DEC 2018 - NO. 4

BAVARIAN JOURNAL OF APPLIED SCIENCES®

A Simulation-Based System for Calculating Optimal Numbers of Forklift Drivers in Industrial Plants

Mohammed Alnahhal* **Diane Ahrens***

ABSTRACT

This article describes an optimization method for the material handling system of forklifts using queueing theory and simulation. The objective is to reduce several types of wastes such as waiting time, capacity costs, delayed job orders, and transportation costs. A certain IT infrastructure is assumed, such as monitors for forklift drivers to check on the current job orders at different work stations. Mathematical equations are used to find initial upper and lower limits for the needed capacity levels over the day. Subsequently, a simulation for different levels of capacity within the range of theoretical results is performed in order to find the exact man hour needed for different jobs sequencing strategies. On the basis of the statistical software R, a tool is devised to provide companies with the results for the different parameters they determine. These results show the effects of using batching, considering limitation of line-side space, and the empty drives reduction strategy on performance measures. The strategy of reducing empty drives by identifying the nearest workstation needing a job order was found to be not efficient since it increases the needed capacity. This is due to the fact that the variability of waiting time and hence the late orders percentage are increased.

Dieser Artikel beschreibt eine Optimierungsmethode für ein Materialtransportsystem von Gabelstaplern mittels Warteschlangentheorie und Simulation. Ziel ist es, verschiedene Arten von Verschwendung bei den Kapazitätskosten, verspäteten Arbeitsaufträgen und Transportkosten zu reduzieren. Es wird eine gewisse IT-Infrastruktur angenommen, wie etwa die Verwendung von Monitoren, um die aktuellen Arbeitsaufträge von verschiedenen Arbeitsplätzen anzuzeigen. Mathematische Gleichungen werden benutzt, um anfängliche obere und untere Grenzen für die benötigten Kapazitätsniveaus zu finden. Danach wird eine Simulation für verschiedene Kapazitätsniveaus innerhalb des Bereichs der theoretischen Ergebnisse durchgeführt, um die genau benötigte Mannzeit für verschiedene Jobsequenzierungsstrategien zu finden. Mit Hilfe der Statistiksoftware R wird ein Tool erstellt, welches Unternehmen für verschiedene Parameter Ergebnisse liefert. Diese Ergebnisse zeigen die Auswirkungen der Verwendung von Batching, unter Berücksichtigung der Begrenzung des Zeilenseitenraums und der Reduzierung der Leerfahrtstrategie für Leistungsmessungen. Die Strategie, das Leerfahren zu reduzieren, indem nach dem nächsten Arbeitsplatz gesucht wird, der einen Auftrag benötigt, ist nicht so effizient, da es die benötigte Kapazität erhöht. Dies liegt daran, dass es die Variabilität der Wartezeit vergrößert und somit den Prozentsatz der verspäteten Bestellungen steigert.

KEYWORDS

Forklifts, material handling, simulation, queueing theory

Gabelstapler, Fördertechnik, Simulation, Warteschlangentheorie

* Deggendorf Institute of Technology

1. Introduction

Material handling is one of the most important factors affecting costs and performance in manufacturing companies [1]. Any delay in moving the needed materials can cause production delay that, in turn, affects productivity. Moreover, the space for line-side inventory can be very scarce and insufficient for much material to be stacked. On the other hand, increasing the needed capacity can lead to major costs and an unorganized allocation of workload. Therefore, material handling is one of the most important activities that need to be optimized. Capacity in this paper means the total needed man hours for forklifts, which affect the maximum number of needed forklifts. Using forklifts is the traditional way in most material handling activities. *Loading units* and *pallets* will be used interchangeably in this paper. In small and medium-sized companies, it is common to see that forklift activities and the needed capacity are manually planned. This paper presents a tool for optimizing a forklift system assuming the availability of data about job orders. The number of drivers needed throughout the day and different performance measures can be obtained by using analytical equations and simulation. Different possibilities of job sequencing by using different priority rules can be done. We present the results of a case study for a company operating in Germany. The tool has been developed at Technology Campus Grafenau (TCG), which is an applied research center affiliated to Deggendorf Institute of Technology. This tool is part of efforts in a project called *Industrie 4.0 Werkstatt Bayerischer Wald* (Industry 4.0 Workshop Bavarian Forest) which is funded by the EU program EFRE (*Europäischer Fonds für Regionale Entwicklung*; European Fund for Regional Development). The tool performs with R Software, a free software that can be used by companies. More information about this software can be found in an article by Ihaka & Gentleman [2].

After this brief introduction about the importance of optimizing material handling systems and the tool developed, a literature review about using queueing theory and simulation in logistics and forklifts systems is presented. Moreover, some previous studies that investigated characteristics of forklifts systems such as batching and time-dependent arrival rates are discussed. After that, a problem definition and the objectives of the study are presented. The methodology, which depends on analytical equations and simulation, is explained. Analytical equations define boundaries of the needed capacity of drivers. Subsequently, simulation of forklift operation with different options and priority rules is done. Finally, results and conclusion are presented.

2. Literature review

2.1. Forklift system as a queueing system

The range of applications of queueing theory includes telecommunications and computer science, manufacturing, air traffic control, military logistics, design of theme parks, and many other areas that involve service systems whose demands are random [3]. Forklift system can be considered as a queueing system. Driving time of forklifts plus loading and unloading activities can be considered as service time. Forklifts and drivers are considered as servers. Pallets are the customers. More details about queueing theory can be found in Shortle et al. [4]. Little was published about using queueing theory to analyze forklift system. An example of applying queueing theory in logistics is the study by Masek et al. [5], where queueing theory was applied in warehouse optimization. Another paper by Munisamy [6] presents a port planning and operations model for capacity planning at a timber terminal. The analytical model was the closed queueing network model. The model evaluated the efficiency of the timber terminal in relation to the cargo handling system and its impact on terminal throughput capacity.

Simulation was used in logistics activities including forklifts. For example, simulation was used in a study by Takakuwa et al. [7], where a procedure to build simulation models for move-store activities of complicated and non-automated distribution warehouses was proposed. Vonolfen et al. [8] developed an optimization model for assembly line supply by integrating warehouse picking and forklift routing using simulation. Their research was motivated by an industrial application where multiple production lines were supplied from various warehouses at the production plant. Material was stored in high rack storage areas and picked by employees utilizing forklifts. Tarczyński [9] analyzed the impact of storage parameters and the size of orders on the choice of the method for routing order picking where the forklift truck's manoeuvers such as turns back and turns aside were considered.

