Data driven supply allocation to individual customers considering forecast bias

International Journal of Production Economics

Seitz, Alexander; Grunow, Martin; Akkerman, Renzo

https://doi.org/10.1016/j.ijpe.2020.107683

Int. J. Production Economics 227 (2020) 107683

Contents lists available at ScienceDirect

International Journal of Production Economics

journal homepage: http://www.elsevier.com/locate/ijpe

Data driven supply allocation to individual customers considering forecast bias

Alexander Seitz [a,*], Martin Grunow [a], Renzo Akkerman [b]

[a] Production and Supply Chain Management, Technical University of Munich, Munich, Germany
[b] Operations Research and Logistics Group, Wageningen University, Wageningen, the Netherlands

ARTICLE INFO

Keywords:
Allocation planning
Order promising
Demand fulfilment
Demand forecast bias
Big data
Supply chain planning

ABSTRACT

We propose a data-driven allocation planning approach, which is designed for use in advanced planning systems as they are widely used in industrial environments. The approach exploits increasingly available data on individual customers and products by allocating supply on a highly granular level at high planning frequencies. It counteracts rationing gaming by customers, which we assume to be the reason for demand forecast biases.

We create an incentive for truthful forecasting by not only allocating supply based on customer profitability but also based on forecast bias. In the long term, this approach gives access to a profit potential and an on-time service level increase. In the short term, however, setting such an incentive does not only have a positive impact on service levels but also leads to a decline in profits. Our methodology quantifies this trade off providing decision support for determining the extent to which the forecast bias should affect the allocation.

In a numerical study based on the semiconductor industry, we demonstrate that the approach has a large long-term profit potential while having limited effect on short-term profits for significant service level incentives. The analysis further shows that the allocation efficiency increases with the granularity level and the predictive quality of the available data.

1. Introduction

Shortening economic and product life cycles lead to increasing demand variations, which are amplified through the supply chain by the so-called bullwhip effect. For example, the growth rates of the semiconductor market (without memory chips), whose companies are typically located upstream in the supply chain, varied between −40% and +50% since the beginning of 2009 (WSTS Inc., 2015). Consequently, periods of supply shortage occur more frequently. Hence, allocation planning (AP), i.e. deciding on when and how to fulfil which customer's demand, gains importance in industrial practice.

AP is part of a demand fulfilment process, typically implemented in software systems such as advanced planning systems (see Fig. 1). It reserves quantities of inventory and planned supply receipts, together termed available-to-promise (ATP), for certain customer segments. These supply reservations, called allocated available-to-promise (AATP), are then used to confirm delivery dates for incoming customer orders in a real-time order promising step. The communication with the customer is often fully automated and occurs at high frequencies, allowing the AP to also be performed at high frequencies.

With the recent advances in big data tools, companies are able to monitor the ordering behaviour of their customers on the granularity level of individual customers and final products. The higher transparency of the customers' ordering behaviour provides opportunities to increase the efficiency of supply allocation. However, conventional AP approaches based on segmentation are not able to exploit the available data because customers are typically clustered according to profitability, disregarding their ordering behaviour.

From the literature it is known that in order to satisfy their demands, the customers strategically inflate their forecasts in supply shortage situations to game the AP procedure of their supplier. Intuitively, one would assume that customers inflate their orders by a fixed bias. However, the data from a large European semiconductor manufacturer, which we use in our case study, shows that the consumers usually communicate a stable forecast, but order only smaller volumes at irregular intervals. Fig. 2 shows the demand errors and the forecast bias resulting from typical customer forecasting and ordering behaviour. The error is defined as the part of the forecast that is not translated into

* Corresponding author.Arcisstraße 21, 80333, Munich, Germany.
E-mail addresses: alexandermseitz@gmail.com (A. Seitz), martin.grunow@tum.de (M. Grunow), renzo.akkerman@tum.de, renzo.akkerman@wur.nl (R. Akkerman).

https://doi.org/10.1016/j.ijpe.2020.107683
Received 7 July 2017; Received in revised form 9 February 2020; Accepted 11 February 2020
Available online 14 February 2020

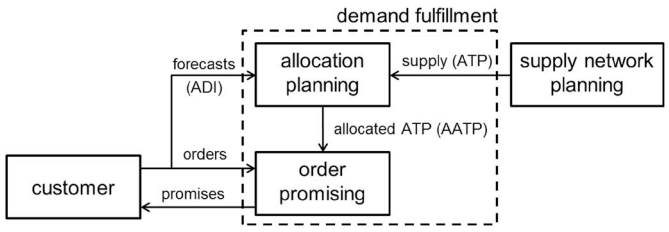

Fig. 1. General structure of a demand fulfilment process.

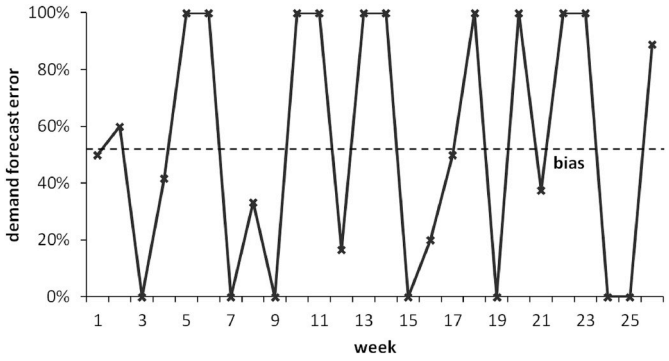

Fig. 2. Demand forecast bias graph for a typical customer demand pattern.

actual orders relative to the total forecast. Its structural component, i.e. its non-random part, is defined as the forecast bias, which we assume to be the result of the strategic order inflation behaviour. In Section 3.1, the exact definitions of the forecast error and the forecast bias are provided.

For a supplier, such strategic customer behaviour, commonly referred to as rationing gaming (see e.g. Lee et al., 2004), leads to the risk of inefficient supply allocation. Scarce supply is allocated to inflated demand of highly profitable customer segments, which is not consumed by subsequent orders. Despite the supply shortage, this results in high stock levels. At the same time, demand of customer segments with lower profitability is not fulfilled in full, leading to low overall on-time service levels.

In order to counteract the negative consequences of the rationing game, new approaches are needed to incentivise customers to truthfully forecast their demand. Herein, the results of big data tools monitoring historical demand forecast biases can be exploited to identify systematic gaming behaviour.

This paper develops a data-driven allocation planning (DDAP) approach that exploits increasingly available data on individual customers and products by allocating supply on a highly granular level at high planning frequencies. It allocates ATP supply to individual customers considering customer forecast bias in addition to customer profitability. In the long term, this AP approach provides an incentive for truthful forecast information sharing, leading to a more efficient product allocation process and unlocking additional profit potential and service level improvements. The incentive depends on the extent to which customers with different forecast biases are served with different service levels. In the short term, however, the introduction of such an incentive leads to a decline in profits compared to an allocation entirely based on customer profitability. In addition, the overall service level across all customers is affected. These long-term and short-term effects must be considered when deciding on the extent to which forecast bias is considered in the allocation. Our paper provides a thorough basis for the decision by developing a framework that structures the different analyses necessary to derive the information required to trade off the effects on long-term profit and service level potentials, incentive size, and short-term profits and service levels.

The approach is developed for the use in advanced planning systems in industrial environments in which standardised goods are mass produced and businesses use a make-to-forecast strategy in supply planning. Because production lead times are larger than order lead times, a notification-release order cycle is used, which first allocates available supply based on demand forecasts, i.e. notifications, provided by the customers and later releases this supply upon order reception.

