only and do not consider any performance measurements yet.

Once we obtain the demand over the period of 52 weeks for each experiment, we evaluate it according to various measurements, namely, its mean, its standard deviation, the minimum and maximum value, its coefficient of variation, the average demand interval and the autocorrelation for lag one to five. Moreover, we classify the demand in smooth, intermittent, erratic and lumpy as described in subsection 4.1.1. By evaluating these measurements we aim to reveal any particular behaviour and relations between the parameters.

The analysis shows two results. The demand generating method does not produce autocorrelated demand. When we evaluate the autocorrelation in lag one to five we obtain no significant autocorrelation with a 5% level of significance. For details on the autocorrelation threshold value see Appendix C.1. Second, it does not create intermittent demand. Intermittent demand is characterized by rather small fluctuations around its mean, and many periods of no demand. A possible reason why the demand generation method is not able to create intermittent demand is, that zero-demand is generated when the fluctuations are high, and thus created by cutting extreme negative values to fulfill the requirement that demand is ≥ 0.

The demand of the two basic types and their sales products do neither show autocorrelation nor are they classified as intermittent as we elaborate in subsection 2.2.1, and subsection 2.2.2. Thus, in our case we do not run into problems when generating demand with this method. Therefore, we can use the current implemented method to create the demand patterns of the sales products. However, when we examine the autocorrelation of some other

products from ATV and PMM, we find that there are basic types showing significant autocorrelation in the first lag as shown in Appendix C.2. Also, products may be classified as intermittent. Hence, on a long term it is necessary to modify the demand generation method such that it allows to create intermittent as well as autocorrelated demand. We provide suggestions how to improve the method in section 5.3. Furthermore, the experimental study of the demand generation method shows that there are no clear recommendations how to set the parameters, such that a particular demand pattern, e.g., smooth, intermittend, erratic, and lumpy, is generated. Hence, one has to try several set ups in order to receive the desired demand behaviour.

5.2 Parametrization of the simulation model and evaluating the fit

The parameters concerning the basic type specifics (system settings) are set and validated according to the knowledge of the two responsible supply chain planners from CCS with whom we lead a discussion. The demand generation parameters, on the other hand, are set by an exhaustive search where the fit between the generated and observed demand is evaluating by a modification of the Kolmogorov-Smirnov procedure in subsection 5.2.1. The best fitted generated demand is subsequently validated by applying the Chi-Square test in subsection 5.2.2.

5.2.1 A modification of the Kolmogorov-Smirnov approach

We follow a three-phase approach shown in Figure 5.5 using Excel. This is done for each sales product. We have six sales products for basic type BT1 and ten sales products for basic type BT2 accounting for $\geq 85\%$ of the volume (in pieces). In the preparation phase we set the parameters in the generation method such that statistical measures as the mean and coefficient of variation show similar values to the observed data. In the second phase, we evaluate various parametrizations for the same observed demand. Thus, we are interested in the question: which parametrization does give the better fit to the observed data. In the last phase, described by subsection 5.2.2, we evaluate the best parameter setting found in the second phase according to the Chi-square test.

We modify the approach of the Kolmogorov-Smirnov test to our needs. In its original form the Kolmogorov-Smirnov test computes the absolute difference between the cumulative distributions of two data series and takes the supreme value as input to the test statistic. Thereby, it assumes that the two empirical distributions are continuous. Note, that there are extensions to the Kolmogorov-Smirnov test for discrete data. However, these are not straight forward as the distribution of the test statistic D is much more difficult to obtain for discrete data since it depends on the null model [5]. As we have discrete data, one would assume that we apply the extension of the Kolmogorov-Smirnov test for the discrete case. But, we omit this for two reasons. First, and that is the main reason, we do not apply the Kolmogorov-Smirnov test in its original sense, but adapt it to our needs and second, since the extension is not straight forward, this may introduce some source of non-applicability due to its need of deeper understanding. Therefore, we stretch the assumption for continuity by making our discrete distributions stepwise continuous. We order both data series, the observed data $X_1, ..., X_n$ and generated data $Y_1, ..., Y_n$ in ascending order and assign cumulative probabilities, $F(x)$ and $G(y)$, respectively, to the ordered values. We define the cumulative probability, $F(x_i)$

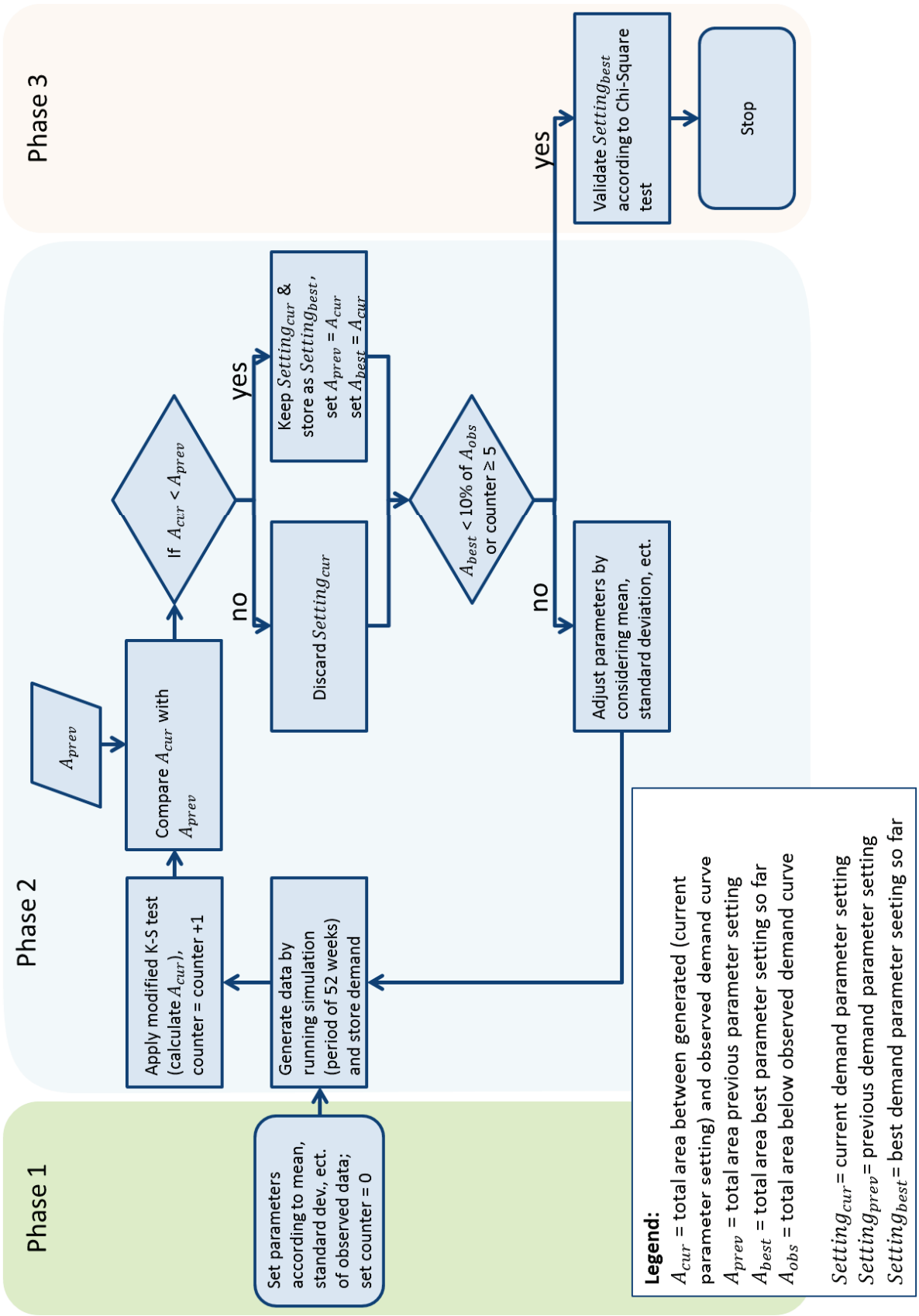


Figure 5.5: Approach of finding the best demand parameter setting

and $G(y_i)$, of the ith data point to be i/N, where i is the ith rank of the ordered series and N is the total number of points in each of the series, note that $N = N_{obs} = N_{gen}$. Next, we define the cumulative probabilities for each integer z (since the data is discrete) in the range of the observed data[a, b] and compare the distributions at these defined points. For example, if the values range from a=0 to b=130, we define the cumulative probabilities at the values of $z = 1, 2, 3, ..., 130$ by assigning $F(z) = F(x)$, where we choose the greatest x for which holds $\leq z$. Elaborating on this example, for the data in the range of $[0, 130]$, we order our observed $X_1, ..., X_n$ and define the cumulative probability of the ith rank to be $F(x_i) = i/130$. As we have discrete data, we may have several probabilities $F(x_i)$ for $\lceil x_i \rceil = \lceil x_{i+1} \rceil = ... = \lceil x_{i+k} \rceil = z$. In this case, we assign $F(z) = F(x_{i+k})$ where x_{i+k} is the greatest integer $\leq z$.

Following the above approach allows us to compute the absolute difference between the two data series at each point z. For further evaluation of two different parametrizations we do not consider the supreme value of the absolute differences as in the Kolmogorov-Smirnov test, since the supreme value may be the same for two different parametrizations as shown in Figure 5.6. This case occurs when we assign the same cumulative probability for some value of z to both generated data series, $G_1(z)$ and $G_2(z)$, where it happens that the absolute difference at this point z is the maximum in the data series. As a result, we would not have an indication which of the two parametrizations performs better. Instead, we assess the total area between the cumulative probability distributions of the observed and generated data as shown in Figure 5.7. That is, parametrization 1 performs better than parametrization 2, if the area between the observed and generated data 1 is smaller than the area between the observed and generated data 2 as stated in the equation below. We can write the sum instead of the integral since we consider discrete data:

$$Difference_1 = \sum_{a}^{b} |F(z) - G_1(z)|dz < Difference_2 = \sum_{a}^{b} |F(z) - G_2(z)|dz \tag{5.1}$$

With this approach we are able to compare various parametrizations of the demand generation method with one another and make suggestions which one of them fits better to the observed data. This is of practical use as it is common in practice to try various parametrizations and check which one fits best. We stop the parametrization either when the total area between the curves of the generated and observed data is $Leq10\%$ of the area below the observed data or after testing more than 5 settings.

However, the approach does not allow to make an overall statement whether a fitted data series is good in general. That is, we receive some value for the total area between the cumulative distribution of the observed and generated data, but do not know whether this value refers to a good or bad fit of the generated data. We can solely compare various parametrizations according to the total area and decide for the parametrization with the smallest total area.

5.2.2 Applying the Chi-Square test

In order to have an indication how well the fit between the observed and generated data performs in general, we apply the Chi-square test for the two-sample case as described in subsection 4.4.1. Since this is a test for discrete data, we do not have to make any modifications to the approach. We use the Chi-Square test to verify that the chosen parametrization according to the modified Kolmogorov-Smirnov procedure is accurate enough. As the Chi-Square

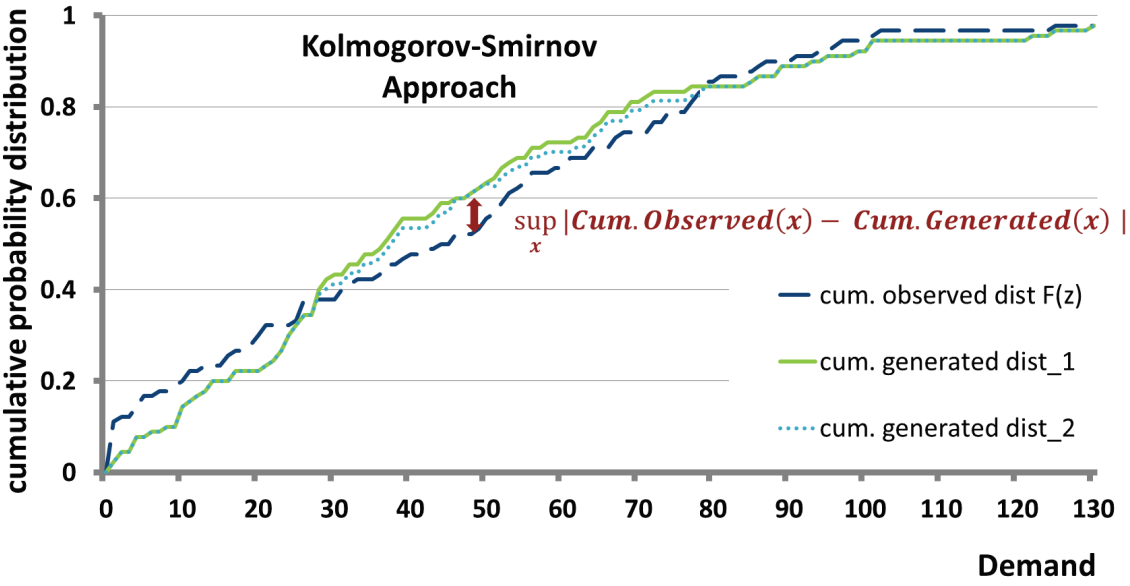


Figure 5.6: Kolmogorov-Smirnov approach

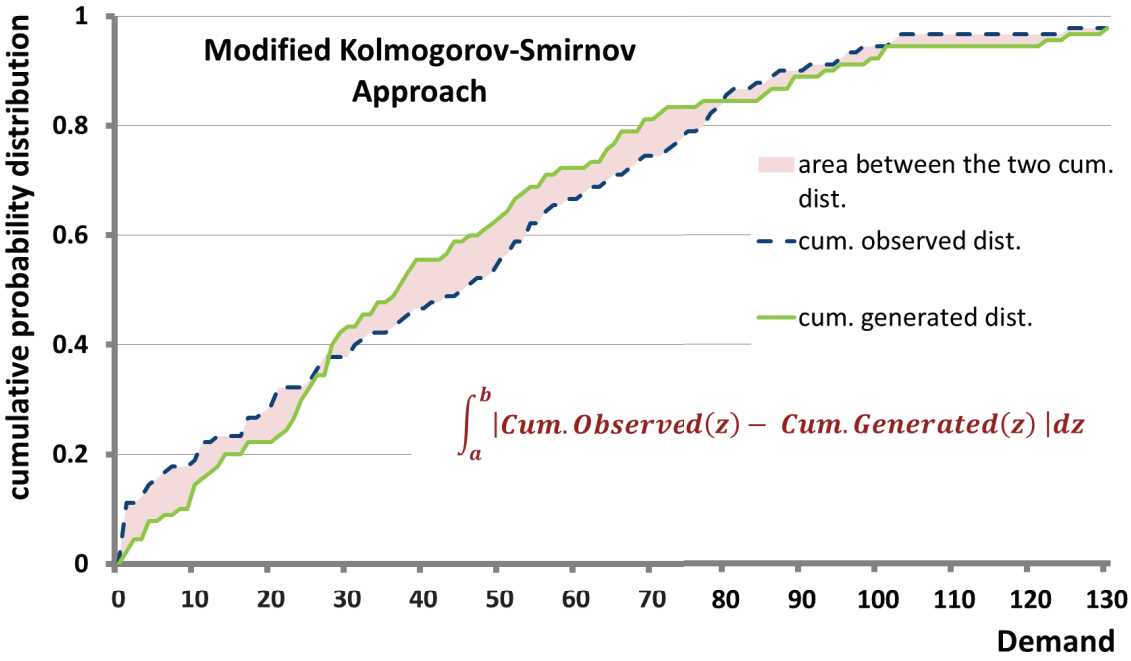


Figure 5.7: Modified Kolmogorov-Smirnov approach for sales product 1 of basic type BT1

statistic has the drawback of binning data and hence loosing information we do not solely rely on the Chi-square test but rather perform it in order to get an indication of the fit between the data.