2.2. Characteristics of the forklift queueing system

When there are several jobs served together, the system is called *bulk* service queueing model [10]. An example is the study by Parimala & Senniappan [3] who investigated the queueing model with single server. The service starts only when *batches* of 'a' customers are present. When the queue length is 'a', but less than or equal to 'b', then the entire queue is taken up for service. If there are more than 'b' customers in the queue, then the server accepts the first 'b' customers. Sometimes the arrivals are coming in batches. In a study by Chakravarthy et al. [11], they consider a multiple server queueing model in which customers arrive according to a batch Markovian arrival process (BMAP). Moreover, Haridass & Arumuganathan [12] concentrated on a bulk arrival and bulk service queueing system with variant threshold policies for vacations.

In this paper, the service time and the inter-arrival times have been assumed to be exponentially distributed, where multiple servers have been assumed. Some studies considered the M/M (a, b)/c queueing system such as Medhi [13], Neuts & Nadarajan [14], and Sim & Templeton [15]. See more details about Kendall's notation in Kendall [16]. Medhi [13] obtained an analytical expression for the customer waiting time distribution in terms of the steady-state probability P(0, 0) that there are no idle servers and no waiting customers. Neuts & Nadarajan [14] estimated the joint distribution of customer waiting time and service group size, and also the distributions of the number of waiting customers and the number of free servers. Sim & Templeton [15] derived an exact recursive algorithm to study the steady-state probability that there are m idle servers and j waiting customers in the queue. They also derived analytical expressions for steady-state system performance measures, in terms of P(0, 0), which can be computed using a recursive algorithm. Adan et al. [17] relaxed the assumption of exponential service time. In the previously mentioned studies, each customer takes the same size from the server's capacity. In this paper, however, the batch size depends on the size of pallets (customers) where different pallets have different sizes. An analytical expression is obtained for the number of servers (forklifts) needed for guaranteeing a certain limit of late percentage. This is only part of the methodology, which depends on analytical investigation and simulation.

The arrival rate of job orders is not only stochastic but also with *time-dependent rates*. Sometimes the jobs arrival rate in the middle of the day is much higher than the rate of jobs arrival in the evening. Therefore, the needed workers capacity must vary from time to time. The study by Massey [18], who analyzed queues with time-varying rates for telecommunication models, is an example of a study employing a time-varying rate. Queueing theory as discussed in that paper was organized and presented from a communications perspective. More details and a survey about time-dependent rates can be found in a study by Schwarz et al. [19], where there are several approaches for the performance analysis of queueing systems with deterministic parameter changes over time.

In queueing theory, it is common to use the First Come First Served (FCFS) rule for sequencing the waiting job orders. However, other *service disciplines* are found in practice. Iyer & Jain [20] modeled the impact of merging capacity in production-inventory systems where there are two warehouses. They presented a priority system in which orders from the lower-variability warehouse are given non-preemptive priority. They prove conditions under which the FCFS operating rule failed to achieve a Pareto improvement over the separate systems because it increases inventory cost at the lower-variability warehouse. Moreover, Silva & Serra [21] studied locating emergency services with different priorities. In urban medical emergency services, calls that involve danger to human life deserve higher priority over calls for more routine incidents. The performance of an emergency center may be judged by the number of persons in queue or by the length of time that a person must wait. These indicators are strongly correlated with the number of centers available and with their locations. In this paper, some job orders have higher priority over others based on some priority rules such as decreasing empty driving.

For a queue to be in steady state, the possibility of service rate must be higher than the arrival rate. However, in real-life peak periods, the number of arriving job orders is sometimes higher than the capacity of forklifts. Therefore, a queueing system with *blocking* is used for investigating this case. This means that the waiting line must not be longer than a certain limit. An example of blocking in queueing theory is the study by Roy et al. [22], in which an autonomous vehicle-based storage and retrieval system (AVS/RS) was investigated. In that study, protocols were developed to address vehicle blocking, and a semi-open queueing network model was proposed to analyze system performance and evaluate design trade-offs. Another example is the study by Koizumi et al. [23]. In their study, "blocking" denotes situations where patients are turned away from accommodations to which they are referred, and are thus forced to remain in their present facilities until space becomes available. Both mathematical and simulation results were presented and compared.

Numerous studies used very simple assumptions to simplify the system, hence it can be very difficult to represent the real situation on the ground. The majority of queueing theory studies assumes very simple situations that cannot be suitable for the forklift-planning problem. The mathematical derivation in this paper has special characteristics due to grouping pallets with different sizes. Little was published about employing queueing theory to study forklifts systems, especially when it comes to using IT infrastructure to send job orders, not using traditional kanbans, but electronic signals. Moreover, the system is not a normal *pull* system with electronic kanbans, but there is the possibility to use priority rules for sequencing job orders according to different criteria. This study combines queueing theory mathematical equations with simulation using R Software with very wide possibilities of combinations of parameters and options to be a tool for optimizing the forklifts systems in terms of capacity planning and job sequencing rules.

3. Problem statement

Usually, forklift system planning in terms of capacity planning and sequencing rules of job orders is performed manually. Especially small and medium-sized companies do so, even if there is an IT infrastructure immediately providing data about job orders. However, an optimized forklifts system may decrease work force costs, waiting times, line-side inventory costs, and congestions problems. The forklifts system contains the following complications, hence it is very difficult to be analyzed without simulation:

- Different demand rates from time to time
- Different used labor capacities during the day
- Working pauses
- Batching (bulking)
- Different sizes of load units
- Different possibilities of batching according to the area in which jobs are waiting
- Blocking in some peak periods
- Different priority rules used

Using simulation, however, can be time consuming and expensive for small and medium-sized companies. Sometimes, companies have no license for expensive simulation software, yet they will need to use simulation in every shift to estimate the needed capacities. Besides the fact that R Software is free, another advantage lies in its capabilities to calculate the results about capacity upper and lower limits from the mathematical equations and then using the results as input in simulation.

A simple way is to upload the Excel file containing the needed data prepared in a certain way using the internet browser. The model exists in the cloud and companies get their results immediately without seeing the code. One of the tool's most important advantages is that it was developed to be very flexible for different layouts and parameters.