Using the approach in a numerical study based on the semiconductor industry, we demonstrate that the proposed method has a significant long-term profit potential and on-time service level potential. In the short term, it also results in lower average stock levels and an increased overall service level, especially for customers with truthful forecasts. Finally, we show that allocation on individual customer level is only valuable when additional data such as forecast bias and lead time is exploited on this granularity level.

The remaining paper is organised as follows: Section 2 provides a literature review. Section 3 explains our AP approach in detail. Section 4 introduces a case from the semiconductor industry and describes the performance measures used in our numerical study. It subsequently introduces an analysis framework for implementation of the DDAP approach, and describes each of the stages in detail. Throughout Section 4, the industry case is used to illustrate the application of the framework. Section 5 provides a numerical study focusing on the value and robustness of the DDAP approach. Finally, we draw conclusions and give an outlook on future research in Section 6.

2. Literature review

Our research is related to the streams allocation planning and order promising, inventory rationing, revenue management and due-date assignment and scheduling. Inventory rationing and revenue management methodologies typically include supply planning decisions (see Liu et al. (2015) and Chevalier et al. (2015) for recent reviews). As described in Section 1, we investigate common industrial environments, in which the planning hierarchy (e.g. in advanced planning systems) separates supply decisions from allocation decisions. In this context, allocation decisions should be made based on the given ATP supply. In contrast, due-date assignment and scheduling approaches are used to plan production at the detailed shop floor level (see e.g. Gordon et al., 2012). We, however, investigate supply chain wide planning. As a consequence, we review literature on allocation planning and order promising.

The literature divides *order promising* models into batch and real-time approaches. For a review of the field and a classification scheme for order promising models, see Ball et al. (2004) and Pibernik (2005). Batch order promising approaches imply customer prioritisation possibilities and are therefore operated without preceding AP. Some of the suggested approaches are of heuristic nature. For example, Pibernik (2006) and Jeong et al. (2002) describe a simple rule-based algorithm and a greedy algorithm respectively. Rabbani et al. (2014) develop a genetic algorithm for order promising and scheduling in a multi-machine flow shop production environment. Majority of the proposed methods, however, are based on optimisation. While a few approaches promise orders based on a given finished product supply (Pibernik, 2005, 2006; Seitz and Grunow, 2017), most contributions integrate order promising with production planning functionalities in assemble-to-order or make-to-order environments (Ball et al., 2003; Zhao et al., 2005; Tsai and Wang, 2009; Lin et al., 2010; Gössinger and Kalkowski, 2015), distribution planning functionalities (Jung, 2010) or a combination of both (Venkatadri et al., 2006; Yang and Fung, 2014).

Integrated batch order promising approaches lead to scattered production schedules that necessitate regular order re-promising (see Dickersbach, 2009). Moreover, they are computationally expensive and myopic in nature since they do not take future orders into account (see Pibernik, 2005). Finally, batch order promising approaches assume that customers are willing to wait until they receive an order promise. This is often not the case since short customer response time is perceived as good customer service. In such scenarios, real time order promising

approaches are needed, which require preceding AP mechanisms.

In practice, common *allocation planning* approaches still use simple business rules (Kilger and Meyr, 2015; Cederborg and Rudberg, 2009; Pibernik, 2006) even though they are known to increase the bullwhip effect (see e.g. Bakal et al., 2011). However, many more sophisticated approaches can be found in the literature. Alarcón et al. (2009), Babarogić et al. (2012) and Ali et al. (2014) propose procedural frameworks that include supply allocation and aim at maximising customer service levels or short-term profits. Other authors study stylised special cases using probabilistic modelling to derive algorithms (Pibernik and Yadav, 2009; Kloos et al., 2018; Kloos and Pibernik, 2020) or structural characteristics of the optimal order acceptance policy (Chiang and Wu, 2011; Gao et al., 2012). Many scholars present linear programming-based approaches in which the aim is to maximise overall profits by integrating AP with production planning in assemble-to-order or make-to-order environments (Ball et al., 2004; Ervolina et al., 2009; Chen and Dong, 2014; Chiang and Hsu, 2014) or in make-to-stock environments (Lebreton, 2015; Meyr, 2009; Cano-Belmán and Meyr, 2019).

Vogel (2014) proposes a method for multi-stage customer hierarchies. He shows that the approach can lead to higher profits compared with profits achieved by an optimal central allocation approach, if demand forecast accuracy is very low. Framinan and Perez-Gonzalez (2016) study the effect of customer forecast accuracy and bias on unused capacity and customer service level in semiconductor manufacturing. They find that forecast bias has a negative effect on the effectiveness of ATP allocations, which is positively correlated to the extent of the bias.

For an up-to-date overview and discussion of demand fulfilment and order promising in semiconductor supply chains, we refer the reader to Mönch et al. (2018).

Table 1 presents an overview of publications on AP and order promising and compares their demand fulfilment approaches to our data driven allocation planning (DDAP) approach in the categories of decision criteria, incorporated demand data, AP frequency and AP level. Even though Meyr (2009) already called for customer ordering behaviour to be included in the profitability evaluation of a customer, neither he nor other researchers have proposed a relevant allocation methodology. The DDAP approach is the first approach that systematically incorporates customer behaviour (i.e. customer forecast bias) as a decision criterion into its allocation planning method. The DDAP approach is also the only approach planning supply allocations in a short-term rolling horizon fashion. Further, the DDAP approach belongs to the very few methodologies incorporating both forecasts and orders into their decision models and considering the importance and profitability of customers.

3. Data driven allocation planning methodology

Fig. 3 gives an overview of our DDAP methodology. It is developed for the single-product case and divided into a mid-term and two short-term planning processes, executed in a rolling horizon fashion. Our approach can also be applied to the multi-product case provided there is no substitution between products. The mid-term customer PAS determination process, described in Section 3.1, uses the profitability and historical forecast accuracy of the customers to determine their individual profitability accuracy score (PAS_i). The resulting score of individual customers is used in the decisions of the short-term processes AP and order promising.

The short-term AP method, which is developed in Section 3.2, reserves supply quantities becoming available in planning period $t \in T$ for forecasted demand being due in period $\tau \in T$. It fulfils the demands according to the PAS of the customers. The AP step is part of the short-term planning of the company. It is so because industrial customers update their demand forecasts in a frequent, short-term manner (e.g. every day). Accordingly, supply allocations are updated with a high frequency in order to make use of the newly available data.

The real-time order promising approach employed in our methodology is adapted from Meyr (2009). It uses the supply allocations to generate order promises $p_{it'\tau'}$ for incoming orders $o_{it'}$ in real-time. Here t' stands for the delivery period, while τ' represents the time period of the requested delivery date. Note that customers do not have to place an order in every time period and there is no predetermined sequence in which the customers order. We follow a nesting policy in the order promising step; i.e. orders can consume AATP quantities, which are reserved for customers being less preferred than the customer placing the order. The order promising model is presented in Appendix B.

Finally, we illustrate the advantages of our AP model over conventional allocation planning (CAP) in Section 3.3.

3.1. Mid-term customer PAS determination

For the DDAP approach, the concepts of the forecast error and the forecast bias have to be distinguished. While the forecast error is simply defined as the deviation of the customer forecast from the final order, the forecast bias denotes the structural or strategic deviation. Hence, the forecast error consists of a random component and a systematic or strategic forecast bias. If the forecast bias of a customer is positive, they systematically or strategically inflate their demand forecasts, which we assume to be the result of rationing gaming behaviour. We hence ignore other possible reasons for the forecast bias. The DDAP approach identifies rationing gaming of the customers and considers it in allocation planning in order to incentivise customers to truthfully forecast their demands.