The evaluation according to the Chi-Square test is done for the parametrization which gives the smallest total area between the observed and generated data. For each sales product of the two exemplary basic types we compute the Chi-Square statistic according to Equation 4.13 and compare it to the critical value $c_{k-1,1-\alpha}$, which is drawn from a Chi-Square distribution with $k-1$ degrees of freedom, where k is the number of non-empty bins and a significance level of α [50]. For the computed values of the Chi-square statistic and the according critical value we refer to Appendix E. The evaluation shows, that the Chi-Square statistic falls in almost all cases except two below the critical value, that is, with a significance level of $\alpha = 1\%$ we can state, that the two empirical distributions for the according sales products are consistent and thereby, validate our approach of the modified Kolmogorov-Smirnov scheme. For the two cases where the Chi-Square statistic is greater than the critical value, that is, for SP6 of the basic type BT1 and SP7 of the basic type BT2, we further check the plot of both cumulative distribution functions and consider the term $(R_i - S_i)$ for each bin i. In case of SP6 we find that there is a large difference in the second bin and for SP7 we find that there is a large difference in the first bin. As both basic types do only account for a small amount of the overall demand, namely 1.8% and 2.4%, we omit further adjustment of the demand parameters as it is quite time consuming. In general, we recommend when observed and generated demand data to not fit according to the Chi-square test to first try another parametrization. If this does not show to improve the fit, we suggest to check how much of the sales's product volume accounts to the overall volume. If this is below some threshold value, e.g. 10%, one may be advised to use the parametrization anyhow. If it is above the threshold value, we suggest to adapt the demand generation method according to the considerations in the next section.

5.3 Improving the demand generation method

We recommend to modify the demand generation method such that it allows for autocorrelation as well as intermittent demand and becomes more user friendly, that is, it makes the parametrization straight forward with parameters that are easy to understand. An overview of relevant characteristics to be considered when improving the method is given in Figure 5.8. In addition, when modifying the demand generation method, one also needs to take into account to change the approach how the simulation forecasts are generated.

We suggest to be able to choose from a set of distributions listed below to create demand even though, the normal distribution often provides a good fit to the observed data as stated in [57]. Tyworth and O'Neill [61] suggest to use the normal distribution if the ratio $\sigma_L/\hat{x}_L$ does not exceed 0.5, where σ_L is the standard deviation of the demand during the replenishment lead time and $\hat{x}_L$ is the expected demand during the replenishment lead time. Otherwise, they propose to use another distribution as well. One may check which one of the available distributions fits best to the observed demand and decide for the appropriate one. In case none of the distributions fit, one may use a simple statistics distribution considering the parameters such as the mean, CV^2, and range to describe its behaviour. In the following we give some distributions to be considered.

- Discrete uniform: For demand that is varying among two integers where we have little

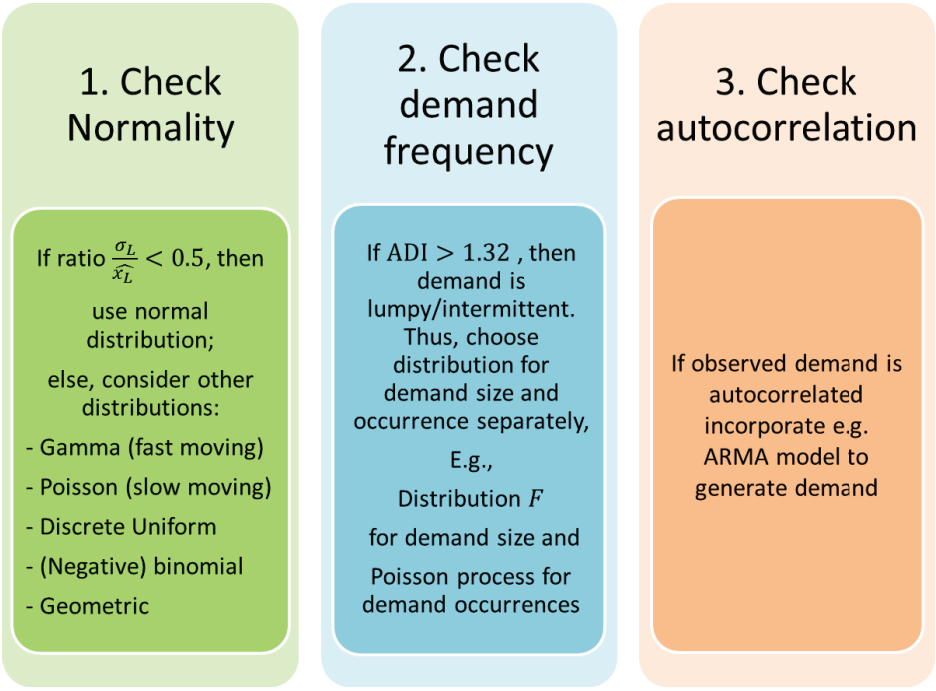


Figure 5.8: Relevant characteristics to be considered when improving the demand generation method

further information [41].

- Gamma distribution: For frequently occuring demand, when the demand distribution is skewed to the right [10].
- Poisson distribution: For low moving items. Demand occurs seldomly (lumpy or intermittent demand) [53].
- Compound Poisson distribution: Demand size is described by some distribution F and demand occurences is described by Poisson process [52].
- (Negative) binomial [41]
- Geometric [41]

In addition, when demand is intermittent or erratic according to the scheme of Syntetos & Boylan [59] explained in subsection 4.1.1, we recommend to describe the demand size and the demand occurrences separately, similarly to the compound Poisson distribution. That is, one chooses a distribution for the volume of the demand and another one for the occurrences. This allows to model slow moving items. Another approach for the current implementation would be to introduce a new parameter that describes the average inter-demand interval (ADI), that is, the average number of periods with no demand over the simulation horizon. In case that the current average demand interval falls below the desired value, zero demand is created for the next one to two weeks, otherwise we omit to create zero demand.

Last, we require to be able to create autocorrelated demand as this influences the stocking strategies. The same stocking strategy may lead to stock outs and thus a lower service level in case of autocorrelated demand compared to non autocorrelated demand as we elaborate

in subsection 2.2.2. Hence, to draw the correct conclusions from the simulation results, it should be included in the simulation model. One possible approach to implement positive or negative autocorrelation in the demand generation method is to use an ARMA model as it is done in [11]. The autoregressive coefficient, which determines the autocorrelation value, can be set as desired.

Table 5.2 summarizes the parameters which we identify that are necessary to describe the demand behaviour. They may be further considered to be implemented in the demand generation method. Note, this is not a restricted enumeration but aims to give ideas for further elaborations. The parameters include the mean, median, range, CV^2, autocorrelation, and ADI for which we give a definition and description in the same table. The mean and median are chosen in order to describe the overall shape of the distribution. That is, to examine whether the demand is skewed to the left or to the right. The range and CV^2 serve to describe the spread and fluctuations of the demand. The autocorrelation is chosen to reflect whether there is an increasing or decreasing trend in the demand over several periods. Last, we find that the ADI is relevant to distinguish between frequently occurring and non-frequently occurring demand.

The recommended actions are a first practical approach to improve the current demand generation method. Nevertheless, they need further elaboration regarding their realization. As our focus lies on the improvement of the stocking levels, we do not further implement the possible improvements due to time constraints. Furthermore, for the two considered basic types we are able to generate sufficient similar demand by checking various set ups of the parametrization and evaluating this to the modified Kolmogorov-Smirnov approach and Chi-Square test.

Measures	Explanation	Definition	Meaning
Distribution Shape: Skewness	Skewness: Absence of symmetry	$\nu = \frac{E\left[(X-\mu)^3\right]}{\sigma^3}$	Data may be skewed to the left or right. If skewed to the left: median < mean If skewed to the right: mean < median
Distribution Location: mean, median	Mean: Average of a sample	$\mu = \frac{\sum x_i}{N}$	Average value of a data set.
	Median: Value in the middle of an ordered data set		Median < mean: data skewed to the left; rather many values (>50%) are smaller than the mean. Median > mean: data skewed to the right; rather many values (>50%) are larger than the mean.
Distribution Variability: Range, CV^2	Range: Difference between the smallest and largest value	$Range = Max - Min$	Describes the spread of the data.
	CV^2: sets the standard deviation in relation to the mean	$CV^2 = \left(\frac{\sigma}{\mu}\right)^2$	$CV^2 \leq 0.49$: smooth, intermittent $CV^2 > 0.49$: erratic, lumpy
Autocorrelation	Measures the internal relation within a data series	$r_k = \frac{E[(X_t-\mu)(X_{t-k}-\mu)]}{\sigma^2}$	$r_k > thresholdvalue$ (autocorrelation): demand increase/decrease over succeeding periods $r_k < thresholdvalue$ (no autocorrelation): demand does not increase/decrease over succeeding periods
Average inter-demand interval (ADI)	Measures the average number of periods between succeeding orders where N=number of total periods and D_N=number of periods where demand occurred	$ADI = \frac{N}{D_N}$	ADI <= 1.32: smooth, erratic ADI >1.32: intermittent, lumpy

Table 5.2: Relevant parameters for describing the demand behaviour, explanations are based on [41]

5.4 Conclusion

We set up the simulation model according to the needs of the basic types and their corresponding sales products as well as validate the input parameters. Input parameters such as the cycle times at the processing steps and the profit margin are validated by discussions with the two responsible supply chain planners. On the other hand, we choose the best set up of the demand generation parameters by using a modified approach of the Kolmogorov-Smirnov statistic. Hereby, we evaluate the total area between the observed and generated cumulative distribution to compare several parametrizations and choose the one with the smallest area. This set up is then validated by applying the Chi-Square test. It allows to make an overall statement whether the fit is good. The evaluation shows that all fitted series except for two sales products are suited according to the Chi-Square test. As the two demand series only account for 1.8% and 2.4% of the overall demand we omit to test further parametrizations. Last, we recommend improvements to the demand generation method. These include to allow for periods with zero-demand, creating autocorrelated demand as well as being able to choose from various distributions for the demand generation. The given recommendations are a first practical approach, however they need further research regarding their realization. Since our focus lies on simulating the supply chain and considering various planning strategies we do not further implement the improvements. Having parametrized the simulation model according to the observed demand with the current method and given a method to assess the fit, we solved the two subproblems:

Parametrize the order arrival pocess in the existing simulation model such that the generated demand data accurately describes observed data to make the simulation results more representative.

Define a method to assess the fit between generated and observed values according to their statistical behaviour.

This enables us to tackle the core problem of improving CCS's planning process in the next chapter.

Chapter 6

Improving the supply chain planning process for two exemplary basic types

We state in our problem definition in section 1.3 to improve the supply chain planning process at Chip Card & Security (CCS) for two exemplary basic types. Therefore, we use the fine tuned simulation model to compare various production release approaches and stocking strategies to provide CCS with recommendations. These suggestions serve as an indication which steps can be taken to enhance their current practices. The planning methods are evaluated according to the service level and cost trade off. That is, the higher the stocks, the faster is the delivery and the more unlikely is it to run into backlog. However, higher stocks increase the costs. Thus, we aim to balance a high service level at relative low costs.

In section 6.1 we start with introducing the examined planning concepts for determining the production releases in front end as well as various stocking strategies. Next, 6.2 shows the experimental design and gives an overview how we determined the number of replications, the warm-up period, and the run length. In section 6.3 we present and discuss the results of the simulation study and section 6.4 concludes the chapter with a sensitivity analysis.

6.1 Planning concepts for determining stocking levels

The planning process at CCS for determining stocking levels of their two exemplary basic types includes two concepts: The production release approach as well as the stocking strategy. The production release approach describes how we determine the amount of wafers to be started in front end in advance where actual orders are not known yet. On the other hand, the stocking strategy defines at which stock points to place inventory and the inventory level. That is, the amount of stock at the stocking points.

6.1.1 Production release approaches

When production is started in front end, one does not know the demand in advance as the customer order decoupling point (CODP) lies further down in the supply chain. The CODP defines where in the supply chain the customer order entry lies. Figure 6.1 illustrates various decoupling points for the supply chain of Infineon, namely make-to-order, assemble-to-order,

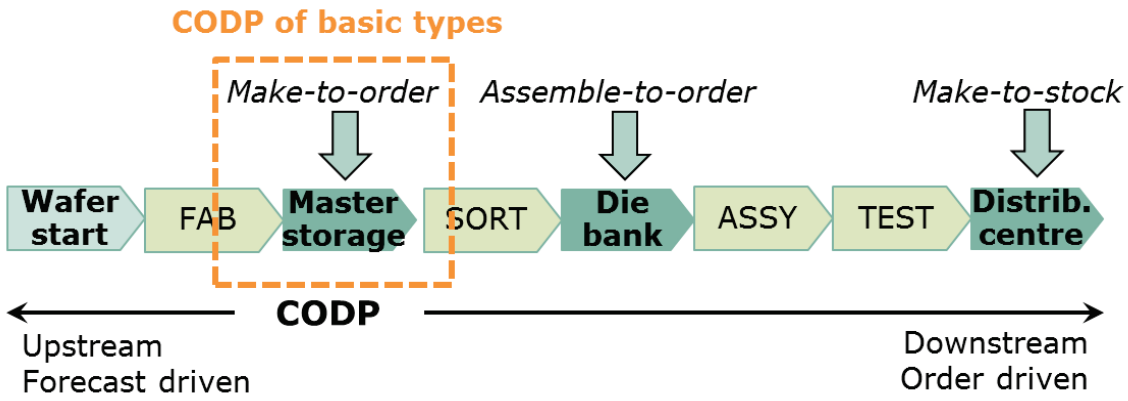


Figure 6.1: Customer order decoupling points at Infineon based on the illustration of [6]

and make-to-stock. Activities upstream of the CODP are forecast driven since customer orders are not known yet, whereas activities downstream of the CODP are order driven where demand is known [6].

The CODP of the two examined basic types lies at the master storage since products become customer specific in the sort step described in section 2.1.1. The amount of wafers to be started in front end (wafer start) up to the master storage is forecast driven. Currently, a four months moving average is used as forecast. As the processing time from wafer start to the master storage takes fairly long, between ten to thirteen weeks, it is important that there are enough products stored at the master storage to satisfy customer demand. In case there is a shortage at the master storage, customers face long lead times and may migrate to competitors.

For the simulation study we define three production release approaches including the current one. Our aim is to examine which one of the approaches results in the best service level and cost balance. The production release approaches are defined below and Figure 6.2 illustrates the time horizon of the used data. Next to the current approach, which uses a four month moving average over the historical data, we define one approach that applies a moving average not only over historical data but also over known orders as well as another approach that uses single exponential smoothing (SES) as explained in subsection 4.1.3.