By using this tool, the following objectives can be achieved:

- Minimize total needed capacity (workforce and maximum number of forklifts)
- Minimize empty driving (lower transportation costs)
- Better utilization (driving with the highest possible forklift capacity by grouping small pallets into one route)
- Enhancing other performance measures such as delay delivery, average waiting time per order, average number of orders waiting in the queue, etc.

4. Methodology

The methodology can be divided into three major steps:

1. First theoretical estimation of the upper and lower limits of the possibly needed number of drivers by using analytical mathematical equations.
2. Simulation

3. Final estimation of the needed capacity of forklift drivers using simulation and trying all the values of capacity between upper and lower limits found in the first step, and taking into consideration the different options selected by the user such as the assumed priority rules.

To handle the problem of different drivers' capacity during the day, the following two points are considered:

- Working time was divided by small time periods (hours). Each hour has the same arrival rate of job orders and the same workforce capacity.
- The system performance during the day will almost be the same. If arrival rate is increased, workforce capacity is also increased to keep almost the same performance. Small deviations in performance are ignored. In this case, the effect of time-dependent rates can be alleviated. In some peak periods, however, the needed demand can be higher than the forklifts available. In this case, more workforce capacity must be used directly after peak periods.

The two boundaries of the needed capacity (number of drivers and forklifts) found by equations are as follows:

- The upper limit assumes no batching, the same size of loading unit, and no breaks.
- The lower limit assumes full possibilities of batching, all waiting batches to be in the same area, loading time to be independent from batch size, and pallets that can be grouped together to have the highest priority to be loaded first. That means the FCFS rule is not applied.

4.1. Case study and study assumptions

This study was motivated by a case study where a company based in Germany uses forklifts to move materials among 12 departments distributed on four floors. Lifts are used to move forklifts among different departments. Each department has its own inventory locations. However, the tool developed in this study is flexible to handle several different layouts without the need for any adjustments. The company uses six forklifts. Demand is different from one hour to another and from one day to another. Therefore, a flexible capacity plan is necessary. Some drivers perform only driving and loading activities. Others are working on other activities besides using forklifts. Job orders are sent to a central system, which is connected to monitors fixed on forklifts displaying the needed jobs and their origins and destinations. Three different sizes of pallets are available: small, medium, and large. A forklift can load a maximum of 4 small pallets, 2 medium pallets, or only one large pallet. It is also possible, for instance, to load one medium pallet and two small pallets together. Each driver usually clicks on the monitor to choose the next job he wants to work on. His choice is based on his experience for the best route. Whenever he finishes a job order, he clicks on the monitor to save the time point of ending the job. Sometimes individual job orders are overlooked for a very long time. The task was to estimate the needed capacity during the day so that job lead time is not longer than 2 hours. Five percent of job orders exceeded this time limit. Another objective was to check the effect of using different strategies for prioritizing the different job orders and the effect on performance measures. Besides this forklifts system, there is another system using small trucks for small bins. For example, a truck can load up to 25 small bins at the same time. This second system was not considered in this study.

The main assumptions in the tool are:

- Forklift drivers are informed about the currently needed job orders.
- All forklifts have the same capacity.
- Workers can do driving, or can also perform other tasks such as order picking.
- One forklift can load one or several loading units if there is enough capacity and if all the pallets are on the same route.
- If the planned working time of a driver finishes while he is still working on an order, he finishes the order before he leaves.
- The forklift can stay at a central station if there are no new orders, or just wait at the last destination.

4.2. Upper and lower limits of workforce needs

Assuming that forklifts can carry only one pallet with exponentially distributed time of loading, unloading, driving, and empty driving, and assuming that the inter-arrival time of job orders is also exponentially distributed, then we have the queueing system M/M/C. In this case, the probability that waiting time in queue (T_q) for more than a certain time t can be found using the following equation (cf. Shortle et al., [4]):

$$Pr\{T_q > t\} = \frac{r^c P0}{c!\,(1-\rho)} e^{-(c\mu-\lambda)t}$$

(1)

Where,

- λ is the number of job orders within a time unit (e.g. an hour)
- μ is the number of job orders that can be fulfilled by one forklift within the time unit
- c is the number of available forklifts/ drivers
- $r = \lambda/\mu$
- $\rho = \lambda/\mu c$
- P0 is the probability that there are 0 customers (job orders) in the system and can be found by:

$$P0 = 1/\left[\sum_{m=0}^{c-1} \frac{(c\rho)^m}{m!} + \frac{(c\rho)^c}{c!\,(1-\rho)}\right]$$

(2)

If the objective is that the waiting time of an order is not to exceed 2 hours minus average service time (e.g. 5 minutes) with a probability of 95%, then t is set to be 1.92 and $\Pr\{T_q>t\}$ is set to be 0.05. Then c, which represents the number of needed forklifts drivers, is simply found by trying all the possible values for c in equation (1). The lowest c value that yields 0.05 or less for the left hand side is used. The result is then rounded up. Note that variability in service time, which is usually small, has been assumed to be ignored if the objective is to concentrate on lead time. To try all the possible values of c, upper and lower limits are needed. The value $\lceil r \rceil$, which equals $\lceil \lambda/\mu \rceil$, can be used to find the lower limit where $\lceil x \rceil$ is the upper bound of x. This simple initial solution ignores several characteristics that can increase or decrease the needed capacity. For example, the following points decrease the needed capacity:

- Batching of job orders
- Production pause

On the other hand, the following points can increase the needed capacity:

- Forklifts drivers pause
- Forklifts breakdown
- Maximum limitation on the numbers of available forklifts at the same time. The effect of this point is for the period after the peak period, where it is sometimes necessary to increase the workforce capacity after the peak period in order to compensate for not using the optimal theoretical value of c because of a limitation on the number of used forklifts in the peak period.
- The line-side inventory limit. Some areas have small space for inventory so they might have a higher priority than the other areas. That might increase the needed capacity.
- Rules other than FCFS

It is important to consider forklift drivers' pauses in the calculations of upper and lower limits. For example, suppose that all drivers are leaving for a 30-minute break at 11:30 o'clock, then if the theoretical needed capacity is three drivers, the actual number should be six because drivers are not working half the time during this hour. Moreover, if it is expected that the needed capacity is even larger than the calculated upper limit – due to limitation on line-side space – then a higher level of capacity can also be tried by simulation. This option was activated by the tool.