As described in Section 1 industrial customers displaying rationing gaming behaviour usually do not inflate their demand forecasts by a fixed ratio. Instead, constant demand forecasts of the (anticipated) maximum possible order size are given; the actual order sizes then experience substantial volatility (see Fig. 2). Hence, an allocation approach will not lead to an efficient supply allocation if it integrates

Table 1
Literature gap analysis.

	decision criteria			demand data		AP frequency	AP level[a]
	customer profitability	customer behaviour	customer importance	forecasts	orders		
Alemany et al. (2015)	x				x	mid-term	seg.
Cano-Belmán and Meyr (2019)	x			x		mid-term	seg.
Chen and Dong (2014)	x			x		mid-term	seg.
Chiang and Hsu (2014)	x			x	x	mid-term	seg.
Ervolina et al. (2009)	x			x		mid-term	seg.
Framinan and Perez-Gonzalez (2016)	x			x		mid-term	cust.
Meyr (2009)	x		x	x	x	mid-term	seg.
Pibernik (2006)				x	x	mid-term	seg.
Pibernik and Yadav (2009)			x		x	mid-term	seg.
Vogel (2014)	x			x		mid-term	seg.
DDAP	x	x	x	x	x	short-term	cust.

[a] Allocation planning (AP) levels; seg.: customer segments; cust.: individual customers.

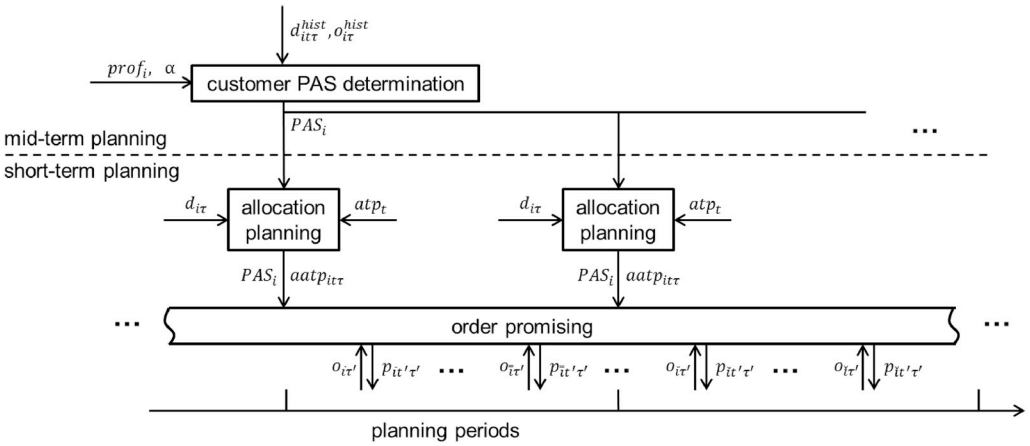

Fig. 3. Rolling horizon scheme for data driven allocation planning.

demand forecast bias information by discounting demand forecasts by a certain predetermined constant that represents the strategic forecast inflation of the customer. However, it is still possible to identify strategic gaming behaviour of the customer by determining the statistically significant positive deviation of the customers' forecasts from their final orders.

In our approach, we determine customer scores PAS_i, which are composed of the historical forecast accuracy acc_i and profitability $prof_i$ of the customers $i \in I$. They represent the priority of a customer for the supply allocation decision. Note that, here, acc_i is defined as the degree to which the forecast of customer i is free of a strategic forecast bias. The measure does not take the random component of the forecast error into account.

A factor $\alpha \in [0;1]$ represents the decision maker's trade-off between forecast accuracy acc_i and profitability $prof_i$ in the determination of PAS_i. At an α -value of 0 and 1, PAS_i is determined solely by $prof_i$ and acc_i, respectively.

The customer scores are determined using a four-step data analysis, which is executed in a mid-term rolling horizon fashion, e.g. once every year. For this analysis, we analyse the customer profitability $prof_i$, the historical demand forecast $d^{hist}_{it\tau}$ provided by customer i in period t for period, and the history of orders o^{hist}_{it} placed by customer i with a due-date in period τ.

The DDAP approach is designed for large industrial suppliers, which usually only serve large customers with a significant revenue contribution directly. Other, smaller customers are served through distributors, who cumulate the demands of numerous small customers and thus also act as large customers. In such environments, usually all direct customers place a demand forecast in every time period. Hence, the ordering frequency of all customers is identical and does not affect the below described determination of the statistically significant forecast bias of individual customers. For each customer, there is hence substantial historical order data available. For new products, there is generally information on similar products available.

Step 1: The error $e_{it\tau}$ of every $d^{hist}_{it\tau}$ is calculated for the past $\tau \in T^{hist}$ consecutive time periods using Equation (1). To base the customer scores on a sufficiently large dataset, T^{hist} should contain more than 30 time periods. However, in order to mitigate the risk of wrong customer scores due to changing customer behaviour over time, should be limited to data of maximum one year.

$$e_{it\tau} = 1 - \frac{o^{hist}_{it}}{it\tau}. \tag{1}$$

Equation (1) is based on the assumption of large industrial customers described above. Hence, $d^{hist}_{it\tau} > 0$ for all i, t and τ. For environments, in which $d^{hist}_{it\tau}$ can be 0, $e_{it\tau}$ could be formulated in accordance to the symmetric mean absolute percentage error (see e.g. Armstrong, 1985 or Ott et al., 2013). In this case, if $d^{hist}_{it\tau} = o^{hist}_{it} = 0$, $e_{it\tau}$ has to be defined as zero as well.

Step 2: In order to separate strategic gaming behaviour from forecasting errors, which increase with the forecast horizon, we determine the average forecast error $\bar{e}_{ih}$ for different forecast horizons h with Equation (2). The horizon h is defined as the time interval between the provision of the forecast and its indicated delivery date. Customers are forecasting their demands every time period for all horizons with a maximum horizon of h^{max}.

$$\bar{e}_{ih} = \frac{1}{\left|T^{hist}\right|} \sum_{\tau \in T^{hist}} e_{i(\tau-h)\tau}. \tag{2}$$

The DDAP methodology aims at measuring the strategic component of $e_{it\tau}$. It therefore allows negative and positive values for $e_{it\tau}$ that origin from random forecasting errors to cancel each other out in Equation (2).

Step 3: To identify rationing gaming behaviour of the customers, it is tested if the values of $\bar{e}_{ih}$ are statistically significantly larger than zero. I. e. the null hypothesis $H0 : \bar{e}_{ih} \leq 0$ is tested using a suitable test statistic. With the result of this statistical test, the forecast bias b_{ih} of customer i for horizon h is determined using Equation (3). If $\left|T^{hist}\right| \geq 30$ the central limit theorem can be applied to assume that the distribution of $\bar{e}_{ih}$ can be approximated by a normal distribution. Then, the student's t-test can be used to determine the values of b_{ih}.

$$b_{ih} = \begin{cases} 0, & \text{if } H0 \text{ is accepted} \\ \bar{e}_{ih}, & \text{if } H0 \text{ is denied} \end{cases} \tag{3}$$

Next, the average demand bias b_i of customer i is determined using Equation (4). In Equation (5), the forecast accuracy of a customer is defined as the complement of the average demand bias.

$$b_i = \frac{1}{h^{max}} \sum_{h=0}^{h^{max}} b_{ih} \tag{4}$$

$$acc_i = 1 - b_i \tag{5}$$

Step 4: Finally, the values of $prof_i$ and acc_i are normalised using Equations (6) and (7) and the values of PAS_i are calculated with Equation (8).

$$prof_i^{norm} = \frac{prof_i - \min_i(prof_i)}{\max_i(prof_i) - \min_i(prof_i)} \tag{6}$$

$$acc_i^{norm} = \frac{acc_i - \min_i(acc_i)}{\max_i(acc_i) - \min_i(acc_i)} \tag{7}$$

$$PAS_i = (1-\alpha)\cdot prof_i^{norm} + \alpha \cdot acc_i^{norm} \tag{8}$$

To ensure an incentive for truthful forecasting also for the most profitable customer i^p, there must be another customer $i \neq i^p$ such that $PAS_{ip} < PAS_i$. To achieve this, the following condition for α must hold:

$$\alpha > \frac{prof_{ip}^{norm} - prof_i^{norm}}{prof_{ip}^{norm} - prof_i^{norm} + acc_i^{norm} - acc_{ip}^{norm}} \tag{9}$$

3.2. Allocation planning model

In allocation planning, the ATP supply atp_t becoming available at the beginning of time period t is allocated to the demand forecast $d_{i\tau}$ of customer i due in period τ. The result of the process is the AATP quantity $aatp_{it\tau}$.