We denote the approaches by 'Hist','Order', and 'SES' to differ between historical demand data, order data, and forecasted demand data using SES, respectively. Historical demand are realised orders of the past ($t = -1, -2, ...$), order data are the incoming orders at the master storage that are known for the next three weeks ($t = 0, t = 1, t = 2$), and forecasted demand data is determined by single exponential smoothing:

1) 'Hist'-approach: The release quantity is determined by a moving average over the last four months of demand data as it is currently done. It considers historical demand data from weeks $t = -16$ up to week $t = 0$.
2) 'Hist&Order'-approach: The release quantity is defined by a moving average over the last two weeks of observed demand as well as the known incoming orders for the next three weeks, that is weeks $t = -2, t = -1, t = 0, t = 1$, and $t = 2$ are taken into account.
3) 'SES'-approach: The release quantity in week t is described by single exponential smoothing of the observed and forecasted demand data in week $t - 1$ where we define

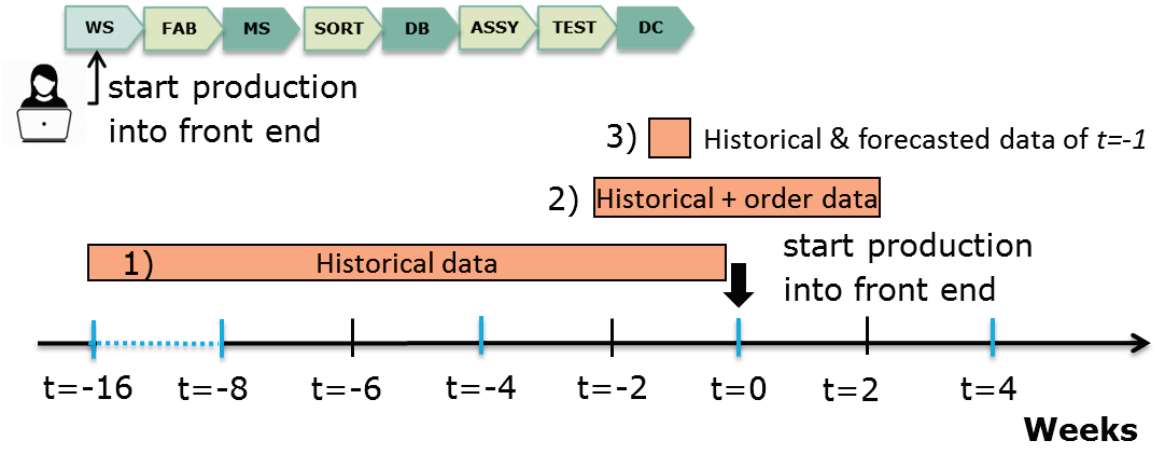


Figure 6.2: Time horizon of the production release approaches

the smoothing parameter $\alpha^{SES} = 0.7$. We choose the value for α such that the forecast for the next data point relies mostly on the observed demand.

We define these three approaches since the first one covers the used practices of today, and the second one uses data that is known in practice and may lead to a better reaction in front end since the horizon is smaller and actual orders are taken into account. The last one is chosen to consider the possibility of using a more elaborated forecast technique instead of a simple moving average, where the forecast takes the observed demand and the forecast error of the previous week into account. The smoothing parameter α^{SES} can be varied between $[0, 1]$. We set it to 0.7 such that a higher weight is applied to the observed demand data and a lower weight to the forecast error. However, this can be adapted if necessary. These three approaches are implemented in the current simulation model using Java. Next to determining the approach of production release, we are interested in the target reach at the storage locations which give the best service level and cost balance.

6.1.2 Stocking strategies

In general, products are stored at three different storage locations, the master storage, the die bank, and the distribution centre. Depending on where we store the products their added value and thus their storage costs is lower or higher. Products stored upstream in the supply chain, for example at the master storage, have a lower value, and hence, the associated storage costs are less expensive. Contrarily, products stored downstream in the supply chain, for example at the distribution centre, have a higher value and thus the storage costs are more expensive. Another factor that influences the decision where to store products is the lead time. The lead time for products stored upstream in the supply chain is longer than for products stored downstream in the supply chain. This in turn may reflect on the customer satisfaction. However, we also need to keep the CODP in mind when taking the decision where to store products. Generally, products that do not become customer specific may be stored at all three storage locations. Products that become customer specific, on the other hand, may only be stored at upstream stock points since producing customer specific products is too risky.

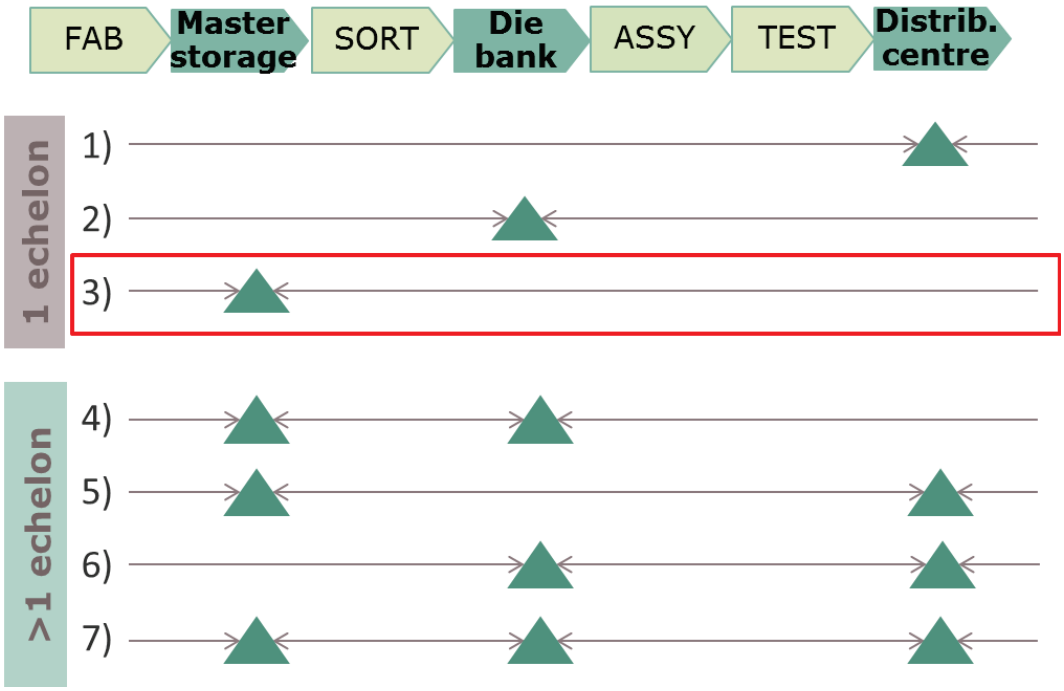


Figure 6.3: Various existing combinations for storing items at the stocking points in Infineon's supply chain

Next to the decision where we store products, we are interested in how much to store at the stocking points to receive a sufficient service level while having appropriate expenses. The more products we store, the more likely we are to satisfy customer demand from stock and hence have a higher service level, however the associated costs also increase. Thus, there is a trade-off between service level and costs when setting the target reach.

When considering the three storage points there are seven combinations where to stock items as shown in Figure 6.3. We can either store at only one of the stock points, at two points each or at all three of them. Furthermore, the target reach may vary for each stock point. It may be favourable to have a higher target reach at the master storage to employ the risk pooling effect.

Usually, with a make-to-order CODP, as it is the case for the two exemplary basic types, one has a stock point solely at the master storage since products become customer specific in the succeeding processing steps and thus, we would need to know which customer specific sales products to produce to die bank or distribution centre. Hence, there is no planned stock at the die bank nor the distribution centre. Nevertheless, there may be cycle stock. Cycle stock are goods that are stored at die bank before further processing due to capacity limits in back end or finished products waiting at the distribution centre for deposition to the customer. Another reason for occurrence of cycle stock may be a large lot size. In production a fixed quantity is produced (lot size). However, in case the lot size is larger than the ordered quantity, more products are produced than needed. Since the CODP of the two products lies at the master storage, we solely take strategy 3) shown in Figure 6.3 into account. Hence, we employ risk pooling, that is, demand variability is reduced by aggregation across sales products [6].

6.2 Experimental design and set up

On the one hand, we consider how to define the amount of wafers to start in front end in advance and on the other hand, we consider which storage locations to use and how high the inventory should be at these locations. The various approaches are evaluated according to the α-service level and respective costs described in section 3.3.2. Section 6.2.1 gives the design of the experiments. However, before running the experiments we define the warm-up period, that is the time interval needed to warm up the system such that the results are not biased and we set the run length as well as the number of replications, which is explained in section 6.2.2.

6.2.1 Experimental design

For each of the three production release approaches, the stocking strategy 3) is taken into account. That is, we vary the target reach at the master storage between one and thirteen weeks (13 set ups) to see the effects when reducing the master storage from the current target reach (13 weeks) to a lower target reach. Combining the set ups with the three production release approaches results in $3 * 13 = 39$ experiments. The range of the target reach at the master storage, die bank, and distribution centre is in accordance with the target reach applied in practice. Table 6.1 summarizes the factors as well as corresponding levels of our experiments.

Table 6.1: Experimental design for the simulation study of two exemplary basic types

Factor	**Levels**	**# of levels**
production release approach	{'Hist', 'Hist&Order', 'SES'}	3
Stocking strategies	*target reach (weeks)*	
master storage	{1, 2,..., 13}	13
die bank	{0}	1
distribution centre	{0}	1
Total number of experiments		**3x13x1x1= 39**

6.2.2 Number of replications, warmup period, and run length

Due to the random nature of a statistical process it is necessary to run an experiment not only once, but several times. Running an experiment once gives a particular realization of the random variables which may have large variations. Thus, to capture the true characteristics of the simulation model, it is necessary to run the experiments various times such that individual outliers do not erroneously influence the simulation output [41].

Determining the number of runs can either be done by an estimation such as the Replication/Deletion Approach described in [41] or by an exact algorithm such as the Sequential Procedure of Law and Carson [42]. First, we use the approximation to determine the number of replications and afterwards verify it with the exact algorithm. The Replication/Deletion Approach as well as the Sequential Procedure result in two to 14 replications per experiment for both basic types. As we do not evaluate all 39 possible experimental settings but restrict ourselves to nine, we add some more replications and decide to run each experiment for 15

independent runs. In Appendix F.1 we elaborate on the Replication/Deletion Approach as well as the Sequential Procedure.

Furthermore, the performance measures of a simulation may either depend on initial conditions, then we speak of transient system behaviour, or they do not depend on initial conditions, then we call it steady-state behaviour [41]. A simulation that models a shop which opens at 9am in the morning where no customers have arrived yet is an example of a transient system since the performance measures are influenced by the number of customers that arrive over the day, but not by customers of the previous day as we consider them to be served. On the other hand, a production line which runs 24x7 where it reaches a steady state after a certain period does not depend on the initial conditions as we assume that there is always work in process in the production line and thus, the performance of an empty system is not representative for the true system behaviour. In case that performance measures do not depend on initial conditions, we need to determine a warmup period. That is, we do not include the first couple of weeks in our performance measures until the system is warmed up. Generally, the longer the warmup period is, the less is the impact of the initial state. However, also the longer is the run length. Hence, one should be aware of the trade off between reducing the impact of the initial state and the run length [41]. The modelled supply chain has a steady state behaviour and thus, we determine a warmup period which will not be included in the performance measures. A simple and general technique is the graphical method of Welch [41]. Welch's method for determining the warmup period is based on the idea to plot the moving averages of the observations and choose the time interval as warmup period where the observations appear to converge. Appendix F.2 describes the procedure in detail. From the graphical method one would choose a warmup period of approximately 40 weeks. Again, as we do not apply this procedure to each experiment but solely to one experiment per basic type we decide to set the warmup period to 52 weeks (hence, to one year) and consider it to be adequately large to exclude initial state behaviour of the performance measures.

Last, the run length is set to 208 weeks (four years) of which 52 weeks (one year) account for the warm up period and the remaining 156 weeks (three years) are taken into account to collect the data and establish the performance measures. We consider three years of collected data to be sufficiently large in order to draw profound conclusions.

6.3 Results

The results of the simulation study provide us with an indication of the system behaviour, which we can use to propose recommendations regarding the production release approach and favourable stocking strategies. They are evaluated according to the α-service level and costs.

We show the main results for both basic types in Figure 6.4 and Figure 6.5. The figures illustrate the three production release approaches, 'Hist', 'Hist&Order', and 'SES'. For each of them, we plot the 13 stocking strategies according to their α-service level given on the x-Axis and the relative total costs shown on the y-Axis, which are scaled according to the overall highest total costs for the reason of confidentiality.

Considering the overall performance of the three production release approaches, we can see that taking historical and order data into account when calculating the moving average or using single exponential smoothing for production release outperforms the current approach as the points lie to the far right where the α-service level increases. Both new approaches

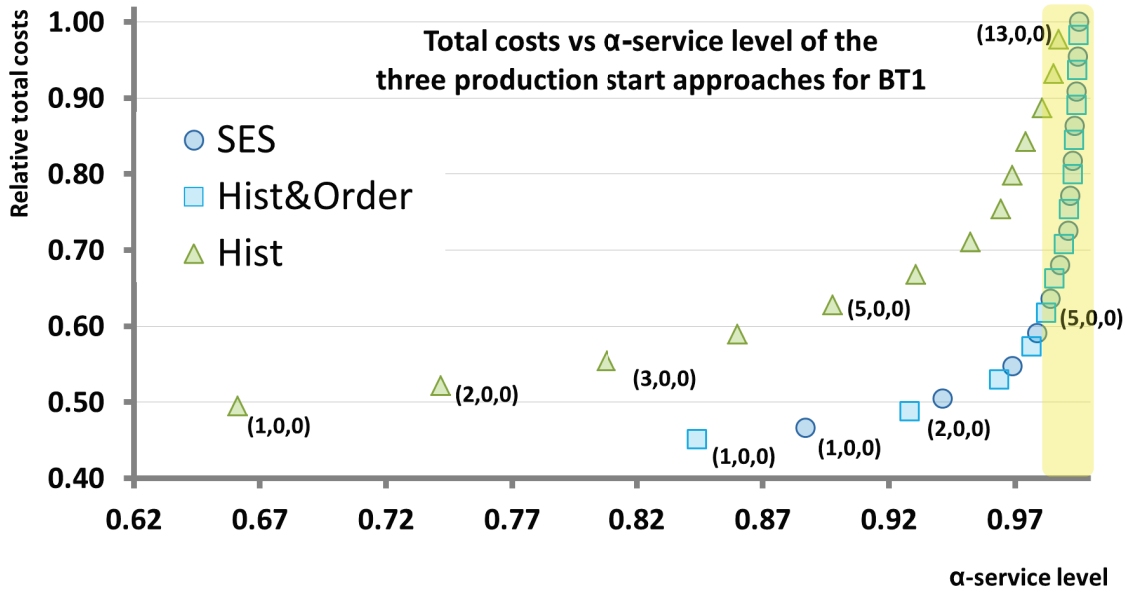


Figure 6.4: α-service level versus total costs of the three production release approaches and considered stocking strategies for basic type BT1