One of the main features of the forklift system is batching more than one pallet on the forklift. Sometimes these pallets come in different sizes. In this paper, the pallet size is measured based on its percentage to the whole capacity of forklifts. For example, a certain pallet can take 50% of the whole loading capacity of a forklift. Sometimes, a pallet is not big, however, it is unstackable. In this case, only one pallet is loaded, and therefore the size of this pallet is considered to be 100%. A complicating feature is that there are different sizes of different pallets in the same facility. A parameter that is very important in the analysis of the system is the average pallet size, which can be found using equation (3):

$$Average\ pallet\ size = \sum_{i=1}^{X} s_i f_i \le 1$$

(3)

Where,

- X: number of different sizes of pallets
- s_i: size of pallet i (as percentage of loading capacity of forklifts)
- f_i: percentage of demand on pallet i. It can be found by dividing the daily demand on i pallets by the daily demand on all pallets.

Batching should reduce the needed capacity. Assuming for simplicity that the time needed for a batch exactly corresponds to the time needed for only one pallet, the system has the same service time µ, however, with a different arrival rate, which is λ' indicating the arrival of batches. Service time here means driving and loading/unloading times. In λ', any arrival of jobs will be ignored until a complete batch is available or when a server (forklift) is available to start a new job. This system is similar to the lift system in which serving one or a group of people will take almost the same amount of time. The lift starts moving when there is only one person or a group of people without exceeding the capacity of the lift. Later a range for λ' is estimated.

Because pallets come in different sizes, forklifts sometimes drive without full capacity. For example, a forklift can leave with two pallets of which one is 50% and the other one 25% of the capacity of forklift. Therefore, the total loaded material corresponds to 75% of the capacity of forklift. It is important to consider the average percentage of loading material without full capacity (P_l) (underutilization percentage). For simplicity, assume that the average size of unutilized capacity is $0.5P_l$. If, for example, all drives of forklifts are with some unused capacity then, on average 50% of forklifts capacity will not be utilized. If, however, sizes of pallets are for example 50%, 75%, and 100%, then it is possible to consider the 75% size as the 100% size, because it is impossible to combine the 75% pallet with the 50% pallet.

Assuming for simplicity that batching will not affect the service time and assuming that whenever there is any number of jobs in the queue that do not exceed the capacity of forklift they will be batched together, then in order to find the number of batches in the queue it is possible to write

$$L_{q\_batch} = 0.5P_l + L_q{}' \sum_{i=1}^{X} s_i f_i$$

(4)

Where L_{q_batch} is the number of batches in the queue, L_q' is the number of jobs in the queue, and P_l is the probability of driving without utilizing the full loading capacity of forklifts. This equation shows the relationship between the number of job orders (pallets) in the queue and the number of batches in the queue. If the driver finds pallets that need part of forklift capacity then he will take the currently available pallets. Therefore, the system behaves as if there was an upper bounding effect for the number of batches. The value $0.5P_l$ was added to compensate for this upper bounding. For example, if the waiting orders in the queue amount to 3 pallets and they form 1.5 of the capacity of the forklift (average pallet size is 50% of forklift loading capacity), it is assumed that there are 2 batches. Moreover, if the number of batches in the queue is always greater than one batch, then $P_l = 0$, which means that forklift loading capacity is fully utilized.

Number of arriving batches within a unit of time (λ')

The same previous method for finding the average number of waiting batches in the queue can be applied for λ':

$$\lambda' = 0.5P_l + \lambda \sum_{i=1}^{X} s_i f_i$$

(5)

For example, if λ = 2 and average pallet size = 30%, and if the number of batches in the queue is always greater than 1, then λ' = 0.6. This means that the forklift will not only take the two pallets, but will also add the rest of loading capacity from the waiting pallets in the queue.

Because $0<P_l<1$, then

$$\lambda \sum_{i=1}^{X} s_i f_i < \lambda' < 0.5 + \lambda \sum_{i=1}^{X} s_i f_i$$

(6)

Since we want to find the lower limit of the number of drivers, the lower limit of λ' is considered. Equation (1) can be rewritten to find the needed c value (number of forklifts drivers):

$$Pr\{T_q > t\} = \frac{(\lambda'/\mu)^c P0'}{c!\,(1-(\lambda'/\mu c))} e^{-(c\mu-\lambda')t}$$

(7)

where

$$P0' = \left(\frac{(\lambda'/\mu)^c}{c!\,(1-(\lambda'/\mu c))} + \sum_{n=0}^{c-1} \frac{(\lambda'/\mu)^n}{n!}\right)^{-1}$$

(8)

The effect of omitting $0.5P_l$ is not great. Using simulation on data in which λ' is from 1 to 60 and μ is from 1 to 30 it was found that the biggest difference that can arise is only an increase of one server. Furthermore, it was found that an increase of 0.5 on λ' can make a difference of less than 6% of possible values of λ' if, for example, μ = 8. The higher the value of μ, the lower this percentage. It is unlikely for the needed capacity to equal the result in equation (7) because of the following reasons:

- Loading a batch will take more time than loading only one pallet.
- In the case of several areas with waiting pallets, it is almost impossible to perform batching every time because the waiting pallets are going to different places and are following different routes. The higher the number of inventory areas and the more complex the layout, the higher the chance of not utilizing batching at its maximum possibility.

Thus, equation (1) and equation (7) provide initial estimations for the upper and lower limits of the number of needed workers. Simulation can be used to find the results for all scenarios between these two limits. For example, assume that service time is exponentially distributed with an average of 12 minutes, and inter-arrival time is exponentially distributed with an average of 5 minutes, then 4 drivers are needed in the case of no batching at all,. Moreover, assuming that batching is possible where all the waiting job orders are in the same area and time for loading a batch is exactly as for loading a pallet, then two drivers are needed. In this case, simulation is required to try scenarios in which the c value is 2, 3, and 4.

4.3. Priority rules for different orders (jobs)

A driver starts to serve a certain job order based on the following rules: (Rules are sorted starting from the highest priority to the lowest one)

1. A job order waiting for longer than a threshold time value.
2. Job orders coming from departments with small areas for line-side inventory.
3. Bulking: If there are several waiting pallets and they can be grouped, they have the priority where:
    1. there is the possibility of loading bigger loading units at first.
    2. the time needed for loading and delivering the whole group is affected by *bulking effect*. For example, if it amounts to 10 %, then grouping two pallets will make the average service time to equal the average time multiplied by 110%.
4. Reducing empty driving time: whenever possible, the forklift should start with the closest place to the previous forklift destination.
5. FCFS

The decision-maker can deactivate any of rules 1 to 4 in the tool.