Our allocation planning model is an adaptation of the AP approach in Meyr (2009). Though a heuristic, such as a greedy algorithm could be used to find an optimal solution to the allocation planning problem, we choose a linear programming model because of its flexibility in terms of adding additional constraints like minimum allocation quantities for customers resulting from contractual obligations of a supplier. Furthermore, the model can easily be extended with more complex constraints, e.g. for service level balancing between customers over time or allocation planning considering substitution of products. Such problems, however, cannot be solved to optimality by a heuristic anymore.

Note that the DDAP approach, which is described by Equations (10)–(13), differs from the model presented in Meyr (2009), which maximises the short-term profit of the supplier. DDAP, in contrast, prioritises the demand fulfilment of preferred customers, thereby explicitly considering customer forecast accuracy in addition to customer profitability. Moreover, while in Meyr (2009) the supply is allocated to customer segments, our model allocates it to individual customers. Finally, while the model in Meyr (2009) allows free ATP quantities, which are available for consumption by all customers, in the DDAP approach the entire available ATP supply must be allocated to the customers. Since we assume significant scarcity of supply at all points in time, this cannot lead to infeasibilities in the model.Maximise

$$z = \sum_i \sum_\tau \left[\sum_t (PAS_i \cdot aatp_{it\tau}) - \sum_{t<\tau} (c_{t\tau}^e \cdot aatp_{it\tau}) - \sum_{t>\tau} (c_{t\tau}^l \cdot aatp_{it\tau}) \right] \tag{10}$$

subject to

$$\sum_t aatp_{it\tau} \leq d_{i\tau} \qquad \forall i \in I, \tau \in T; \tag{11}$$

$$\sum_i \sum_\tau aatp_{it\tau} = atp_t \qquad \forall t \in T; \tag{12}$$

$$aatp_{it\tau} \geq 0 \qquad \forall i \in I, t \in T, \tau \in T. \tag{13}$$

The objective function (8) maximises the customer-score-weighted supply allocation and penalises early and late demand fulfilment with the factors $c_{t\tau}^e$ and $c_{t\tau}^l$, respectively. It ensures that demands of the customers with high PAS_i -values are satisfied with priority. Constraints (11) ensure that the generated AATP quantities do not exceed customer demand forecasts. Constraints (12) state that the sum of allocated supply quantities must equal the total available ATP quantities. Constraints (13) represent non-negativity constraints.

For the penalty costs for early and late order fulfilment, normally values are assigned such that early order fulfilment (i.e. temporary stock building) is preferred over late order fulfilment. Hence, $c_{t\tau}^e$ and $c_{t\tau}^l$ such that $max(c_{t\tau}^e) < min(c_{t\tau}^l)$, $max(c_{t\tau}^l) < min(PAS_i)$, $c_{t^1\tau}^e < c_{t^2\tau}^e$ for $t^1 > t^2$ and $c_{t^1\tau}^l < c_{t^2\tau}^l$ for $t^1 < t^2$.

3.3. Illustration of approach

Table 2 illustrates the potential advantages of the DDAP approach compared to the CAP approach for a simple single-period AP example. CAP reserves ATP to customer segments. It thereby solely considers segment profitability and satisfies demand forecasts of segments in order of their profitability. All time indices are omitted in Table 2 because a single-period problem is investigated. Furthermore, for simplicity, the

Table 2
Single-product, single-period example: CAP vs. DDAP.

customer i	1	2	3	4	5	Total
profitability $prof_i$ (per-unit)	15	14	13	12	11	–
d_i	100	100	100	100	100	500
o_i	70	60	90	80	100	400
ordering sequence	5	4	1	2	3	–
atp	–	–	–	–	–	350
segment k	1 (customers 1, 2)		2 (customers 3, 4, 5)OT: Keep “(customers 3, 4, 5)” in single line.			–
$prof_k$	14.5		12			–
$aatp_k^{CAP}$	200		150			350
$promise_i^{CAP}$	70	60	90	60	0	280
ending stock	–	–	–	–	–	70
SL^{CAP}	100%	100%	100%	75%	0%	70%
$profit_i^{CAP}$	1050	840	1170	720	0	3780
α	–	–	–	–	–	0.6
$prof_i^{norm}$	1.0	0.75	0.5	0.25	0.0	–
acc_i	0.7	0.6	0.9	0.8	1.0	–
acc_i^{norm}	0.25	0.0	0.75	0.5	1.0	–
PAS_i	0.43	0.24	0.81	0.62	1	–
$aatp_i^{DDAP}$	50	0	100	100	100	350
$promise_i^{DDAP}$	50	0	90	80	100	320
ending stock	–	–	–	–	–	30
SL^{DDAP}	71.4%	0%	100%	100%	100%	74.3%
$profit_i^{DDAP}$	750	0	1170	960	1100	3980

example assumes that $b_i = e_i$.

Note that both our multi-period DDAP approach and the multi-period CAP approach in Meyr (2009) allow re-allocation of ATP to any segment after each period. Within the same period, re-allocations do not happen, which is why we do not consider re-allocation in the single-period example below.

We measure the effect of CAP and DDAP on the demand fulfilment performance by using the customer service level, the profit generated from sales and the ending stock level. All customers forecast a demand d_i of 100 units. However, their order quantities o_i differ from this forecast. The sequence of order reception and the per-unit profit $prof_i$ of the customers are given in the table. We assume a total ATP quantity of 350 units.

CAP allocates 200 units to segment 1 and 150 units to segment 2 and promises a total amount of 280 units, which leads to an ending stock level of 70 units, a total service level of 70% and a profit of 3780.

For the DDAP approach we assume the α -level to be 0.6. The normalised customer profitability $prof_i^{norm}$ and forecast accuracy acc_i^{norm} are calculated on the basis of d_i and o_i using Equations (1) to (7) without time indices. The customer scores PAS_i follow from Equation (8). Based on these customer scores, the AATP quantities given in Table 2 are obtained when the DDAP model is run. After order promising, the DDAP approach leads to a total of 320 units of satisfied orders, an ending stock level of 30 units, a total service level of 74.3% and a profit of 3980, i.e. lower stocks, higher service level and higher profit than the CAP approach. In a multi-period setting, the ending stock is reallocated by both the CAP and the DDAP approach. This supply can thus be used to fill demands later. However, the lower stock level for DDAP leads to a higher on-time service level.

Such results are due to the consideration of forecast bias data on a highly granular level. Note that, for this exemplary case, DDAP approach leads to higher profits than the CAP approach even though the model does not exclusively aim at fulfilling the demands of the most profitable clients. The example further shows that DDAP approach incentivises the customers to forecast their demands truthfully, since higher values of acc_i result in higher service levels. For example, for customer 5 ($acc_5 = 0.9$), SL^{DDAP} is 100% and SL^{CAP} is 0%, while for customer 2 ($acc_2 = 0.6$), SL^{DDAP} is 0% and SL^{CAP} is 100%.