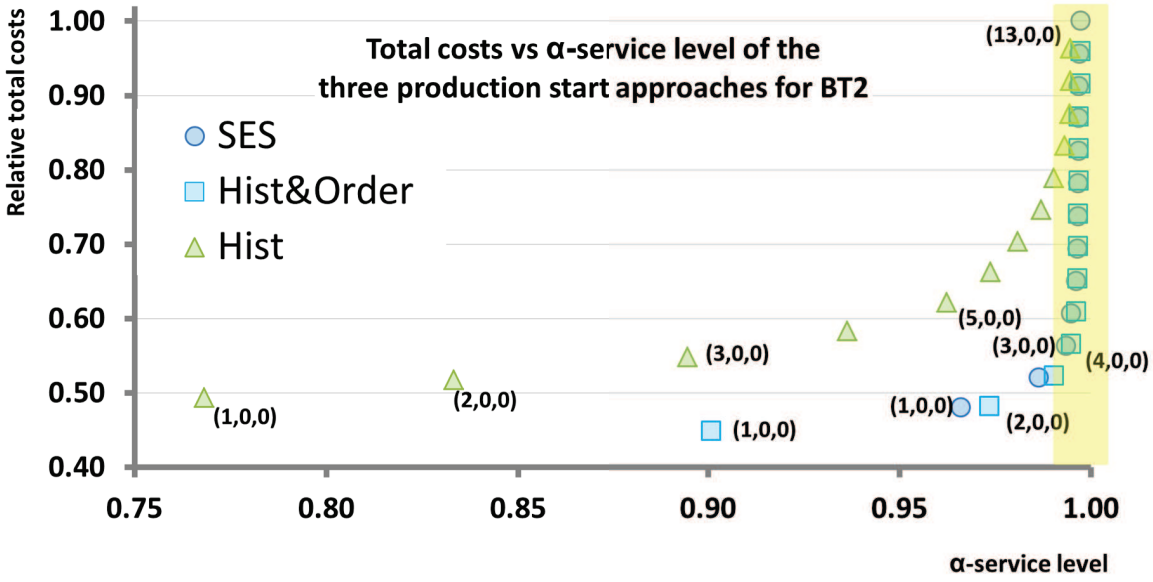


Figure 6.5: α-service level versus total costs of the three production release approaches and considered stocking strategies for basic type BT2

deviate only slightly in their service level and costs. This valuable insight demonstrates that a clever chosen production release approach in front end can already lead to an improved α-service level. On average, for product BT1 the 'SES' approach improves the α-service level towards the 'Hist&Order' approach by 1% and towards the 'Hist' approach by 8%. Comparing the 'Hist&Order' and 'Hist' procedure, shows that using historical and known order data outperforms the use of historical data by 7%. Similar results are found for product BT2. The 'SES' approach outperforms the 'Hist&Order' approach by 1% and the 'Hist' approach by around 5%.

Both new approaches have in common to consider less observed data points and thus react faster to variations in demand where fluctuations are not as strongly smoothed out as in the 'Hist' approach. The 'SES' approach is based on the last observed demand point and its respective forecasted value and the 'Hist&Order' approach includes a moving average over five weeks. A larger time window, on the other hand, would suggest of having a rather stable amount of wafers to be released in front end since demand fluctuations are smoothed out whereas a shorter time window suggests more fluctuations in the amount of wafers to be started. A rather stable amount of wafers to be released in front end may be favourable when production is fluctuating since it stabilizes the overall production flow.

Looking closer at the individual stocking strategies, we are interested in those, which achieve a similar service level to the current one (98%) but at lower costs. These are highlighted in Figure 6.4, and Figure 6.5. Note, we denote the strategies by (X, Y, Z) where X, Y, and Z are the target reach at the master storage, die bank and distribution centre, respectively. For both basic types we do not have stocks at the die bank, and distribution centre, thus $Y = 0$ and $Z = 0$. Figure 6.4 illustrates that the master storage for product BT1 can be decreased to a target reach of five weeks when using either the 'Hist&Order' or the 'SES' approach and still achieving a similar α-service level to the current one of 98%. This result is shown more clearly in Figure 6.6. Figure 6.6 illustrates the change in the α-service level when decreasing the target reach at the master storage for product BT1. For strategies 'SES' and 'Hist&Order' the change in the α-service level is only marginal when decreasing the target reach down to five weeks. A significant drop by more than 3% in the service level occurs when further decreasing the target reach to two weeks. Thus, it is advisable to keep a target reach of at least five weeks. Furthermore, when continuing to use the current approach, 'Hist', the target reach can also be reduced to eight weeks without a strong change in the α-service level. Those reductions in the target reach for all three approaches result in a cost decrease compared to the current target reach of 13 weeks as shown in Figure 6.7. For both new approaches a cost reduction of about 40% is achieved at a target reach of five weeks and for the current approach a cost reduction of 20% is achieved at a target reach of eight weeks. It also shows that a 10% reduction of costs can be already achieved by decreasing the target reach to eleven weeks. Hence, with a target reach reduction of solely two weeks cost savings are comparably high.

Similarly, for product BT2 the target reach at the master storage can be reduced down to four weeks when using the 'Hist&Order' approach and to three weeks when employing the 'SES' approach which results in a cost decrease of around 50%. Thus, it is highly recommended to employ another production release approach as it safes costs by more than 50%. Detailed results for BT2 are given in Appendix H.

Up to now we checked graphically how much we may reduce the target reach for each production release approach to keep the current service level. Next, we analytically check by how many weeks we can reduce the target reach without a significant change in the α-

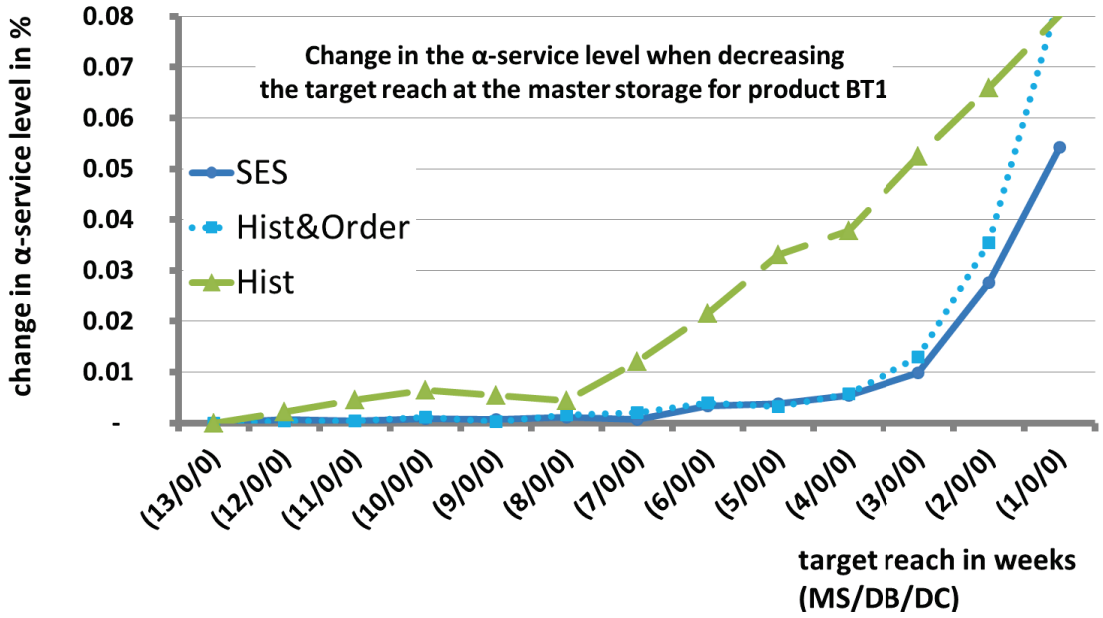


Figure 6.6: Change in the α-service level when decreasing the target reach at the master storage for each of the production release approaches of basic type BT1

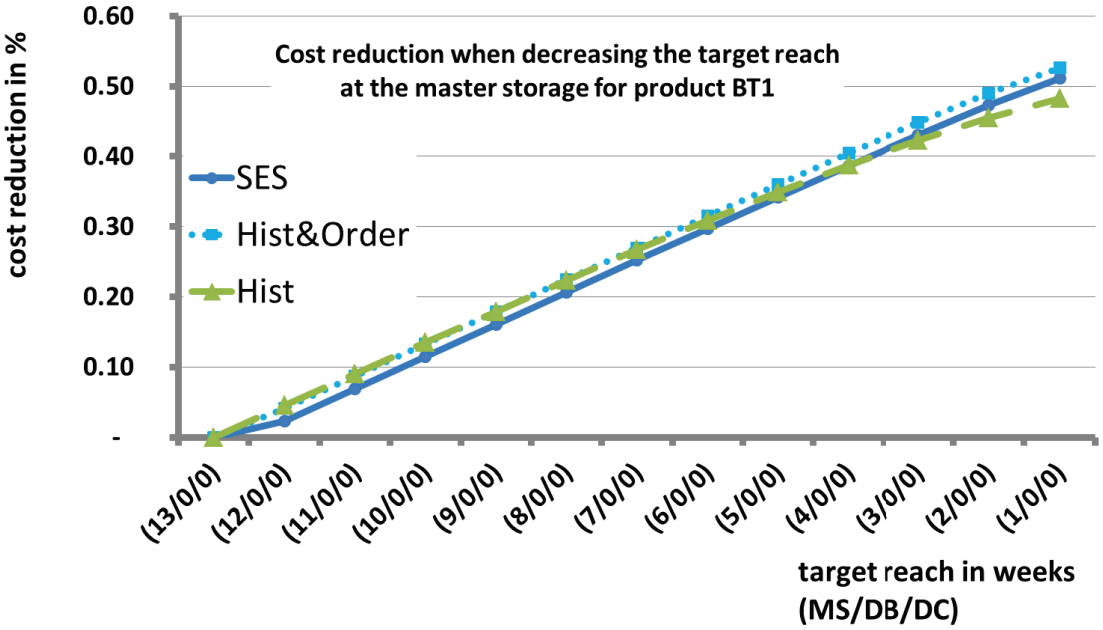


Figure 6.7: Reduction in costs compared to the current costs when decreasing the target reach at the master storage for basic type BT1

Table 6.2: 95% confidence intervals for the α-service level for basic Type BT1

	Hist 95% CI		**Hist&Order 95% CI**		**SES 95% CI**	
stocking strategy	LB	UB	LB	UB	LB	UB
$(13,0,0)^{Hist}$ vs (11,0,0)	0.004	0.009	- 0.010	- 0.005	- 0.010	- 0.005
$(13,0,0)^{Hist}$ vs (8,0,0)	0.019	0.027	- 0.006	- 0.002	- 0.007	- 0.003
$(13,0,0)^{Hist}$ vs (5,0,0)	0.081	0.099	0.002	0.008	0.000	0.006

service level by constructing a confidence interval (CI) using the paired-t approach. If the two system configurations do significantly differ, the confidence interval will not contain zero, otherwise, if the system configurations do not significantly differ, the confidence interval will contain zero. Note that, two systems are correlated if they use the same common random numbers, that is, the seed values of the various simulation replications are the same for both configurations. The advantage of using the same common random numbers is the introduction of useful correlation which in turn reduces the variance among the two system configurations and leads to a smaller confidence interval [41]. For a detailed explanation of the paired-t approach we refer to Appendix G.

We construct the 95%-CI comparing the current strategy $(13,0,0)^{Hist}$ with (11,0,0), (8,0,0), and (5,0,0) for each production release approach given in Table 6.2. This shows for the current approach that statistically there is a significant difference in the α-service level when reducing the target reach to eleven, eight, or five weeks as the confidence interval does not contain zero. However, in praxis one may still reduce the target reach at the current approach as the confidence intervals are fairly small and the drop of the α-service level is marginal. For strategies 'Hist&Order' as well as 'SES' the CIs for a target reach of eleven or eight weeks are negative. That is, both approaches are significantly better than the current approach. Moreover, for a target reach of five weeks the CIs contain zero, that is, there is no significant change in the α-service level. Hence, this strengthens the advice of applying a new production release approach and simultaneously reducing the target reach when the current α-service level is considered to be sufficient. In case that the α-service level should be improved at the same costs, a new production release approach can be applied while keeping a target reach of 13 weeks.

The question remains whether one of the new approaches is significantly better than the other one. Therefore, we compare the 'Hist&Order' and 'SES' approach by constructing the 95% CI for the α-service level and the costs for strategies (11,0,0), (8,0,0), and (5,0,0). Table 6.3 provides the CIs. The CIs show that the 'Hist&Order' approach has significantly lower costs at the same α-service level for a target reach of eleven or eight weeks. Moreover, for a target reach of five weeks the costs are still significantly lower, however, also the α-service level is significantly lower. That is, the 'Hist&Order' approach outperforms the 'SES' approach regarding the costs for a target reach of eleven, eight, or five weeks. Note, for a target reach of five weeks it has an impact on the α-service level compared to the 'SES' approach. However, comparing it to the current approach 'Hist', there is no significant impact on the α-service level at a target reach of five weeks. Hence, we would advice to choose the 'Hist&Order' approach. Furthermore, the 'Hist&Order' approach has the advantage that it is fairly easy to apply and to understand.

In conclusion, the results show that a careful chosen time window over which we calculate

Table 6.3: 95% confidence intervals for the costs and α-service level comparing the 'Hist&Order' with the 'SES' approach for basic type BT1

	95% CI for costs		**95% CI for α**	
strategies	**LB**	**UB**	**LB**	**UB**
$(11,0,0)^{Hist\&Order}$ **vs** $(11,0,0)^{SES}$	- 0.020	- 0.018	- 0.001	0.001
$(8,0,0)^{Hist\&Order}$ **vs** $(8,0,0)^{SES}$	- 0.023	- 0.022	- 0.001	0.000
$(5,0,0)^{Hist\&Order}$ **vs** $(5,0,0)^{SES}$	- 0.028	- 0.026	- 0.001	- 0.001

the moving average to release the according quantity in production is favourable, since one can achieve the same service level as for the current approach but at lower costs. This leads us to the question whether a shift of the time window improves the 'Hist&Order' approach and whether varying the smoothing parameter α^{SES} outperforms the current 'SES' approach. This is further examined in the next section.

6.4 Sensitivity Analysis

The results of the previous section showed that a smaller chosen time window ('Hist&Order), or using single exponential smoothing ('SES') performs better than the current approach which uses a simple moving average over four months. We are now interested, what impact a shift of the time window for the 'Hist&Order' approach as well as another value for the smoothing parameter α^{SES} of the 'SES' approach have on their performances. Thus, we determine the following three additional approaches:

1) 'Hist&Order$_{shift}$': Shifting the time window backward: Moving average over the last five weeks (weeks $t = -5, ..., -1$) such that known incoming orders are not considered.
2) 'SES$_{\alpha=0.1}$': Setting the smoothing parameter $\alpha^{SES} = 0.1$. That is, the forecasted point in week $t + 1$ relies mainly on the previous forecasted value of week t.
3) 'SES$_{\alpha=0.9}$': Setting the smoothing parameter $\alpha^{SES} = 0.9$. That is, the forecasted point in week $t + 1$ relies mainly on the previous observed value of week t.