Line-side inventory

Factors affecting the response time for a job order, i.e. available man hour capacity, available forklifts, forklift loading capacity, empty transportation time, other job orders, and traffic jam of forklifts may increase line-side inventory.

The workplace priority index is used. It can be dynamic and may change from one hour to another. Therefore, it is possible to write:

$$PI_{ph} = \frac{d_h O_{ph} \sum_{i=1}^{X} s_i f_i}{LSIC_p}$$

(9)

Where:

- PI_{ph}: priority index for workplace *p* in hour *h*
- *dh*: average hourly demand in hour *h*. In other words, the average arrival rate of job orders in hour *h*.
- O_{ph}: percentage of job orders for workplace *p* in hour *h*
- $LSIC_p$: capacity of line-side space for workplace *p* measured in units of capacity of forklifts. For example, if the maximum area capacity is two pallets of a certain size and forklift can load these two pallets together, and it cannot take more than them, then $LSIC_p = 1$.

In the case of the percentage of demand for different workplaces being stable over the day, O_p is used instead of O_{ph}. In this case, the priority index for each workplace will be the same over the day. The term "$d_h\, O_{ph}$" represents the workplace's hourly demand in terms of pallets. If it is available, d_{ph} can be used instead. The nominator represents the hourly demand or arrival rates for the workplace *p* in hour *h* in units of forklifts. If the index equals one, then, on average, the line-side area will be filled with pallets within one hour if there are not any forklifts to take these pallets during that hour. Whenever the priority index for a certain workplace is more than for example 0.5, any order coming from that workplace must have a higher priority.

4.4. Simulation with R

Apart from simulation purposes, R is also used to perform the theoretical calculations about the initial limits of capacity needs. It is easy to find these limits for the *c* value using R, where all the possible values are tried until the appropriate value is found. Therefore, R was used to perform the whole tool. Performing simulation with R renders the model independent from the facility layout as in traditional simulation software. That saves a considerable amount of time of simulation development. The data must be entered into an Excel file in a predefined way and uploaded via a specific link on the internet, and then the user can obtain the results.

The package "queueing" can be used to get exact results for some types of queues. Moreover, the package "queuecomputer" may be used to perform simulation of some basic types of queues. There is the possibility to set variable server capacity to be used by the package. Because forklift systems have so many unique characteristics that make it impossible to directly use such a package to simulate the system, a new code has been written to be suitable for such a system. Some parts of the package code may also be used. The starting capacity of servers (drivers) can be set employing the two limits found in mathematical equations. Because the arrival rate of job orders is different over the day, the model must take into consideration changeable workforce capacity over the day. For the waiting job orders, the priority rules in section 4.3 are considered, and performance measures are calculated. All the parameters and data can be set by the user him/herself. The user can for example employ the tool for simulation instead of optimization. In this case, he/she must determine a capacity plan. Any one of the priority rules can be activated or deactivated. Moreover, bigger pallets can be loaded at first in the case of batching.

It is important to know that for any system to be in steady state, ρ that equals $\lambda/\mu c$ must be less than one. Therefore, if λ/μ = integer, then *c* must be at least $\lambda/\mu+1$. In this case, the performance of the system is usually good and most of the jobs wait for a short time only. Figure 1 shows the probability that waiting in the queue takes longer than a certain time *t* if $\lambda = \mu$ and if $c = 2$. Figure 2 shows the case when λ/μ is an integer which can be greater than 1, and waiting in queue is 2 hours. The two figures show the fact that when λ/μ is an integer and *c* is set in a way to turn the system into steady state, then waiting time in the queue is not very long. For example if $\lambda=\mu=1$ and $2<=c<=$any value greater than 2, and if $t = 2$ hours, then there will be no way that $\Pr\{T_q>t\}$ is >0.05. Therefore, $c=\lambda/\mu+1$ can immediately be chosen.