4. Implementation of DDAP: framework and illustration

As depicted in Fig. 3 AP and order promising are run in a rolling horizon scheme. In every planning period, first, AP generates AATP quantities based on ATP and customer demand forecast data. Afterwards, customer orders are realised and promised based on the allocated supply. Then, the planning horizon is rolled over, new demand forecasts and ATP quantities become available and AP is performed again.

The approach is implemented in Java. IBM ILOG CPLEX V12.6.0 is used to solve the linear programming models for AP and order promising. The analysis was performed on a personal computer with an Intel Xeon E7-4860 v2 processor with 2.6 GHz and 32 GB RAM on a 64-bit Microsoft Windows 7 installation.

In this section, we introduce a generic analysis framework for the implementation of the DDAP approach. To ease the understanding, we integrate the discussion of an illustrative case study throughout the discussion of the framework stages. Therefore, we first introduce the characteristics of the industry case (Section 4.1) and the performance measures (Section 4.2) before we outline the framework structure (Section 4.3) and discuss each of its stages in detail (Section 4.4 to 4.7).

Table 3
Dataset for numerical case study.

Products	P1	P2	P3	P4	P5	P6	Total
Number of customers	12	41	23	25	18	26	145
Number of orders	182	1772	723	944	727	820	5168
Average of demand biases b_i	8%	10%	5%	6%	12%	13%	9%
Share of customers with positive bias	50%	82%	48%	83%	71%	72%	67%
Average error of $acc_i^{norm}(\overline{err}^{acc})$	0.029	0.029	0.037	0.039	0.045	0.049	0.038

4.1. Characteristics of illustrative industry case

We illustrate the DDAP approach with a numerical study using historical demand data from the semiconductor manufacturing industry. Here, supply shortage situations appear frequently due to long production cycle times, high capacity investment cost and high demand volatility (see e.g. Ehm et al., 2011). The customers of the industry under review display rationing gaming behaviour.

We use data from a large European semiconductor manufacturer. The dataset contains orders and demand forecasts for six standard products from the automotive and industrial segments of the company. Table 3 gives an overview of the large dataset containing 78 weeks of forecast data for 145 customers and the corresponding 5168 orders, including their arrival time details. The first 52 weeks (in sample) are used to generate the customer scores PAS_i with the four-step data analysis described in Section 3.1. The last 26 weeks (out-of-sample) are used for the numerical study presented in Section 5. The customers in the dataset order with an average lead time of 3 weeks. The case company groups them into three segments.

We calculate the customer forecast bias on a sample of 52 weeks and obtain 52 observations of the demand forecast error $e_{i(\tau-h)\tau}$ for every horizon h and every customer i. We can consequently apply the central limit theorem to assume that the distribution of $\overline{e}_{ih}$ can be approximated by a normal distribution. We therefore use the student's t-test with a significance level of 10% to determine the values of b_{ih} in Equation (3).

When calculating the forecast accuracy for the out-of-sample time period $acc_i^{norm}(out-of-sample)$, we can derive the average error $\overline{err}^{acc}$ of the historical forecast accuracy acc_i using Equation (14).

$$\overline{err}^{acc} = \frac{\sum_{i\in I}\left|acc_i^{norm} - acc_i^{norm}(out-of-sample)\right|}{|I|} \tag{14}$$

Smaller values indicate a higher predictive quality of acc_i^{norm} for the out-of-sample time period. Table 3 shows that the historical forecast accuracy values of the customers are of different predictive quality since the value of $\overline{err}^{acc}$ differs significantly between the products. However, the small values of $\overline{err}^{acc}$ in Table 3 show that for a typical industrial environment, the historical forecast accuracy calculated over a period of 52 weeks is usually of high predictive quality. This observation validates our approach to perform allocation planning at the individual customer level.

For confidentiality reasons, the real profitability of the customers is not provided in the dataset. However, we do have information on the relation of profitabilities of the customers within the dataset. For our numerical study, we assume two scenarios for the real profitabilities of customers in the dataset. In the *extreme case scenario*, the per-piece profitabilities of the most and the least profitable customers are €1 and €0, respectively; in the *realistic case scenario*, these profitabilities are € 0.1 and € 0.067, respectively.

4.2. Assumptions and performance measures

To be able to measure the capability of the DDAP approach to cope with biased demand forecasts from the customers, the following assumptions are made, which eliminate other sources of uncertainty:

1. The supply quantities atp_t are given.
2. Orders will not be cancelled or rescheduled by the customers once they enter the system.

Furthermore, we assume that:

3. Orders can be fulfilled partially and with multiple shipments.
4. If a part of an order cannot be promised when it is received, that part is lost. This is often demand from wholesalers who cannot substantially postpone their downstream sales.

To measure the demand fulfilment performance, we use the on-time service level (OTSL), the total service level (TSL), the profit generated from sales and the average level of stock resulting from excess allocation. To calculate OTSL and TSL, we use Equations (15) and (16), respectively.

$$OTSL = \frac{\sum_{i\in I}\sum_{\tau\in T} p_{i\tau\tau}}{\sum_{i\in I}\sum_{\tau\in T} o_{i\tau}} \tag{15}$$

$$TSL = \frac{\sum_{i\in I}\sum_{\tau\in T}\sum_{t\in T} p_{it\tau}}{\sum_{i\in I}\sum_{\tau\in T} o_{i\tau}} \tag{16}$$

To analyse the incentive for customers i to provide truthful demand forecasts, we use the impact on the service level measures $\Delta OTSL$ and ΔTSL (relative to a situation in which forecast bias is not considered in the allocation, i.e. $\alpha = 0$).

The average level of stock resulting from excess allocation $\bar{s}$ is defined as average over all periods t of s_t, which is the ending stock level at the end of period t. Similar to the service level measures, we introduce $\Delta\bar{s}$ as the average stock level relative to relative to a situation in which $\alpha = 0$.

4.3. Framework structure

The DDAP approach developed in Section 3 allocates ATP supply to individual customers considering customer forecast bias in addition to customer profitability. The use of the approach has several long-term potentials and short-term impacts. To be able to make an informed decision on the extent to which the forecast bias should be included in the allocation process, these potentials and impacts need to be quantified. In this section, we therefore develop an analysis framework consisting of the following stages:

1. Determine the long-term profit and OTSL potentials of truthful customer forecasts (Section 4.4).
2. Determine the size of the impact on service levels and the resulting incentives to provide truthful forecasts (Section 4.5).
3. Analyse the short-term profit impact and overall service level effects across customers (Section 4.6).
4. Use of the data from the previous three sections to determine the extent to which the forecast bias should affect allocation decisions (Section 4.7).

In the following sections, we outline the generic approach before illustrating it for the industry case. Note that the DDAP approach is developed for the single product case. Hence, when applying it to multiple products, the values of PAS_i and α can be determined separately for all products. Hence we illustrate stages 2, 3 and 4 for an example product data set.

4.4. Long-term profit and OTSL potentials

A first indication of the long-term potentials is derived by calculating the demand biases b_i to determine the extent to which the customers exhibit rationing gaming behaviour. The potentials of truthful customer forecasts, i.e. the long-term effect on company profits and OTSL are quantified by eliminating the bias from all customer forecasts in the dataset and using the DDAP approach with an α level of 0. I.e. we set the PAS_i values to $prof_i$ to only consider customer profitability in the customer scoring. Then, we calculate the resulting service levels and profits for the extreme case scenario and compare the results to the service levels and profits resulting from using the DDAP approach with an α level of 0 without modification of customer forecasts.