Figure 6.8 and Figure 6.9 show the results of the sensitivity analysis when shifting the time window backward and varying the smoothing parameter.

Shifting the time window backward means that we do not take incoming orders into account. That is, we can assess the impact of considering known incoming orders for production release. Figure 6.8 illustrates both approaches, the 'Hist&Order' approach including incoming known orders for the simple moving average and the 'Hist&Order$_{shift}$' approach considering solely historical data. There is a slight shift to the left when incoming order data is omitted meaning that the 'Hist&Order' approach remains better. Thus, taking known incoming orders into account when defining the release quantity in front end is advisable even though the information is lagged by a period of up to eleven weeks. To give an example, we provide in Table 6.4 the started quantities in front end, the quantities arriving at the master storage, and the incoming orders at the master storage in week t where we assume a processing time of 13 weeks. In week $t = 0$, we start the production in front end according to the average demand of weeks $t = -2$, up to $t = 2$ as shown in row one, column one, where the demand of weeks $t = -2$ and $t = -1$ are observed orders and the demand of weeks $t = 0$, $t = 1$, $t = 2$

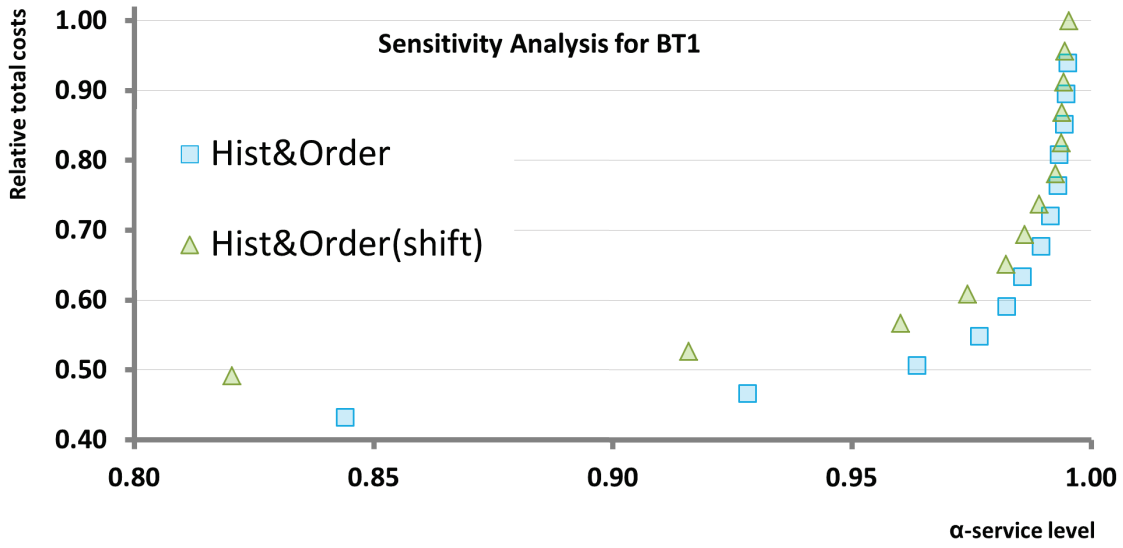


Figure 6.8: Sensitivity analysis for BT1

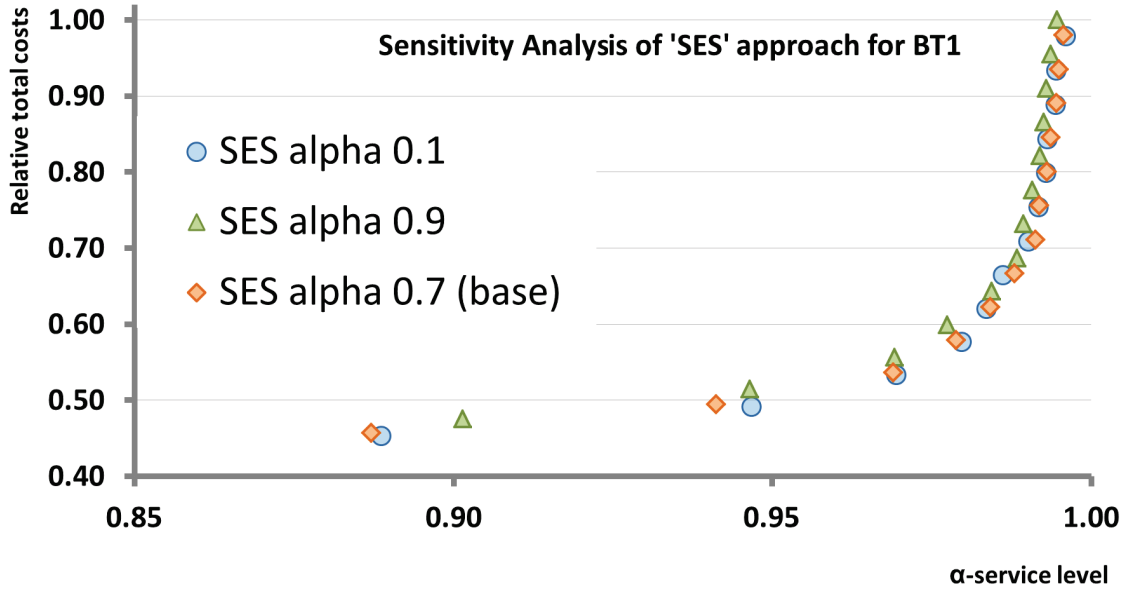


Figure 6.9: Sensitivity analysis for BT1

Table 6.4: Example of production release in front end and incoming orders at the master storage

		master storage	
week t	**production release in front end**	**arriving quantities from production**	**incoming orders**
0	$(o_{-2}+o_{-1}+o_0+o_1+o_2)/5$	$(o_{-15}+o_{-14}+o_{-13}+o_{-12}+o_{-11})/5$	o_0
1	$(o_{-1}+o_0+o_1+o_2+o_3)/5$	$(o_{-14}+o_{-13}+o_{-12}+o_{-11}+o_{-10})/5$	o_1
2	$(o_0+o_1+o_2+o_3+o_4)/5$	$(o_{-13}+o_{-12}+o_{-11}+o_{-10}+o_{-9})/5$	o_2
	...	...	...
13	$(o_{11}+o_{12}+o_{13}+o_{14}+o_{15})/5$	$(o_{-2}+o_{-1}+o_0+o_1+o_2)/5$	o_{13}

are known incoming orders. (Note, an order that arrives at the master storage is known three weeks in advance). Since the processing time from production release up to the master storage takes 13 weeks, this started quantity including the orders of weeks $t = 0$, $t = 1$, and $t = 2$ will arrive at the master storage in week $t = 13$ shown in row five column three. However, the known orders, o_0, o_1, and o_2 already arrive at the master storage in weeks $t = 0$, $t = 1$, and $t = 2$. Thus, the release in front end lags behind the incoming orders by up to eleven weeks. Nevertheless, according to the results in Figure 6.8 it seems advisable to include the known incoming order data.

Varying the smoothing parameter has only little influence on the performance of the 'SES' approach as one can see in Figure 6.9 where we show the 'SES' approach for an α^{SES} of 0.7 (base case), 0.1, and 0.9. It slightly indicates that a smaller chosen α^{SES} performs better than a larger α^{SES} as both approaches, 'SES' with 0.7 and 'SES$_{\alpha=0.1}$', outperform 'SES$_{\alpha=0.9}$'. That is, the forecasted value in week $t+1$ should mainly rely on the forecast error of week t but not on the previous observed value in week t. This result is reasonable as we detected that there is no autocorrelation within the time series. That is, succeeding time periods do not show any correlation.

6.5 Conclusion

Summarizing the chapter, we study various production release approaches as well as stocking strategies for the two exemplary basic types in order to provide Chip Card & Security with recommendations according to their planning procedure. Currently, the production releases in front end are driven by a moving average over historical data for a rather large time horizon, namely of 16 weeks. However, we find that considering a smaller time window of about five weeks ('Hist&Order approach) to compute the moving average or using single exponential smoothing ('SES' approach) decreases the costs by around 40% at the same service level of currently 98%. The cost reduction is due to a lower target reach at the master storage since a 98% service level can be achieved with a target reach of five weeks for the new approaches whereas the current approach has a target reach of 13 weeks. When deciding to keep the current approach, 'Hist', the master storage can still be reduced to around eight weeks at only a slight drop of 0.5% in the α-service level which in turn results in a cost reduction of 20%. Morover, we can conclude from the results that calculating the release quantity by a simple moving average over a large time window, which would lead to a more stable amount of products released in front end, is not preferable since it reacts slower to demand fluctuations.

Thus, we recommend to rather choose a smaller time window, e.g. of five weeks, to determine the amount of wafers to be released in front end in order to accelerate the reaction. This time window should be chosen such that incoming orders which are known three weeks in advance are taken into account.

Chapter 7

Conclusions and recommendations

In section 1.3, we formulate our problem definition along with two sub problems. These require to parametrize the demand generation method in the existing simulation model such that the generated demand accurately describes the observed demand. In order to assess the fit between the observed and generated demand we need to define an appropriate method. The parametrization of the simulation model can then be adjusted such that the fit between the data meets the requirements of the defined method. As a result, we can use the parametrized model to study two exemplary basic types of Chip Card & Security according to their supply chain planning process and propose recommendations as well as improvements of their current practises. To solve these problems we determine our research questions in subsection 1.4.3.

7.1 Conclusion

We summarize the findings regarding the research questions which lead us to solve the problem definition in section 1.3 of this research project. Detailed explanations for each of the questions are given at the end of every chapter.

How is the supply chain planning carried out?
Infineon is a semiconductor company producing chips, sensors, and microcontrollers. These products are characterized by a long cycle time which ranges between one to four months as they require an elaborate processing. To provide a competitive advantage towards other vendors, Infineon aims to keep lead times to customers low. This requires a smooth planning process where capacity and demand are matched accordingly. Due to the long cycle times production usually starts before demand is known in order to reduce the lead time to costumers. That is, one has to make an estimation of how much to start in front end. For the two considered products, a contactbased and a contactless payment chip, the quantity to release in production is defined by using a four months moving average over the historical demand. The released quantity is then processed up to the first storage location in Infineon's supply chain, the master storage, since the products become customer specific in the downstream processing steps. The question arises whether there is a more clever way to define the quantity to release in production and the amount of products to store at the stocking location which we answer during this research by conducting a simulation study. However, the input parameter of the simulation need to be fine tuned according to the considered products such that they reflect reality. We choose simulation for practical reasons since on the one hand there is

an existing simulation model and on the other hand simulation allows for highly complex in interconnected processes as of the supply chain process of Infineon.

How does the demand data of the representative products from CCS behave?
Before we start to answer the questions how to define the amount of wafers to start in production and how much to store at the stocking locations, we aim to analyse the demand data in depth such that we can generate statistical similar demand to use as input in our simulation model. The data analysis is done for each of the two basic types and their corresponding sales products that make up $\geq 85\%$ of the volume. We perform among others a time series decomposition, classification into smooth, intermittent, erratic, and lumpy demand as well as calculate various statistical measures, such as mean, standard deviation, and autocorrelation which provides us with a first insight into the behaviour of the data. The time series decomposition shows that there is neither a trend nor seasonality in the data, that is, fluctuations occur due to a rather random customer demand behaviour than due to reoccurring events. Furthermore, it is noticeable that most of the sales products are classified as lumpy, thus, the demand is fluctuating with rather many periods of no demand. Aggregating the sales product on basic type level shows that fluctuations level of each other and demand occurs more frequently such that it is classified as smooth. Finally, we emphasise to consider autocorrelation in order to detect subsequent periods with increasing demand since they stress the system, that is, the probability of stock outs increase which in turn leads to a lower delivery performance. Our considered products do not show autocorrelation, however there are products from e.g. ATV and PMM showing autocorrelation and thus may require another stocking strategy.

How is the simulation model set up?
At Infineon various approaches exist to determine stocking levels. These range from basic rule of thumbs to enhanced inventory methods using common safety stock calculations on the assumption of a normal distributed lead time to elaborated approaches such as discrete event simulation. Discrete event simulation has the advantage of being able to capture randomness not only in the lead time but also in the demand and process steps and incorporating interactions within the supply chain among manufacturing and stocking locations. In addition, simulation is able to capture time dependent events and conditions. We use an existing discrete event simulation model to examine stocking strategies for two exemplary basic types since the supply chain of Infineon is highly complex and in order to give accurate recommendations we need to capture the interactions of the production process. The simulation model which we adapt to our needs was built by the scenario & econometrics team of Infineon. It models the whole supply chain of Infineon with its various processing steps and stocking points from the releases of the wafers in front end to the finished products in back end. When adapting the simulation model to the two exemplary basic types, we require to parametrize the demand generation method such that it produces demand similar to the observed one in order to receive accurate simulation results.

What solution approaches exist in literature to assess the fit between generated and observed demand data?

The generated demand should match the observed demand such that the output of the simulation model is precise and recommendations are valid. There exists a variety of approaches in literature to compare observed and forecasted data according to their fit. Techniques are forecast accuracy measures, time series similarity measures as well as hypothesis tests. Fore-

cast accuracy measures are point-to-point methods which compare the observed data at time period t with the generated data at time period t. Thus, they do not allow differences in the observed and generated data for time period t even though their overall behaviour according to the mean, standard deviation, and autocorrelation may be similar. However, as we want to capture the overall behaviour they are not suitable. Furthermore, we consider time series similarity measures. Similarly to forecast accuracy measures do these ones compare one to many points, thus giving more freedom, or are based on the periodicity of the data. Since our data does not show periodicity and we consider a one to many point approach as unsuitable to capture the whole behaviour of the data we further look into goodness of fit tests. They allow to test the hypothesis whether two data sets come from the same underlying distributions. As we aim to compare the overall statistical behaviour of two data series we find the two-sample hypothesis test sufficiently for our needs. Hereby, we decide to use the Kolmogorov-Smirnov as well as the Chi-Square test, which are commonly used in practice and rather easy to apply.

How do we need to parametrize the simulation model to create accurate demand data?
The parameters of the demand generation methods need to be set such that we receive generated demand similar to the observed one. In an experimental study we examine whether there are any indications how to set the parameters in order to receive a particular demand behaviour. However, we find that there are no clear rules for parametrizing the demand generation method. Thus, we use an iterative approach by searching for the parameter settings that fit the desired demand data comparing the outcome using a modification of the Kolmogorov-Smirnov approach. Eventually, we evaluate the final chosen demand by the Chi-Square test which shows that the fit between the data series is sufficient. However, not only for the reason that this procedure is quite time consuming but also that the demand generation method is not able to create intermittent nor autocorrelated demand, we recommend to improve the demand generation method. Suggestions are to allow for various distributions, explicitly setting the average-inter-demand interval, as well as incorporating the possibility of autocorrelated demand.