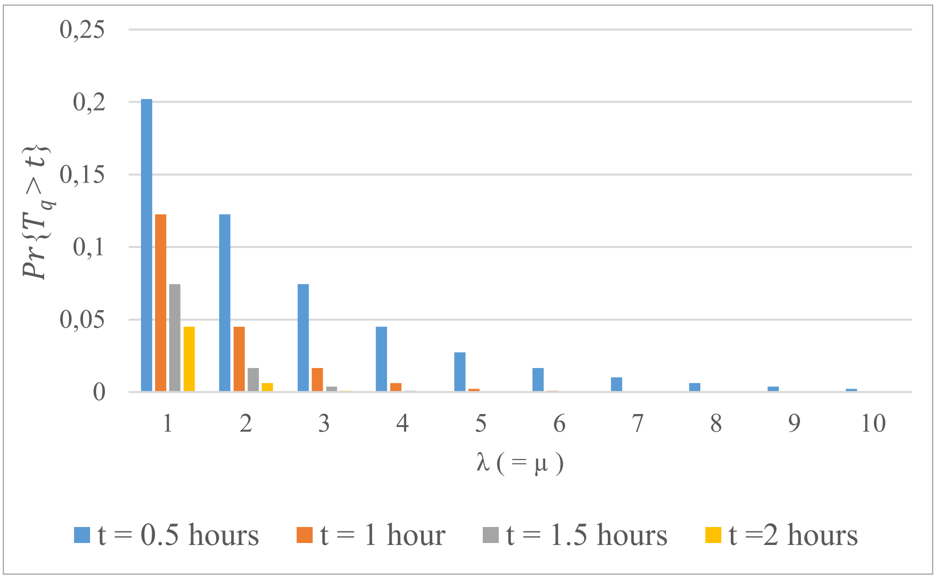

Figure 1: Probability that waiting in the queue is more than *t* hours when c=2

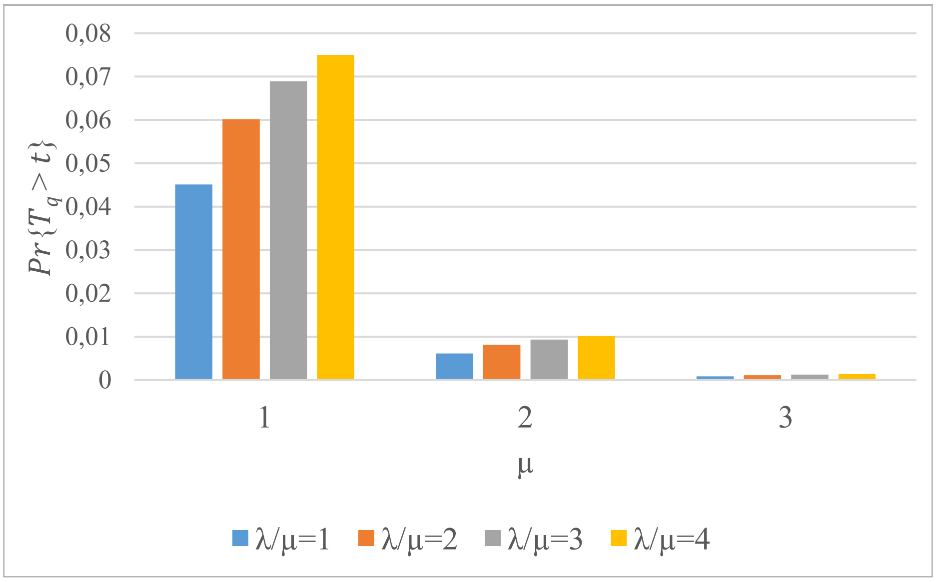

Figure 2: Probability that waiting in the queue is more than *t*=2 hours when c= λ/μ+1

Because the size of capacity is different during the day, a possible solution to go through the range between the two limits found by mathematical equations is to start from the highest one and then reduce the capacity for example 5% in each iteration until getting the lowest possible one. Because capacity must be integers, upper bounding of the needed capacity must be done. Because of this upper bounding, the system performance will not be the same over the day. There must be some differences from time to time.

For the case study, data about driving time and the percentage of different loading packages was drawn from historical data of more than one year. The arriving rates, however, are different from one day to another. Service time was sometimes found to be exponentially distributed, and sometimes it follows different types of distributions. The theoretical equations were found based on the assumption of exponential distribution. However, the simulation itself for values within the range found by theoretical equations provides the decision-maker with the possibility to define the appropriate distribution such as exponential, uniform, normal, etc. for each different driving time. For example, in some places where there is a lot of traffic, the variability in driving time can be high. However, the variability in other places where there is not much traffic can be low. For instance, driving time can be exponentially distributed from workstation 1 to workstation 2, and normally distributed from workstation 2 to workstation 3. There is transportation with forklifts within the same workstation, especially when it is relatively large. For example, figure 3 shows the histogram for the full load driving time within workstation 5. KolmogorovSmirnov test was also done and the p-value was found to be 0.112 when testing the exponential distribution hypothesis. That means that the available historical data about this driving time show that it follows exponential distribution.

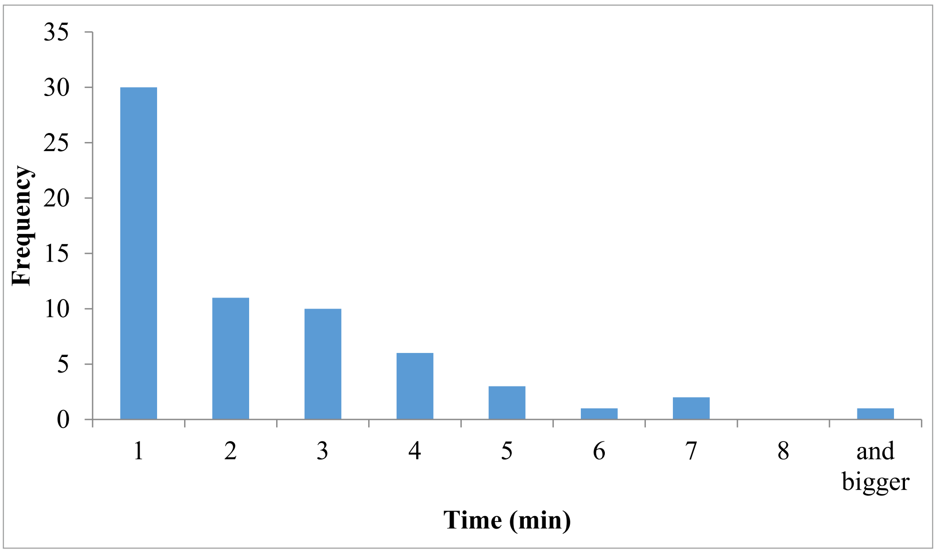

Figure 3: Histogram for full load driving within workstation 5

5. Results and analysis

In this section, input data are presented and then the results and their analysis are shown. The pallets data was as in table 1. Table 2 displays arrival rates of job orders. For the 12 workplaces, demand percentages were variable and can be 0, 5%, 10%, 20%, 30%, 40%, and 50% from the total hourly demand. A matrix of driving times from each workspace to another based on their origins and destinations was used. Another matrix was used for empty driving time. The empty driving time showed a slightly lower average. Driving time here also includes the loading and unloading time of pallets.

Pallet types	Percentage of demand	Size (compared to forklift loading capacity)
Small	36.4%	25 %
Medium	57.3%	50 %
Large	6.3%	100 %

Table 1: Pallets types in case study

Time	from	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
	to	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Inter-arrival time (min)		4.3	2.7	4	2.7	2.7	2	1.6	1	1.3	5	12	15	15	15	15	15	15

Table 2: Job arrival rate

The following scenarios were tried out by the tool:

Scenario 1: All priority rules are activated.
Scenario 2: Only batching was deactivated.
Scenario 3: Only reducing empty driving was deactivated.
Scenario 4: Five places have capacity for line-side inventory of 4 forklifts capacities.

Batching effect on time was set to be 10%. The number of iterations was set to be 20 and no break was considered. Maximum allowable lead time was set to be 60 min. Only 5% of job orders were allowed to be late (take more than 60 minutes). No limitation was selected on the maximum number of used forklifts in any hour. Moreover, work ends in the middle of the night.

For the case study containing 12 workstations, the calculation time using R Software took about 1 minute on a normal personal computer. Results about order fulfillment time (lead time) are shown in figure 4. Working hours mean the starting of an hour. For example, the working hour 6:00 means the time period 6:00-7:00 h. The average values of order fulfillment time can be found in table 3. The average values are almost the same for scenario 1 and scenario 2. As expected, the performance of the system is not stable over the day, even if the number of drivers is variable to alleviate the changes in demand. This is due to the aforementioned upper bounding effect for the capacity determined.