Note that for the bias free forecasting scenario using an α level of 0 leads to the same results as using any other value for α. This is because, when customer forecast truthfully, acc_i equals 1 for all customers in the dataset and, thus, for all possible α levels (except 1), the differences in the PAS_i values between different customers are only determined by their individual profitability $prof_i$.

Note that this analysis only quantifies the potential benefits of truthful forecasting in allocation planning. Benefits of other planning processes, e.g. mid-term allocation planning, certainly increase the benefits even more. Therefore, the shown analysis strongly motivates the development of demand fulfilment processes, like DDAP, which incentivise customers to forecast their demands truthfully. In addition, more accurate customer forecasts lead to efficiency improvements in other planning processes like, for example, mid-term production planning. These positive effects on the supplier's profitability are not further quantified.

For the illustrative industry case, the positive values of the average of demand biases b_i in Table 3 illustrates that the customers in the dataset exhibit rationing gaming behaviour. However, the share of customers with a positive b_i shows that not all customers in the dataset show strategic gaming.

The quantification of the potential is done at a supply shortage level of 20%, which is defined as the level to which the total customer demand exceeds the total available supply. Table 4 presents the result for all six product data sets. It shows that truthful customer forecasts increase the OTSL and the total profits by on average 41% and 4%. TSL are not affected because all supply is always consumed in a severe supply shortage situation.

4.5. Incentive for customers to forecast truthfully

In this section, we analyse the impact of truthful forecasting on the service level experienced by customers with high and low biases. With this analysis, we aim to show the existence of an incentive for customers to provide truthful forecasts. When customers are aware of the impact of their forecast bias on their service level, they are incentivized to reduce their bias.

To this end, we measure ΔTSL and $\Delta OTSL$ for customers with low and high forecast biases for different levels of α. We first determine the PAS_i values for historical data and then run the DDAP method for each α. For the illustrative industry case, we use the 52 weeks in-sample data and run the DDAP method on this data varying the level of α between 0 and 1 in five equidistant steps. Again, we use a supply shortage level of 20% for the analysis. We show the result for the example of dataset P4. The results for all other products show the same pattern. Therefore, the conclusions drawn for P4 can be generalized for all six investigated

Table 4
Potential benefits of truthful customer forecasts.

Product	Relative difference OTSL	Relative difference TSL	Relative difference profitability
P1	26%	0%	1%
P2	55%	0%	3%
P3	33%	0%	4%
P4	37%	0%	4%
P5	32%	0%	5%
P6	62%	0%	9%
Average	**41%**	**0%**	**4%**

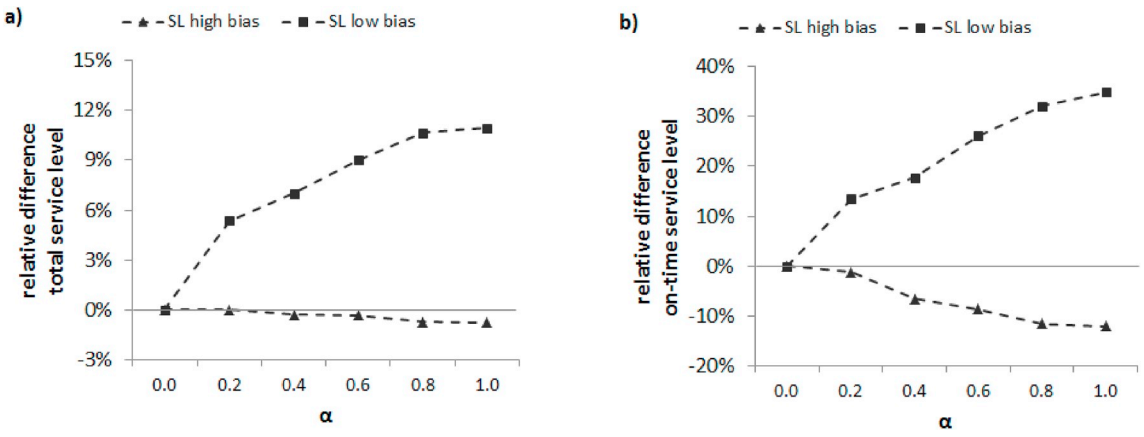

Fig. 4. a) total service level and b) on-time service level for customers with high and low forecast bias.

datasets. Fig. 4 illustrates the influence of the α level on the service level for the first quartiles of customers with the lowest and highest forecast biases. Fig. 4 analyses ΔTSL and $\Delta OTSL$ relative to the DDAP performance at an α level of 0.

As DDAP prioritises customers with low forecast biases more as α increases, ΔTSL and $\Delta OTSL$ increase monotonically in α for customers with low forecast bias and decrease monotonically in α for the customers with high forecast biases. Hence, the gap in ΔTSL and $\Delta OTSL$ between customers with low and high forecast increases in α. At high levels of α, differences in forecast biases between customers have a large impact on their service levels.

As a result of this impact of forecast biases on service levels, especially customers with relatively low profitability have an incentive to increase their customer score PAS_i by increasing their forecast accuracy. To ensure also an incentive for the most profitable customer exits, α must be chosen such that equation (9) holds (in our case, (9) does not impose any constraints because of a number of equally profitable customers).

Note that customers always have to consider that their service level is determined by their own and all other customers' behaviour. Thus, at every point in time, all customers are incentivized to reduce their forecast bias because they have to expect all other customers to do the same. The incentive specifically also exists for the customer with the currently highest *PAS* value as they have to assume that the other customers' reduction of forecast bias will impact their service level negatively in the future if they do not change their behaviour. If, over time, this incentive would remove all bias, then a substantial overall improvement in OTSL is achievable as the analysis of the potential benefit in the previous section has demonstrated. A large α may help in achieving a larger share of this long-term potential sooner. However, the selection of α depends also on its short-term impacts – e.g. on profitability.

4.6. Impact on short-term profits and OTSL

As outlined in the previous section, the incentive depends on the extent to which customers with different forecast biases are served with different service levels. This has obviously an impact on the overall short-term service level across all customers. In addition, the introduction of DDAP leads to a decline in short-term profits compared to an allocation entirely based on customer profitability, unless the forecast bias and profitability are perfectly negatively correlated (i.e. adding forecast bias to the allocation procedure has no impact on the allocation). In reality, this situation does not occur as highly profitable customers often also show a high demand bias, strategically gaming their suppliers' AP.

To quantify these short-term impacts, we measure the impact on profit as well as ΔTSL and $\Delta OTSL$ across all customers for different levels of α using the same analysis employed in the previous section. We additionally measure the relative average stock level $\Delta\bar{s}$. Short-term profits decrease monotonically in α, as DDAP stronger prioritises customers with low forecast biases at the expense of profitable customers as α increases. The overall ΔTSL does not depend on the level of α. The entire ATP supply is always consumed, because supply is short, we allow the fulfilment of orders after their requested delivery date, and we nest AATP quantities.

Increased allocations to customers with more accurate forecasts also mean a decrease of redistributions of supply that was allocated to customers who subsequently ordered less than forecasted. There are three possibilities for redistribution: (1) If redistributions fulfil demand due in the same period, they have no impact on the performance measures. However, (2) if they fulfil demand due in earlier periods, they lead to a decrease of *OTSL*, and (3) if they serve demand due in later periods, they lead to an increase of $\bar{s}$. Hence, the decrease of redistribution volumes resulting from an increased α leads to an increase of *OTSL* and a decrease of $\bar{s}$. We call this the *volume effect*.