How can the planning process of CCS be improved?
After parametrizing the simulation model accurately we conduct a simulation study to improve the planning process for two exemplary basic types of Chip Card & Security. Thereby, we consider the questions how to define the amount of wafers to start in production and how much to store at the various stocking locations. We look at three production release approaches, the current one which uses a simple moving average over four months, a second one which applies a simple moving average over historical and order data for a time horizon of five weeks and a third one which uses single exponential smoothing. Next to this, we vary the amount of stock at the master storage. Note, for both basic types stock is kept solely at the master storage where the CODP lies since products become customer specific in the downstream processing steps. The simulation study shows that using a smaller time window for calculating the simple moving average or applying single exponential smoothing to quantify the amount of wafers to be released in front end outperforms the current approach. That is, fluctuations are not as smoothed out as with a large time window and thus reactions are faster. Hence, we recommend CCS to shorten their time horizon to at least eight weeks over which they calculate the moving average. With both new approaches the target reach can be reduced from currently 13 to around five weeks while keeping an α-service level of around 98%. This in turn reduces the costs by 40%. In addition, when deciding to keep the

current approach, the target reach can still be reduced to around eight weeks at a marginal drop in the α-service level, however a large cost reduction of 20%. The tendency of being able to reduce stocks was also seen by the supply chain planners of CCS, however there was no particular idea how much stocks can be reduced. Thus, the simulation output confirms their gut feeling and provides an analytical approach of determining the target stock.

7.2 Recommendations

We gave suggestions for improving the supply chain planning of CCS and found a way to create accurate input data to the simulation model by assessing the fit between two time series. However, this research also has some limitations that may be overcome by further research:

- *Demand generation.* We are able to generate the demand using the current demand generation method, however this method is not able to capture all demand patterns, such as autocorrelation and intermittent demand. Therefore, we suggest to improve the demand generation method such that it allows for intermittent and autocorrelated demand. Furthermore, it may be desired to create demand from other distributions such as the Poisson or Gamma distribution.
- *Complex production release approaches.* We quantify the production release amount by a simple moving average over various time horizons and using single exponential smoothing. However, we do not consider other forecasting approaches. Thus, it may be suitable to apply more elaborate forecasting methods such as advanced exponential smoothing techniques, Holt Winter analysis, or ARIMA models [22] in order to quantify the production amount.
- *Machine capacity.* Currently, the simulation model under study presents the processing steps without a capacity restriction for machine availability. That is, machines are always available and we do not build queues in front of the machines. For the considered basic type BT2 this may be a further point to consider, since the production in front end of this basic type is restricted. Thus, we cannot release a higher quantity than the given limit, which in turn restricts the applicability of the production release approaches.
- *Idle costs.* The defined total costs are based on the WIP and the weighted average cost of capital, but do not consider idle costs. Idle costs occur when machine capacity is not fully used for further production. Another approach of evaluating strategies may then be determined by balancing bind capital in inventory versus reducing idle costs.
- *Autocorrelated products.* We find in an constructed example that the probability of stock outs for autocorrelated products is approximately 55.48% higher than for non-autocorrelated products. Therefore, we strongly recommend to analyse the autocorrelation for the observed and generated data as planning strategies may differ. To confirm this observation we suggest to simulate the planning procedures with autocorrelated and non-autocorrelated demand such that one can detect whether different strategies perform better in either case.
- *Simulation at ATV/PMM/IPC.* Simulation results are individual and currently done for two products of CCS. The remaining divisions ATV, PMM, and IPC, may also benefit from using discrete event simulation for analysing their planning processes. This may

not be done for two particular products but products may be grouped into categories with certain specifics which may then be simulated and results can be taken as an overall tendency. At present, a project with ATV is initiated to examine the target levels of product groups using the build simulation model.

- *Comparing results with simple approaches.* So far, we proposed to use simulation in order to improve the planning process. Nevertheless, simple approaches such as 'one size fits all', which estimates the target stock levels by a rule of thumb, as well as the 'enhanced inventory management', that uses commonly known stocking policies such as the (R, S) policy, require considerably less effort and may still lead to similar results. Thus, it is advisable to assess the outcome of different approaches according to their performance improvement and effort needed.

Bibliography

[1] A. M. A. Adriaansen. *Balancing inventory and equipment contingencies by a flexible semiconductor supply chain model*. Master thesis, University of Twente, Enschede, August 2015. 8

[2] T. W. Anderson and D. A. Darling. Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212, 1952. 46

[3] T. W. Anderson and D. A. Darling. A test of goodness of fit. *Journal ofthe American Statistical Association*, 49(268):765–769, 1954. 46

[4] H. André-Jönsson and D. Z. Badal. Using signature files for querying time-series data. In Jaime G. Carbonell, Jörg Siekmann, G. Goos, J. Hartmanis, J. Leeuwen, Jan Komorowski, and Jan Zytkow, editors, *Principles of Data Mining and Knowledge Discovery*, volume 1263 of *Lecture Notes in Computer Science*, pages 211–220. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997. 42

[5] T. B. Arnold and J. W. Emerson. Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3(2):34–39, 2011. 52

[6] S. Axsäter. *Inventory control*, volume 90 of *International series in operations research & management science*. Springer, New York, 2nd ed. edition, 2006. xvii, 17, 62, 64

[7] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KKD Workshop: 1994*, pages 359–370. 42

[8] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time series analysis: Forecasting and control*. Wiley series in probability and statistics. John Wiley & Sons, Inc., Hoboken, NJ, 4th edition edition, 2008. 22, 36, 37

[9] A. O. Brown, H. L. Lee, and R. Petrakian. Xilinx improves its semiconductor supply chain using product and process postponement. *Interfaces*, 30(4):65–80, 2000. 2, 3

[10] T. A. Burgin. The gamma distribution and inventory control. *Operational Research Quarterly (1970-1977)*, 26(3):507, 1975. 17, 57

[11] J. M. Charnes, H. Marmorstein, and W. Zinn. Safety stock determination with serially correlated demand in a periodic-review inventory system. *The Journal of the Operational Research Society*, 46(8):1006, 1995. 23, 58

[12] L. Chen, M. T. Özsu, and V. Oria. Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 491–502. 42

[13] M. Comella López-Pinto. *Simulation-based Supply Chain Segmentation.* Master thesis, Technical University of Munich, Munich, October 2017. 3, 8, 17, 27, 33, 49

[14] J. D. Croston. Stock levels for slow-moving items. *Operational Research Quarterly (1970-1977)*, 25(1):123–130, 1974. 21

[15] R. B. D'Agostino and M. A. Stephens. *Goodness-of-Fit-Techniques.* Marcel Dekker, Inc., New York, 1986. 46

[16] D. A. Darling. The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, 28(4):823–838, 1957. 43

[17] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008. 41, 42

[18] A. H. C. Eaves. *Forecasting for the ordering and stock holding of consumable spare parts.* Phd thesis, Lancaster Univesity, Lancaster, UK, 2002. 35

[19] Ehm, H., Infineon Technologies AG. General introduction to scor at infineon technologies ag: The supply chain operations reference model: e-learning. xvii, 11, 12, 14, 15

[20] R. Ehrhardt. The power approximation for computing (s , s) inventory policies. *Management Science*, 25(8):777–786, 1979. 17

[21] S. Eirich. Segmentation simulation @as se of infineon technologies ag: Internal power point presentation. 17

[22] T. England. Time series decomposition, Spring Semester 2015/16. 21, 36, 37, 38, 39, 41, 78

[23] N. Erkip, W. H. Hausman, and S. Nahmias. Optimal centralized ordering policies in multi-echelon inventory systems with correlated demands. *Management Science*, 36(3):381–392, 1990. 22, 23

[24] P. Esling and C. Agon. Time-series data mining. *ACM Computing Surveys*, 45(1):1–34, 2012. 41, 42

[25] L. Fortuin. Five popular probability density functions: A comparison in the field of stock-control models. *The Journal of the Operational Research Society*, 31(10):937, 1980. 17

[26] S. Fotopoulos, M.-C. Wang, and S. S. Rao. Safety stock determination with correlated demands and arbitrary lead times. *European Journal of Operational Research*, 35(2):172–181, 1988. 23

[27] T. Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011. 41

[28] P. Goodwin and R. Lawton. On the asymmetry of the symmetric mape. *International Journal of Forecasting*, 15(4):405–408, 1999. 40

[29] J. N. D. Gupta, R. Ruiz, J. W. Fowler, and S. J. Mason. Operational planning and control of semiconductor wafer production. *Production Planning & Control*, 17(7):639–647, 2006. 1, 3

[30] H. Heerkens. A methodological checklist for the high-tech marketing project, February 2004. 5

[31] B. Heiermann. Interview with b. heiermann (ccs) by f. federmann and s. lingelbach: Supply chain and demand planning at ccs. pdf file, 21.12.2016. 15, 16, 19, 25

[32] R. J. Hyndman. Another look at forecast accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting*, (4):43–46, 2006. 39, 40, 41

[33] R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts.com, April 2014. 21, 22, 38, 39, 40

[34] R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. 39, 40, 41

[35] Infineon Technologies AG. Smape: Symmetric mean absolute percentage error: Internal power point presentation. 40

[36] Infineon Technologies AG. Annual report 2016, 30. September 2016. 2, 15

[37] J. Jackson. Interview with j. jackson (ccs) by s. lingelbach: Demand planning at ccs. pdf file, 23.11.2016. 14, 15, 16

[38] G. J. Janacek, A. J. Bagnall, and M. Powell. A likelihood ratio distance measure for the similarity between the fourier transform of time series. In Tu Bao Ho, David Cheung, and Huan Liu, editors, *Advances in knowledge discovery and data mining*, volume 3518 of *Lecture notes in computer science. Lecture notes in artificial intelligence, 0302-9743*, pages 737–743. Springer, Berlin and Great Britain, 2005. 42

[39] E. Keogh and S. Kasetty. On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371, 2003. 41

[40] H.-S. Lau and M.-C. Wang. Estimating the lead-time demand distribution when the daily demand is non-normal and autocorrelated hon-shiang lau. *European Journal of Operational Research*, 29(1):60–69, 1987. 23

[41] A. M. Law. *Simulation modeling and analysis*. McGraw-Hill series in industrial engineering and management science. New York, 5th edition edition, 2015. xix, 17, 27, 43, 44, 45, 57, 59, 65, 66, 70, 96, 97, 99, 101

[42] A. M. Law and J. S. Carson. A sequential procedure for determining the length of a steady-state simulation. *Operations Research*, 27:1011–1025, 1979. 65

[43] H. L. Lee, V. Padmanabhan, and Whang. S. The bullwhip effect in supply chains. *Sloan Management Review*, 38(3):93–102, 1997. 2

[44] S. Makridakis. Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, 9(4):527–529, 1993. 40

[45] G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8):114, 1965. 2

[46] NIST/SEMATECH. *e-Handbook of Statistical Methods*. U.S. Department of Commerce, 2003. 37, 38, 45

[47] D. Panchenko. Kolmogorov-smirnov test, Fall 2006. 46

[48] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series*, 5(50):157–175, 1900. 43

[49] A. N. Pettitt. A two-sample anderson-darling rank statistic. *Biometrika*, 63(1):161–168, 1976. 46

[50] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The art of scientific computing*. Cambridge University Press, Cambridge and New York and Port Chester and Melbourne and Sydney, 2nd edition, 1992. 44, 56, 96

[51] V. J. Presutti and R. C. Trepp. More ado about economic order quantities (eoq). *Naval Research Logistics Quarterly*, 17(2):243–251, 1970. 17

[52] K. Ramaekers and G. K. Janssens. On the choice of a demand distribution for inventory management models. *European J. of Industrial Engineering*, 2(4):479, 2008. 57

[53] G. K. Rand, R. Peterson, and E. A. Silver. Decision systems for inventory management and production planning. *The Journal of the Operational Research Society*, 30(12):1133, 1979. 57

[54] L. Sachs. *Angewandte Statistik: Anwendung statistischer Methoden*. Springer, Berlin, Heidelberg, 11th edition, 2004. 46

[55] C. Schiller, T. Ponsignon, and G. Yachi. Plan porcess at infineon technologies ag: Introduction of planning; supply planning; demand planning: e-learning. xvii, xvii, 11, 12, 13, 14

[56] R. E. Shannon. *Systems simulation: The art and science*. Prentice-Hall, Englewood Cliffs and London, 1975. 1

[57] E. A. Silver, D. F. Pyke, and D. J. Thomas. *Inventory management and production planning and scheduling*. John Wiley & Sons, Inc, third edition edition, 1998. 17, 39, 40, 56

[58] A. Stuart, J. K. Ord, S. F. Arnold, and M. G. Kendall. *Kendall's advanced theory of statistics. Volume 2A, Classical inference and the linear model*. Kendall's library of statistics. John Wiley & Sons, Inc., Chichester, 6th ed. / alan stuart, j. keith ord, steven arnold edition, 2008. 43

[59] A. A. Syntetos, J. E. Boylan, and J. D. Croston. On the categorization of demand patterns. *Journal of the Operational Research Society*, 56(5):495–503, 2005. xvii, 20, 35, 36, 57

[60] H. Tempelmeier. *Bestandsmanagement in Supply Chains*. Books on Demand, Norderstedt, 2012. 32, 94, 95

[61] J. E. Tyworth and L. O'Neill. Robustness of the normal approximation of lead-time demand in a distribution setting. *Naval Research Logistics*, 44(2):165–186, 1997. 17, 56

[62] M. Vlachos, P. Yu, and V. Castelli. On periodicity detection and structural periodic similarity. *SDM*, 5:449–460, 2005. 42

[63] T. M. Williams. Stock control with sporadic and slow-moving demand. *Journal of the Operational Research Society*, 35:939–948. 35

[64] J. M. Wooldridge. *Introductory econometrics: A modern approach*. South-Western Cengage Learning, Mason OH, 5th edition edition, 2013. 21, 23, 89

Appendix A

Correlation among sales products

In Table A.1, we give the correlation among the six biggest sales products for basic type BT1. The correlation coefficient, with a range of [-1, 1] indicates whether two data series are positive or negative correlated, that is, whether they influence one another or not. For example, if there is a high positive correlation between two sales products one could conclude, that these two sales products get often ordered together. However, in our case the correlation coefficient fluctuates around 0, and hence we can conclude that there are no correlations among the sales products.

Table A.1: Correlation matrix for the six biggest sales products of the basic type BT1

	SP1	**SP2**	**SP3**	**SP4**	**SP5**	**SP6**
SP1	1.00	0.35	-0.01	-0.36	-0.33	-0.09
SP2	0.35	1.00	-0.04	-0.22	-0.15	-0.01
SP3	-0.01	-0.04	1.00	-0.09	-0.23	-0.24
SP4	-0.36	-0.22	-0.09	1.00	0.48	0.12
SP5	-0.33	-0.15	-0.23	0.48	1.00	0.10
SP6	-0.09	-0.01	-0.24	0.12	0.10	1.00

Appendix B

Decomposition of time series

Decomposition of time series for BT1 into a trend, seasonal component and error term.