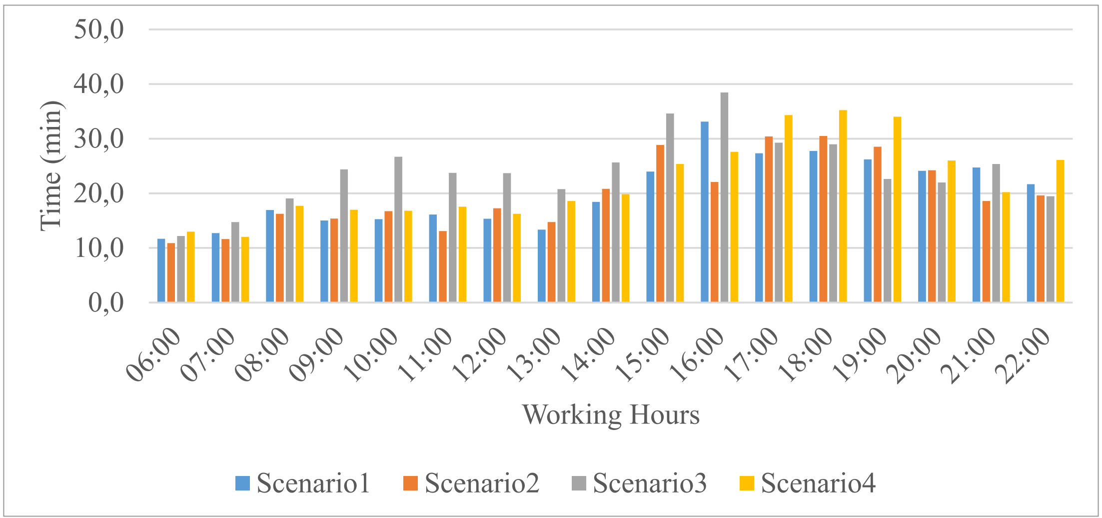

Figure 4: Average hourly order fulfillment time over the day for different scenarios

Figure 5: Optimal number of forklift drivers versus theoretical upper and lower limits

Figure 5 shows capacity variable levels. The optimal level of capacity is the lowest level of capacity that does not cause a lead time to be longer than the maximum allowed one with a probability of 95%. That means only 5% of orders can take an order fulfilment time which is longer than the maximum allowed one. It is clear that in all the different scenarios, the capacity level is within the expected theoretical range. The level of capacity in scenario 2 was the highest. An interesting point in table 3 is that the number of needed forklift drivers was found to be the lowest in scenario 3. To explain that, table 4 shows a comparison between the optimal iteration in scenario 1 and a previous iteration in which the man hour is 41 for scenario 3. One new performance measure was added, which is the standard deviation of waiting time. It is higher for scenario 1, when the option for reducing empty driving was activated. This is due to the simulation model trying to find the closest workstation for the forklift on the expense of the job orders waiting for longer times, but being for farther workstations. Therefore, the late orders percentage is higher in scenario 1. Thus, the decision-maker may want to deactivate this option to get the lowest man hour and that may reduce the maximum needed forklifts. This is especially useful when there are maintenance tasks for forklifts.

Performance Measures	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Percentage of late orders (%)	1.92	2.97	2.97	3.14
Average order fulfillment time (min)	17.32	17.07	23.04	19.14
Mean waiting time in queue (min)	13.12	13.03	18.68	14.99
Man hour	41	44	35	41
Utilization (%)	83.65	86.9	86.86	84.59
Total empty driving time (min)	951.17	1051.45	873.8	988.63
Average number of orders in the queue	4	3.99	5.71	4.54
Maximum number of forklifts	6	7	5	6

Table 3: performance measures for different scenarios

Performance Measures	Scenario 1 Optimal Iteration	Scenario 3 Previous Iteration
Percentage of late orders (%)	1.92	0.33
Average order fulfillment time (min)	17.32	16.96
Mean waiting time in queue (min)	13.12	12.75
Standard deviation of waiting time (min)	12.28	9.66
Man hour	41	41
Utilization (%)	83.65	83.64
Total empty driving time (min)	951.17	1054.35
Average number of orders in the queue	4	3.85
Maximum number of forklifts	6	6

Table 4: Performance measures comparison for scenario 1 and scenario 3 for the same man hour value

Conclusion

In this paper, a forklift material handling system was analyzed to find the optimal capacity levels over the day and the best priority rule for sequencing the job orders. The analysis depends on performance measures such as man hour, waiting time, lead time, and percentage of delayed orders. Queueing theory and simulation were used to find the best parameters of the system. Even though the study considered a case study for obtaining results, its methodology can be applied to other environments in which forklifts are used. It has been assumed in this study that workers work on driving, or might also do driving from time to time as well as other tasks such as order picking. Another limitation lies in the theoretical mathematical equations where service times and inter-arrival times are assumed to be exponentially distributed. However, the tool provides the decision-maker with the possibility to choose other types of distributions in simulation, such as normal, uniform, and triangular distributions. Results showed that batching orders has a significant effect on enhancing performance measures. Results also showed that trying to decrease the needed empty driving times might not be the optimal decision because it increases the variability in waiting times. It is therefore recommended for future research to find other priority rules that consider waiting times. That is possible when the IT infrastructure allows that by informing drivers about the optimal sequence of job orders via monitors attached to forklifts.

References

[1] Baudin, Michel (2005): Lean logistics. The nuts and bolts of delivering materials and goods. New York, NY: Productivity Press.

[2] Ihaka, Ross; Gentleman, Robert (1996): R: A Language for Data Analysis and Graphics. In: *Journal of Computational and Graphical Statistics* 5 (3), S. 299–314.

[3] Parimala, R. Sree; Senniappan, Palaniammal (2015): An analysis of bulk service queueing model with servers various. In: *International Journal of Advances in Science and Technology* 4 (2), S. 18–24.

[4] Gross, Donald; Shortle, John F.; Thompson, James M.; Harris, Carl M. (2013): Fundamentals of queueing theory. Hoboken, N.J: Wiley & Sons, Inc. (Wiley Series in Probability and Statistics).

[5] Masek, Jaroslav; Camaj, Juraj; Nedeliakova, Eva (2015): Application the Queuing Theory in the Warehouse Optimization. In: *International Journal of Social, Behavorial, Eduational, Economic, Business and Industrial Engineering* 9 (11), S. 3744–3748.