Possibilities for redistributions depend on customer order lead times. If customers with shorter lead times receive priority, a redistribution to other customers with demand in the same period (redistribution possibility (1)) is limited. As a consequence, redistribution possibilities (2) and (3) have to be used to a larger extent and $\Delta OTSL$ and $\Delta\bar{s}$ are negatively affected. We call this the *lead-time effect*.

For the illustrative industry case, we again use the 52 weeks in-sample data and run the DDAP method varying the level of α between 0 and 1 using a supply shortage level of 20%. Fig. 5 illustrates the influence of the α level on the overall service levels (ΔTSL and $\Delta OTSL$), average stock levels, and short-term profits resulting from excess allocations for product data set P4. The graph confirms that profits (for the extreme case scenario and the realistic case scenario) are monotonically decreasing in α while the overall ΔTSL does not depend on the level of α.

The results for the overall (average) $\Delta OTSL$ show the impact of the volume effect and the lead-time effect described above. For $\alpha < 0.6$, the volume effect dominates. Excess allocations and resulting redistributions are reduced, leading to a higher $\Delta OTSL$ and a lower $\Delta\bar{s}$.

For $\alpha > 0.6$, the lead-time effect substantially reduces the volume effect. In our case, the customers with low demand biases tend to place their orders later than others. When placing large emphasis on forecast accuracy (high α), the supply allocation for these customers increases. Even though their forecasts are more accurate, they still include a demand bias. However, due to the short lead time, this bias cannot be compensated for by redistributing excess allocation to other demand without reducing the $\Delta OTSL$ or increasing $\Delta\bar{s}$. Therefore, even though the $\Delta OTSL$ for the customers with low forecast bias increases with increase of α (as shown in the previous section), the overall $\Delta OTSL$

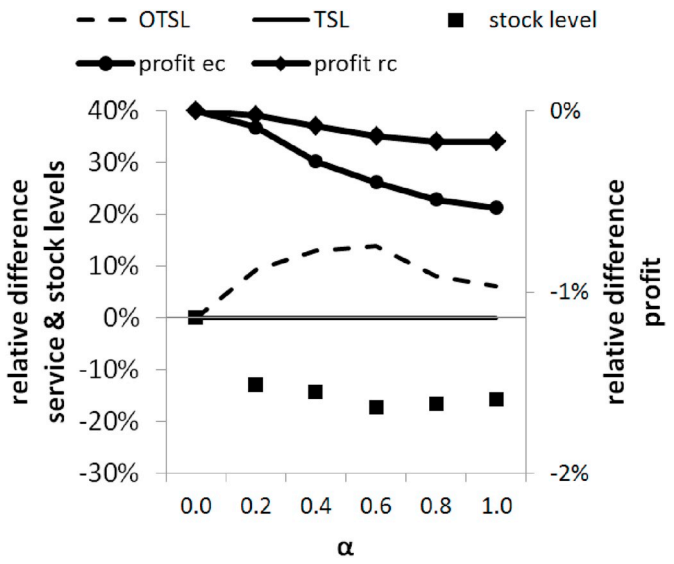

Fig. 5. Service levels, stock levels and short-term profits of P4 depending on.α.

deteriorates for α levels above 0.6. For the same reason, the average stock levels resulting from excess allocation show a minimum at an α -level of 0.6.

4.7. Sizing the incentive: reconciling long-term potentials and short-term impacts

In the previous sections, we went through Stages 1, 2, and 3 of our analysis framework. We discussed the dependence of long-term potentials, incentives, and short-term impacts on the extent to which the forecast bias is considered in our DDAP approach. Along with generic insights, we provided case-specific results. For the implementation of the DDAP approach, we now in Stage 4 have to choose for each product a value for the extent to which the forecast bias is considered in addition to the profitability. In other words, we need to select a value for α when applying equation (8) as part of an implementation of DDAP in an advanced planning system environment. We call this selected value α^*.

In general, if the biases b_i or the long-term potentials determined in Stage 1 are small, then the use of the DDAP approach cannot lead to significant improvements over traditional allocation procedures (in other words, we set $\alpha^* = 0$). The larger the long-term potentials are, the larger the α^* should be.

The analysis for Stage 2 in Section 4.5 showed that a large α^* creates a larger incentive, which may help achieve a larger proportion of the long-term potentials sooner. To ensure that also the most profitable customers are incentivized, the value for α^* should comply with the minimum condition in equation (9).

In addition to the long-term potentials, short-term impacts determined in Stage 3 also influence the selection of α^*. As outlined in Section 4.6, the decrease of short-term profits in α favours a small value for α^*. In relation to the short-term OTSL the best value for α^* depends on the interaction between the volume effect (which favours large values for α^*) and the lead-time effect (which mediates the volume effect).

Stages 1, 2, and 3 of our framework provide the required quantitative information for decision maker to reconcile the above mentioned drivers in determining α^* in Stage 4. In order to relate long-term potentials to short-term impacts, the decision maker has to assess the extent to which the potential can be achieved as well as the temporal dynamics behind achieving the potentials and their dependence on α^*. This assessment depends on numerous situational characteristics such as the sensitivity of customers to incentives (which in turn depends on size and type of the customers) or the communication with the customers, including the communication of the incentives. Due to the uncertainties inherent in these characteristics, the reconciliation with short-term impacts also needs to reflect the risk preferences of the decision makers.

For the illustrative industry case, we demonstrated in Section 4.4 that there are significant long-term potentials (Stage 1), resulting from the gaming behaviour reflected in the biases listed in Section 4.1. A decision maker would therefore aim for a large incentive. From Section 4.5, it is clear that even for smaller levels of α^* (0.2 and above), the incentive is large. The minimum condition in equation (9) is also satisfied for these levels of α^* (Stage 2).

For the short-term impacts, Section 4.6 shows only a very small decrease of α for the short-term profit. In relation to the short-term OTSL, Section 4.6 shows a significant increase up to $\alpha = 0.6$, after which it decreases (Stage 3).

Considering the above results, the short-term profit reductions are relatively small (<0.5%). In contrast, we have shown in Section 4.4 that the profit increase potential is 4% when customers forecast truthfully, combined with a large increase of OTSL (37%). Therefore, α^* is chosen such that the short-term OTSL is maximised, i.e. $\alpha^* = 0.6$ for product dataset P4 (Stage 4). An application of our analysis framework to the other product datasets leads to the following levels of α^* for datasets P1 to P6: 0.6, 0.2, 0.4, 0.6, 0.8, and 0.4.

5. Value and robustness of the DDAP approach

In this section, we analyse the value and robustness of considering forecast bias data in allocation planning. In Section 5.1, we analyse how a DDAP approach would perform under different supply shortage levels. We then compare the performance of DDAP and CAP in Section 5.2 depending on the predictive quality of the historical data. In Section 5.3, we conclude our analysis with investigating the effects of moving the demand fulfilment level from customer segment to individual customer and considering demand bias data separately.

5.1. Value of considering forecast bias data

To investigate the benefits of considering forecast bias data in demand fulfilment, we use the DDAP approach on the out-of-sample data and vary the level of supply shortage from 10% to 30% using, first, the six α^* values determined in Section 4.7 (note that all analyses use α^* values determined for a supply shortage of 20%) and, second, an α level of 0.

Fig. 6 shows the service level averages over all products for the DDAP approach at different levels of supply shortages. The findings are displayed relative to the performance of the DDAP approach at an α level of 0, i.e. only considering customer profitability.