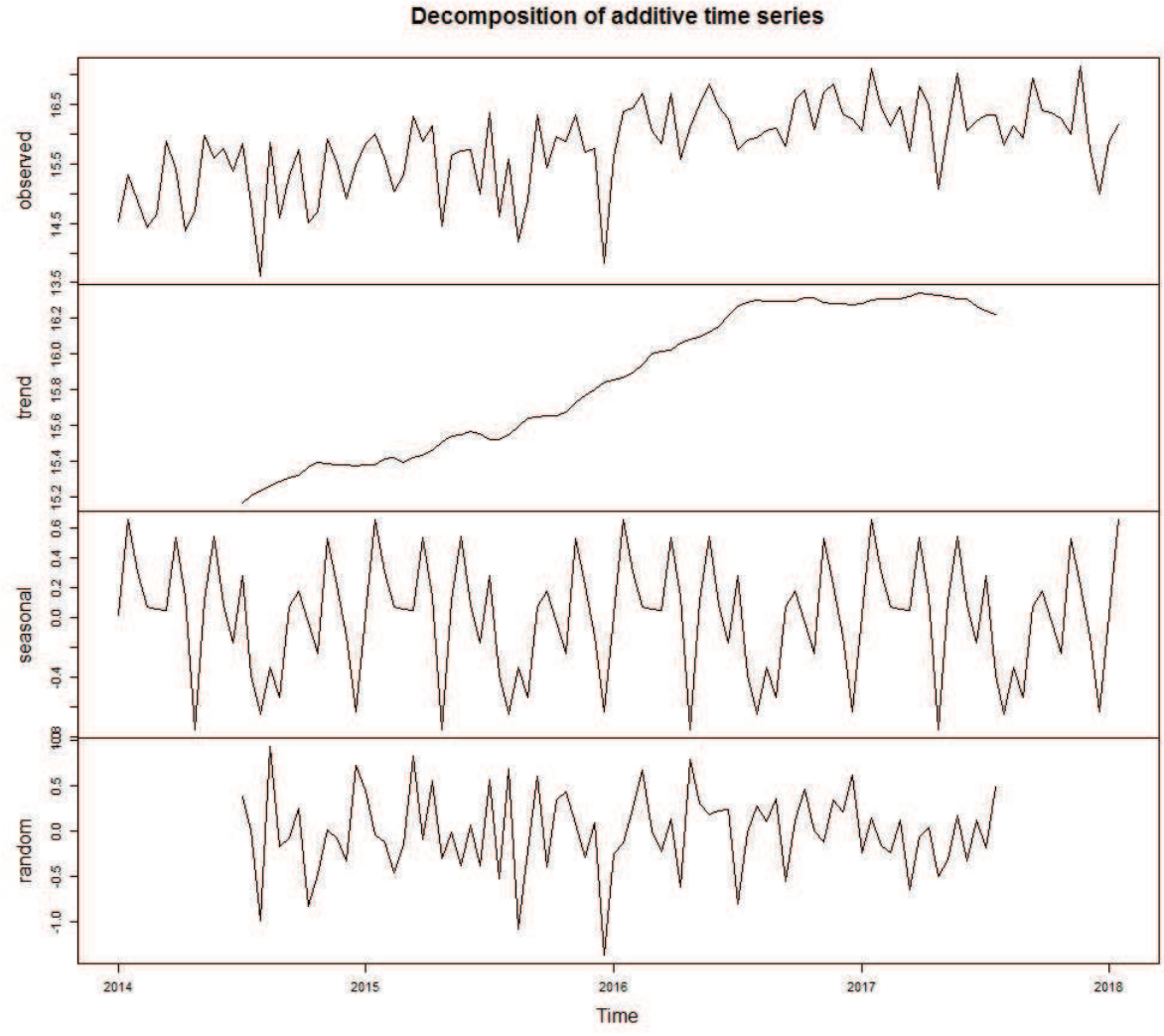


Figure B.1: Additive decomposition of time series data for product BT1

Appendix C

Autocorrelation

C.1 Autocorrelation threshold value

In order to determine whether there is significant autocorrelation within a data series one can compute a threshold value. This threshold value is approximated by an Hypothesis test, which tests $H_0 : r_j = 0$ against $H_1 : r_j \neq 0$. The null hypothesis would be rejected with a 5%-level of significance if:

$$|r_k| > \frac{1.96}{\sqrt{T}} \tag{C.1}$$

under the assumption that $Y_t \sim N(0, \sigma^2)$ and where T is the number of data points [64].

C.2 Autocorrelated products at Infineon

In Figure C.1, Figure C.2,and Figure C.3, we show the autocorrelation for lags zero to 20 of three exemplary products computed in R by the function 'acf'. The dashed line indicates whether the autocorrelation is significant. That is, if the autocorrelation exceeds the dashed line as it is the case for the three products, the autocorrelation is significant. Further note, that lag zero always has a autocorrelation of one since the time series with itself (no lagged values) is perfectly correlated. How the significance level is specified is described in section C.1.

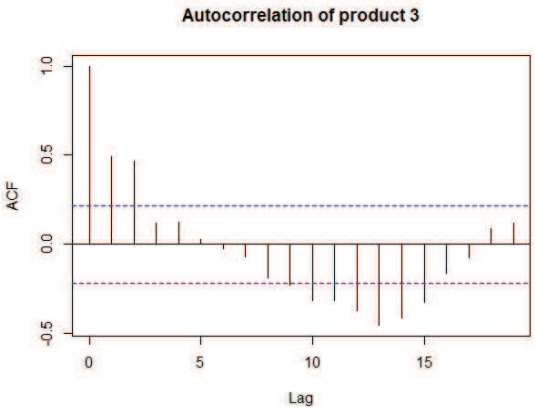


Figure C.1: Autocorrelation for lags 1 to 20 of product 1

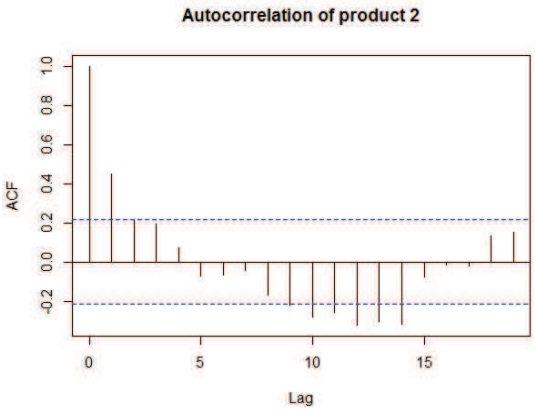


Figure C.2: Autocorrelation for lags 1 to 20 of product 2

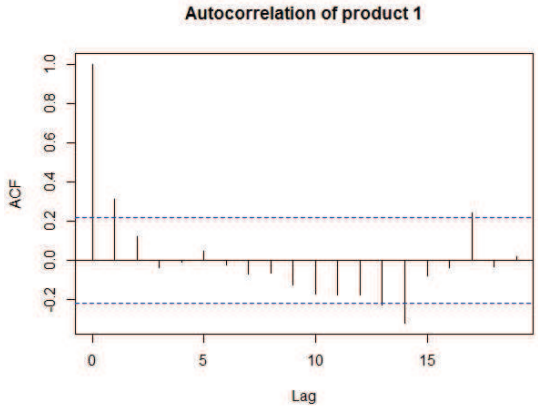


Figure C.3: Autocorrelation for lags 1 to 20 of product 3

C.3 Stock outs in existence and non existence of autocorrelation

Table C.2 and Table C.3 show the analysis of the number of stock outs, denoted as 'Sum LostOpp' for generated data with and without autocorrelation. We show the first 30 time periods out of 1040 periods. The autocorrelation for the first four legs is given in Table C.1.

From our previous analysis in subsection 2.1.2 we determine the cycle time to be approximately 10 weeks from the start of production in the fabrication up to the Masterlager where customer demand arrives. In order to produce a sufficient amount of basic type units during this lead time of 10 weeks we determine the production quantity in time period $t = 11$ to be the average demand over 10 weeks that is from $t = 1$ up to $t = 10$. The production output in time period $t = 20$ is then defined as the average demand over time periods $t = 1$ up to $t = 10$. The production output for the periods one to twenty is the average demand over all periods. The demand for the non autocorrelated series is generated by drawing a random number between the minimum and the maximum value of the autocorrelated series, where

the autocorrelated series is generated by a sine function and some random term. The stock at each period t is calculated by subtracting the demand of period t from the stock at period $t-1$ plus the production output at t. In case that demand exceeds the stock of period $t-1$ plus the production output at t, it is set to 0. The number of stock outs is calculated by summing the number of periods where the stock is 0.

Table C.1: Autocorrelation of first four legs for non autocorrelated and autocorrelated demand

	AC lag1	**AC lag2**	**AC lag 3**	**AC lag 4**
Non autocorrelated demand	-0.016	-0.032	0.009	-0.007
Autocorrelated demand	0.311	0.310	0.281	0.230

Table C.2: One example of the first 30 periods out of 1040 periods for non autocorrelated demand

time period t	**Production output**	**Non autocorrelated demand**	**Stocks**	**Lost opportunity**	**Sum LostOpp**
1	151	117	302	0	5
2	151	158	295	0	
3	151	210	236	0	
4	151	79	308	0	
5	151	91	368	0	
6	151	181	338	0	
7	151	121	368	0	
8	151	146	373	0	
9	151	183	341	0	
10	151	83	409	0	
11	151	178	382	0	
12	151	76	457	0	
13	151	222	386	0	
14	151	141	396	0	
15	151	201	346	0	
16	151	198	299	0	
17	151	96	354	0	
18	151	138	367	0	
19	151	195	323	0	
20	151	175	299	0	
21	137	148	288	0	
22	143	161	270	0	
23	135	217	188	0	
24	136	85	239	0	
25	142	94	287	0	
26	153	158	282	0	
27	155	127	310	0	
28	152	216	246	0	
29	152	151	247	0	
30	153	162	238	0	
...	...	...	...	...	
1040	...	...	...	...	

Table C.3: One example of the first 30 periods out of 1040 periods for autocorrelated demand

time period t	**Production output**	**Non autocorrelated demand**	**Stocks**	**Lost opportunity**	**Sum LostOpp**
1	150	135	301	0	13
2	150	158	293	0	
3	150	133	310	0	
4	150	156	304	0	
5	150	179	275	0	
6	150	168	257	0	
7	150	153	254	0	
8	150	153	251	0	
9	150	127	274	0	
10	150	212	212	0	
11	150	156	206	0	
12	150	145	211	0	
13	150	153	208	0	
14	150	199	159	0	
15	150	160	149	0	
16	150	148	151	0	
17	150	155	146	0	
18	150	159	137	0	
19	150	227	60	0	
20	150	197	13	0	
21	157	220	0	1	
22	160	190	0	1	
23	158	208	0	1	
24	160	207	0	1	
25	165	145	20	0	
26	163	190	0	1	
27	161	141	20	0	
28	161	131	50	0	
29	161	125	86	0	
30	171	143	114	0	
...	...	...	...	...	
1040	...	...	...	...	

Appendix D

The β- and γ-service level

The β-service level provides the proportion of demand (not including backorders) during a period t that is fulfilled by on-hand inventory and thus delivered without delay [60]. To give an example, if the average demand during a period is 50 units and the average backorder quantity during a period is 5 units, then the β-service level becomes 90% which gives the proportion of demand that is fulfilled immediately. Often the β-service level is also described as fill rate. It is calculated by [60]:

$$\beta\text{-sevice level} = 1 - \frac{\text{E(backlog at the end of time period t)}}{\text{E(demand during time period t)}} \tag{D.1}$$

In the simulation model the β-service level is described by considering the change in the backorders between week $t-1$ and week t. In case that the backorders at the end of week t increase compared to week $t-1$ we could not deliver all demand from stock, thus the β-service level will decrease accordingly. The backorders bo_{rpT} at the end of week T are the difference between the orders o_{rpt} up to the end of week T and the deliveries d_{rpt} up to the end of week T:

$$bo_{rpT} = max\left[\sum_{t=1}^{T}\left(o_{rpt} - d_{rpt}\right), 0\right] \tag{D.2}$$

The β-service level in week t for product p and replication r is then defined by the delta in the backorders between week $t-1$ and week t. Thereby, it compares the amount of backorders for week t to the ordered quantity in week t. If the backorders are rather small compared to the ordered quantity, then the β-service level will be rather high and vice versa, if the backorder quantity is rather large compared to the ordered quantity, the β-service level will be rather small.

$$\beta\text{-sevice level}_{rpt} = max\left[1 - \frac{max\left(bo_{rpt} - bo_{r,p,t-1}, 0\right)}{max\left(o_{rpt}, 1\right)}, 0\right] \tag{D.3}$$

In order to compare the various stocking strategies, we aggregate the β-service level across all sales products ($p = 1...P$), weeks ($t = 1...T$), and replications ($r = 1...R$):

$$\beta\text{-sevice level} = \frac{\sum_{r=1}^{R}\sum_{p=1}^{P}\sum_{t=1}^{T}\beta_{rpt}}{R * P * T} \tag{D.4}$$

The γ-service level extends the β-service level by also taking backorders from previous weeks into account. Thus, it measures not only the amount of backorders, but also the time needed to fulfill all backorders and thus to recover from a high demand period. That is, it relates the expected sum of the accumulated backorders at the end of period t (backlog level) to the expected demand during period t. It is defined by [60]:

$$\gamma\text{-sevice level} = 1 - \frac{\text{E(backlog level at the end of time period t)}}{\text{E(demand during time period t)}} \tag{D.5}$$

E.g., during a period of five weeks with an expected demand of 50 units per period backlog occurs in week four and five of 30 and 40 units each. The accumulated backlog in week four is 30 units and in week five is $30 + 40 = 70$ units. The expected backlog level is then the sum of the accumulated backlogs divided by the period length: $(30 + 70)/5 = 20$. Thus, the γ-service level can be calculated by $1 - (20/50) = 60\%$ [60].

In the simulation model the γ-service level is described for each week t, product p, and replication r by the backorders and the incoming orders up to week T for product p and replication r:

$$\gamma\text{-sevice level}_{rpT} = max\left[1 - \frac{bo_{rpT}}{max(o_{rpT}, 1)}, 0\right] \tag{D.6}$$

When aggregating it over all sales products ($p = 1...P$), weeks ($t = 1...T$), and replications ($r = 1...R$), we define the service level as follows:

$$\gamma\text{-sevice level} = \frac{\sum_{r=1}^{R}\sum_{p=1}^{P}\sum_{t=1}^{T}\gamma_{rpt}}{R * P * T} \tag{D.7}$$

Appendix E

Chi-Square test for evaluating the fit between the observed and generated data

We apply the Chi-Square test to the observed and generated data that gives the smallest total area between the cumulative distribution functions according to the modified Kolmogorov-Smirnov scheme. The Chi-Sqare statistic is computed from [50]:

$$\chi^2 = \sum_{i=1}^{k} \frac{(R_i - S_i)^2}{R_i + S_i} \tag{E.1}$$

where R_i is the number of observations in the ith bin for the observed data and S_i is the number of observations in the ith bin for the generated data. The critical value is drawn from a Chi-Square distribution with $k-1$ degrees of freedom and a significance level of α, where k is the total number of bins. The number of bins is defined by the square root rule $\sqrt{n}$ where n is the total number of observations [41]. Table E.1 and Table E.2 provide the computed values for the sales products of the two exemplary basic types. We choose a significance level of $\alpha = 1\%$.