[6] Munisamy, Susila (2010): Timber terminal capacity planning through queuing theory. In: *Maritime Economics & Logistics* 12 (2), S. 147–161.

[7] Takakuwa, Soemon; Takizawa, Hiroki; Ito, Kumiko; Hiraoka, Shinichiro (2000): Simulation of distribution systems: simulation and analysis of non-automated distribution warehouses. In: Paul A. Fishwick (Hg.): Proceedings of the 32nd conference on Winter simulation (10-13 Dec, 2000; Orlando, FL, USA). San Diego, CA: Society for Computer Simulation International (ACM Digital Library), S. 1177–1184.

[8] Vonolfen, Stefan; Kofler, Monika; Beham, Andreas; Affenzeller, Michael; Achleitner, Werner (2012): Optimizing assembly line supply by integrating warehouse picking and forklift routing using simulation. In: Proceedings of the 2012 Winter Simulation Conference (WSC) [Dec 2012; Berlin], S. 1–12.

[9] Tarczyński, Grzegorz (2012): Analysis of the impact of storage parameters and the size of orders on the choice of the method for routing order picking. In: *Operations Research and Decisions* 22 (4), S. 105–120.

[10] Jaiswal, N. K. (1960): Letter to the Editor—Bulk-Service Queuing Problem. In: *Operations Research* 8 (1), S. 139–143.

[11] Chakravarthy, Srinivas; Dudin, Alexander N. (2002): A multi-server retrial queue with BMAP arrivals and group services. In: *Queueing Systems* 42 (1), S. 5–31.

[12] Haridass, M.; Arumuganathan, R. (2011): Analysis of a batch arrival general bulk service queueing system with variant threshold policy for secondary jobs. In: *International Journal of Mathematics in Operational Research* 3 (1), S. 56–77.

[13] Medhi, Jyotiprasad (1979): Further Results on Waiting Time Distribution in a Poisson Queue Under a General Bulk Service Rule. In: *Cahiers Centre Rech. Opnl.* 21 (2), S. 183–189.

[14] Neuts, Marcel F.; Nadarajan, R. (1982): A multi-server queue with thresholds for the acceptance of customers into service. In: *Operations Research* 30 (5), S. 948–960.

[15] Sim, S. H.; Templeton, J.G.C. (1985): Steady state results for the M/M(a, b)/c batch-service system. In: *European Journal of Operational Research* 21 (2), S. 260–267.

[16] Kendall, David G. (1953): Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. In: *The Annals of Mathematical Statistics* 24 (3), S. 338–354.

[17] Adan, Ivo J.B.F.; Resing, Jacques A.C. (2000): Multi-server Batch-service Systems. In: *Statistica Neerlandica* 54 (2), S. 202–220.

[18] Massey, William A. (2002): The Analysis of Queues with Time-Varying Rates for Telecommunication Models. In: *Telecommunication Systems* 21 (2-4), S. 173–204.

[19] Schwarz, Justus Arne; Selinka, Gregor; Stolletz, Raik (2016): Performance analysis of time-dependent queueing systems: survey and classification. In: *Omega* 63 (September), S. 170–189.

[20] Iyer, Ananth V.; Jain, Apurva (2004): Modeling the impact of merging capacity in production-inventory systems. In: *Managment Science* 50 (8), S. 1082–1094.

[21] Silva, F.; Serra, D. (2008): Locating emergency services with different priorities: the priority queuing covering location problem. In: *Journal of the Operational Research Society* 59 (9), S. 1229–1238.

[22] Roy, Debijt; Krishnamurthy, Ananth; Heragu Sunderesh S.; Malmborg, Charles J. (2014): Blocking effects in warehouse systems with autonomous vehicles. In: *IEEE Transactions on Automation Science and Engineering* 11 (2), S. 439–451.

[23] Koizumi, Naoru; Kuno, Eri; Smith, Tony E. (2005): Modeling patient flows using a queuing network with blocking. In: *Health Care Management Science* 8 (1), S. 49–60.

Dr. Eng. Mohammed Alnahhal

Mohammed Alnahhal holds Bachelor and Master degrees in Industrial Engineering. He was employed at the Islamic University of Gaza for three years. In 2015, he received his PhD degree from the University Duisburg-Essen for his thesis on efficient material flow in mixed-model assembly lines. For one year subsequent to his PhD graduation he was a logistics consultant in a German company called Müller – Die lila Logistik AG. He is currently working as a research fellow at Deggendorf Institute of Technology.

Dr. Mohammed Alnahhal hat einen Bachelor- und einen Masterabschluss in Wirtschaftsingenieurwesen absolviert. Er war drei Jahre lang an der Islamischen Universität Gaza tätig. Im Jahr 2015 promovierte er an der Universität Duisburg-Essen zum Thema effizienter Materialfluss in gemischten Modellmontagelinien. Im Anschluss an seine Promotion arbeitete er ein Jahr lang als Logistikberater in der Firma Müller – Die lila Logistik AG. Aktuell ist er als wissenschaftlicher Mitarbeiter an den Technischen Hochschule Deggendorf beschäftigt.

Contact / *Kontakt*
✉ **mohammed.alnahhal@th-deg.de**

Prof. Dr. Diane Ahrens

Diane Ahrens studied Business Management at the University of Passau. After receiving her doctorate, she acquired industry experience at the Central Global Supply Chain and Procurement Department of Siemens AG. She was appointed Professor of International Management at Hof University of Applied Sciences and Deggendorf Institute of Technology in 2003 and 2009, respectively. Apart from her teaching activity, she leads a 40-person research team specializing on supply chain management, data analytics and applications, as well as smart region projects.

Prof. Dr. Diane Ahrens studierte Betriebswirtschaftslehre an der Universität Passau, an der sie auch promovierte. Industrieerfahrung sammelte sie in der Zentralstelle Global Supply Chain and Procurement der Siemens AG. 2003 wurde sie an die Hochschule Hof, 2009 an die Technische Hochschule Deggendorf als Professorin für Internationales Management berufen. Neben ihrer Lehre leitet sie dort ein 40-köpfiges Forschungsteam, spezialisiert auf Supply Chain Management, Data Analytics and Applications sowie Smart Region.

Contact / *Kontakt*
✉ **diane.ahrens@th-deg.de**