Table E.1: Chi-square statistic and critical value for the sales products of basic type BT1

BT1	**SP1**	**SP2**	**SP3**	**SP4**	**SP5**	**SP6**
Chi-square statistic	7.16	6.08	14.62	3.91	4.85	41.63
Critical value	18.48	21.67	21.67	16.81	18.48	23.21

Table E.2: Chi-square statistic and critical value for the sales products of basic type BT2

BT2	**SP1**	**SP2**	**SP3**	**SP4**	**SP5**	**SP6**	**SP7**	**SP8**	**SP9**	**SP10**
Chi-square statistic	6.61	17.15	11.81	4.38	8.31	14.14	66.57	13.91	7.19	2.77
Critical value	21.67	23.21	21.67	13.28	16.81	16.81	20.09	15.09	15.09	24.72

Appendix F

Number of replications and warmup period of simulation study

F.1 Defining the number of replications

In the following, we describe in detail how we define the number of replications using the Replication/Deletion Approach, which is an approximation, as well as verifying this by the exact algorithm of the Sequential Procedure. For these explanations we refer to [41].

The output of a simulation are realizations of identical independent distributed random variables which vary due to the stochastic nature of the simulation. Due to the variations in the output it is necessary to run the simulation several times such that we can capture the true characteristics of the simulation model. Therefore, we aim to reduce the width of the confidence interval for the mean of the performance measure X such that it becomes sufficiently small. The width can be reduced by increasing the number of replications n. The half width of the confidence interval is given by:

$$\frac{t_{n-1,1-\alpha/2}\sqrt{\frac{S_n^2}{n}}}{\bar{X}_n} \tag{F.1}$$

where the sample mean $\bar{X}_n$ and sample variance S_n^2 of the performance measure can be computed from:

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \text{ and } S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2 \tag{F.2}$$

The confidence interval should become sufficiently small. In this context 'sufficiently small' is given by the relative error $\gamma = |\bar{X}_n - \mu|/\mu$. That is, that the sample mean $\bar{X}_n$ should not deviate more than γ from the true μ. Since we have to estimate γ by $\gamma = |\bar{X}_n - \mu|/\bar{X}_n$, we use the corrected target value $\gamma\prime = \gamma/(1+\gamma)$. This is then be used to determine the minimum number of replications n^* for which the estimated relative error is $\leq \gamma\prime$:

$$n^* = \min\left\{ i \geq n : \frac{t_{i-1,1-\alpha/2}\sqrt{\sqrt{S_n^2/i}}}{|\bar{X}_n|} \leq \gamma\prime \right\} \tag{F.3}$$

Rewriting the above term as follows, we can estimate n^* by Equation F.8:

$$\left(\frac{t_{i-1,1-\alpha/2}\sqrt{S_n^2/i}}{|\bar{X}_n|\gamma\prime}\right) \leq 1 \tag{F.4}$$

$$\left(\frac{t_{i-1,1-\alpha/2}\sqrt{S_n^2/i}}{|\bar{X}_n|\gamma\prime}\right)^2 \leq 1^2 \tag{F.5}$$

$$\left(\frac{t_{i-1,1-\alpha/2}}{|\bar{X}_n|\gamma\prime}\right)^2 * \frac{S_n^2}{i} \leq 1 \tag{F.6}$$

$$\left(\frac{t_{i-1,1-\alpha/2}}{|\bar{X}_n|\gamma\prime}\right)^2 * S_n^2 \leq i \tag{F.7}$$

$$\left(\frac{t_{i-1,1-\alpha/2}}{|\bar{X}_n|\gamma\prime}\right)^2 * S_n^2 \leq n^* \tag{F.8}$$

In case of the Replication/Deletion Approach, we start with running the experiments for some n (not too large) and estimate $\bar{X}_n$ as well as S_n^2 for our performance measure, the α-service level. Note, we do not define the number of replications for each single experiment due to the large amount of experimental settings, but restrict ourselves to three stocking strategies per production release approach. That is, we look at nine experimental set ups. The chosen stocking strategies, which we denote by (X, Y, Z) where X, Y, and Z are the target reach at the master storage, die bank and distribution centre, are strategies $(1, 0, 0)$, $(4, 0, 0)$, and $(13, 0, 0)$. We run the experiments for $n = 15$ replications. Computing the values of $\bar{X}_{15}$ and S_{15}^2 for each experiment, we can calculate n^* from Equation F.8.

Table F.1 provides the number of replications for both basic types according to the Replication/Deletion Approach.

Table F.1: Number of replications according to the Replication/Deletion Approach for both basic types

	BT1			**BT2**		
Estimated # replications	**Hist**	**Hist&Ord**	**FC**	**Hist**	**Hist&Ord**	**FC**
(1,0,0)	12	7	7	4	5	4
(4,0,0)	8	3	10	2	4	3
(13,0,0)	6	2	6	2	3	2

The Sequential Procedure does not run many replications at once, but evaluates the width of the confidence interval given by Equation F.1 after each additional replication according to the corrected target value $\gamma\prime$. That is, after each run, we evaluates the confidence interval and stop for the n where it holds that:

$$\frac{t_{n-1,1-\alpha/2}\sqrt{\frac{S_n^2}{n}}}{\bar{X}_n} \leq \gamma\prime \tag{F.9}$$

In Table F.2 we give the number of replications according to the Sequential Procedure for both basic types. They are slightly different to the Replication/Deletion Approach, but

are not much larger. Nevertheless, as we do not evaluate the number of replications for each experiment, we decide to take these values as an indication and run each experiment for 15 replications.

Table F.2: Number of replications according to Sequential Procedure for both basic types

	BT1			**BT2**		
Exact # replications	**Hist**	**Hist&Ord**	**FC**	**Hist**	**Hist&Ord**	**FC**
(1,0,0)	14	7	8	7	5	6
(4,0,0)	10	5	8	6	3	5
(13,0,0)	8	4	6	5	3	4

F.2 Determining the warmup period

For a system with steady state behaviour, we do not collect the performance measures from the beginning of the simulation where the system is still empty, but start collecting data when the system is warmed up and hence capture the true characteristics of the steady state behaviour. Welch's method for determining the warmup period is based on the idea to plot the moving averages of the observations and choose the time interval as warmup period where the observations appear to converge. Describing it in more detail, we start with calculating the mean of the ith observation ($i = 1, 2, ..., m$) from the jth replication ($j = 1, 2, ..., n$). The mean $\bar{Y}_i = \sum_{j=1}^{n} Y_{ji}/n$ per observation i is taken since it reduces the variance of the averaged process $\bar{Y}_1, \bar{Y}_2, ...$ but still reflects the transient behaviour when the system is empty. Next, to smooth out high-frequency oscillations in the averaged process but keeping the longterm trend, we calculate the moving averages $\bar{Y}_i(w)$ with a window w. These moving averages are then plotted and the time period where the $\bar{Y}_1(w), \bar{Y}_2(w), ...$ seem to converge is chosen as warmup period [41]. When applying this procedure to our case, we take $m = 156$ observations, that is, we collect the data each week over three years and choose $n = 10$ replications as this is suggested to be sufficiently large. Next, we calculate the moving averages for the windows $w = 4$, $w = 10$, and $w = 15$ and plot them into a graph shown in Figure F.1 to define the warmup period. From the graph one can see that the moving averages $\bar{Y}_i(w)$ converge where i is approximately 40. As we do not evaluate all experimental settings according to their warmup period, we decide to increase the warmup period up to 52 weeks, which we consider to be sufficiently large.

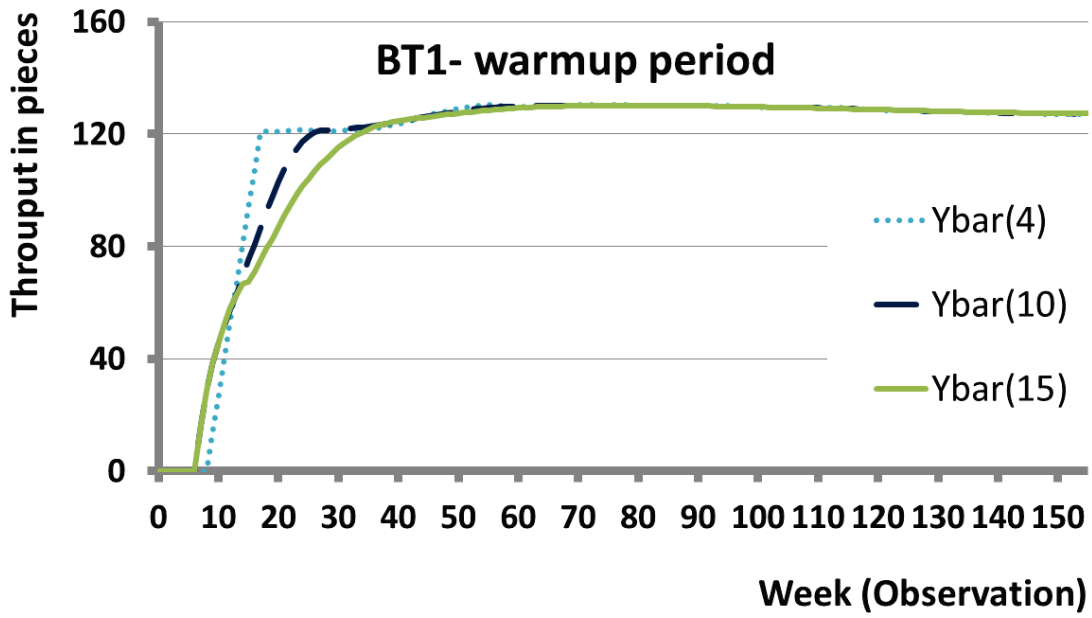


Figure F.1: Graphical method of Welch for determining the warmup period on the example of the basic type BT1

Appendix G

Comparing two system configurations using the paired-t approach

Two system configurations can be compared on the basis of some performance measure by forming the confidence interval for the difference in the expectations [41]. In case they do significantly differ, the confidence interval will not contain zero. On the other hand, in case they do not significantly differ, the CI will contain zero. When two systems are correlated, that is, they use the same common random numbers, one can apply the paired-t approach. Introducing correlation between two system configurations has the advantage of reducing the variance which in turn leads to a smaller confidence interval. For the two system configurations with observations X_j (j=1,...,n) and Y_j (j=1,...,m), where $n = m$, $\mu_X = E(X)$, and $\mu_Y = E(Y)$, we can construct the 100(1-α)% confidence interval for the difference in the expectations $\zeta = \mu_X - \mu_Y$. Therefore, we define

$$W_j = \bar{X}_j - \bar{Y}_j \tag{G.1}$$

$$\bar{W} = \frac{1}{n}\sum_{j=1}^{n} W_j \tag{G.2}$$

$$Var(\bar{W}) = \frac{1}{n(n-1)}\sum_{j=1}^{n}[W_j - \bar{W}]^2 \tag{G.3}$$

where the confidence interval can be found from:

$$\bar{W} \pm t_{n-1,1-\alpha/2}\sqrt{Var(\bar{W})} \tag{G.4}$$

Appendix H

Further results of simulation study for product BT2

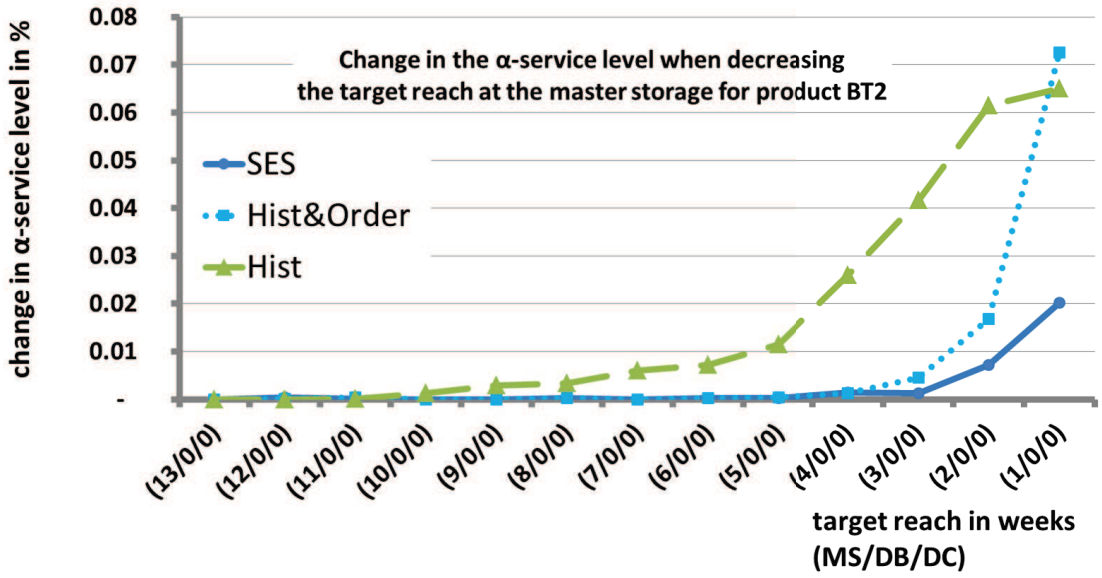


Figure H.1: Change in the α-service level when decreasing the target reach at the master storage for basic type BT2

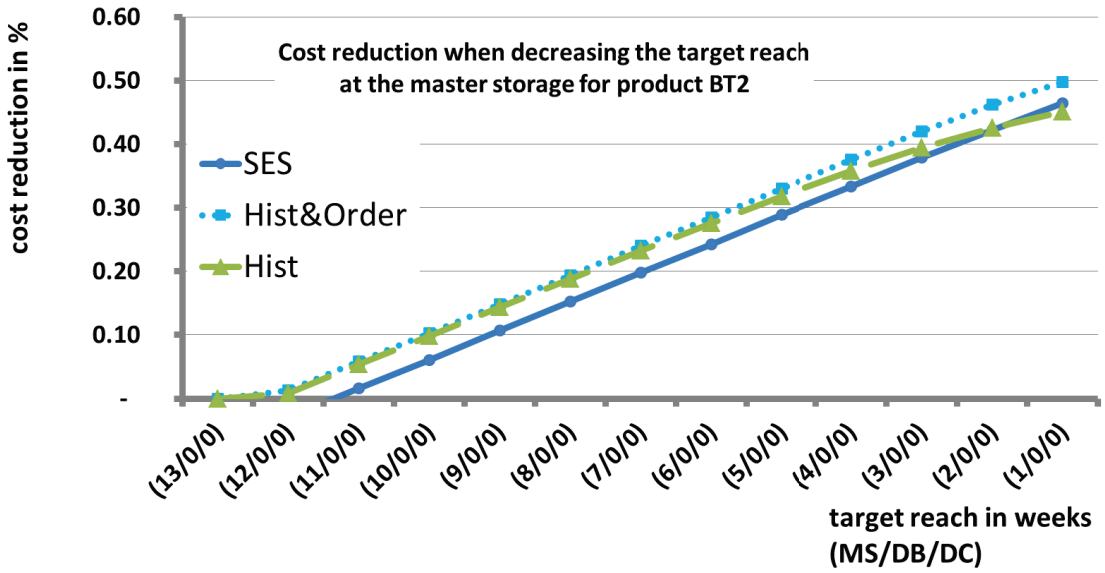


Figure H.2: Reduction in costs compared to the current costs when decreasing the target reach at the master storage for basic type BT2