The results in the figure clearly confirm the effectiveness of the incentive and the increase of OTSL for out-of-sample data. This also holds for supply shortage levels of 10% and 30% even though the results were calculated with α^* values determined for the 20% supply shortage level. This shows that the results are robust against fluctuations in the supply shortage level. This simplifies an implementation of the DDAP approach in an advanced planning system environment. The graphs also show that the positive effect of considering forecast accuracy data on the overall OTSL increases with the level of supply shortage. The reason for the amplification of the given effects lies in the scarcity of supply itself. The influence of allocating supply more efficiently among customers on the service levels grows with the level of supply shortage.

5.2. Value of DDAP depending on the predictive quality of historical data

Table 3 in Section 4.1 shows that the historical forecast accuracy values of the customers are of different predictive quality since the value of $\overline{err}^{acc}$ differs significantly between the products. In this section, we compare the demand fulfilment performance of DDAP and CAP

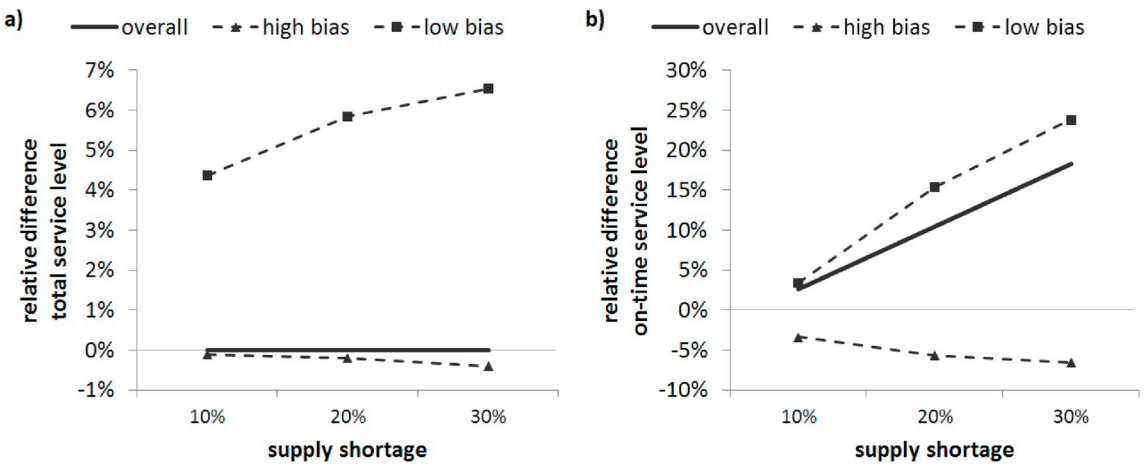

Fig. 6. a) total service level and b) on-time service level in dependence of the level of supply shortage.

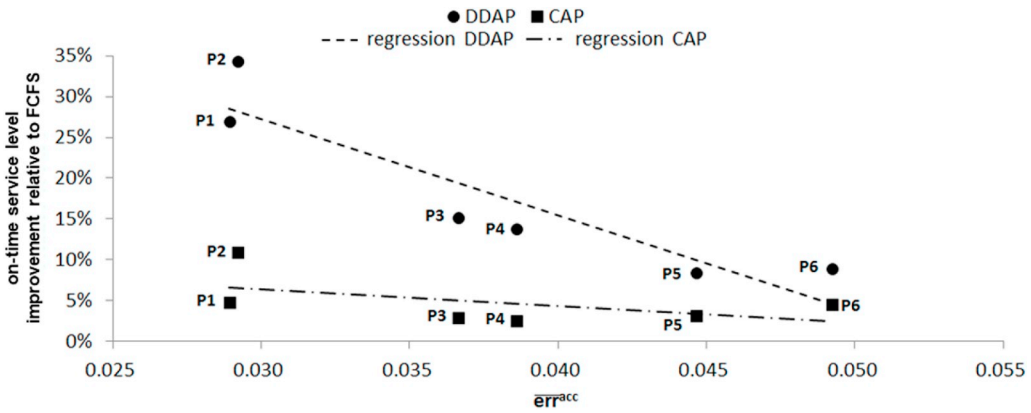

Fig. 7. Service levels of DDAP and CAP in dependence of the predictive quality of data.

depending on the predictive quality of the historical data. A first-come-first-served (FCFS) real-time order promising without preceding AP serves as a benchmark. For comparability reasons, we set the frequency of the CAP approach to the frequency of the DDAP approach, i.e. one week. Fig. 7 depicts the OTSL improvements of DDAP and CAP relative to the FCFS approach over the average error of the historical forecast accuracy values $\overline{err}^{acc}$.

Like the approach in Meyr (2009), CAP allocates supply to customer segments solely based on customer profitability data. The models used for allocation planning and order promising in the CAP approach are presented in Appendices A and B. The case company groups their customers into three different customer segments. These are used in the CAP approach.

The graph shows that the DDAP approach outperforms the CAP approach for all the studied product data sets, i.e. predictive qualities of the historical forecast accuracy values, because of its more efficient supply allocation; thus, exploiting the availability of highly granular data. However, the advantage decreases with increasing $\overline{err}^{acc}$. This is intuitive since high levels of predictive quality indicate stable forecasting behaviour of the customers. Consequently, the chosen α^* values also lead to high benefits in the out-of-sample data. On the other hand, in a situation where $\overline{err}^{acc}$ is high, the customers' forecasting behaviour cannot be predicted accurately and the potentials of highly granular data cannot be exploited as efficiently by the DDAP approach.

Table 5
Effects of demand fulfilment granularity and consideration of demand bias on OTSL and profits (results relative to the performance of the CAP approach with customer segments).

	product	customer segments		individual customers	
		OTSL	profit	OTSL	profit
CAP($\alpha = 0$)	P1	–	–	+0.0%	−0.01%
	P2	–	–	+0.9%	0.01%
	P3	–	–	+1.0%	−0.01%
	P4	–	–	+0.1%	0.00%
	P5	–	–	+0.3%	0.00%
	P6	–	–	+1.1%	0.01%
DDAP ($\alpha = \alpha^*$)	P1	+3.3%	−0.03%	+8.9%	−0.13%
	P2	+2.4%	−0.02%	+9.3%	−0.12%
	P3	+4.0%	−0.07%	+5.0%	−0.09%
	P4	+3.1%	−0.02%	+5.1%	−0.06%
	P5	+2.2%	−0.01%	+2.4%	−0.02%
	P6	+3.2%	−0.03%	+1.5%	−0.02%

5.3. Effects of demand fulfilment on customer service level and consideration of demand bias

The DDAP approach differs in two respects from the CAP approach. First, AP and order promising are done on individual customer level. Second, the demand bias data is taken into account. In this section, we evaluate the effects of both these aspects separately.

To be able to measure the effect of considering forecast bias data when fulfilling demands at the customer segment level, we adapt the

DDAP model to do AP and order promising on customer segment level. For this, first, the average of the in-sample forecast bias b_k and the profitability $prof_k$ of the customer segment k are calculated by taking the average of the b_i values and the $prof_i$ values for all the customers i belonging to the segment k. Here, again, the customer segments of the case company are used. Afterwards, the normalised forecast accuracy acc_k^{norm} and profitability $prof_k^{norm}$ of the segments as well as the segment scores PAS_k^{seg} are generated using Equations (6)–(8) for customer segments. Finally, the so generated PAS_k^{seg} are used as $score_k^{seg}$ in the allocation planning. The resulting AP model is described in Appendix A.

Table 5 presents the effects of fulfilling demands at customer segment or individual customer levels as well as using profitability of customers for customer scores alone or additionally considering demand bias data. When reading the table from left to right, the effect of changing the demand fulfilment granularity from customer segment to individual customer is shown. Reading the table from top to bottom shows the effect of considering forecast bias information in customer scores. All numbers are relative to the performance of the CAP approach.

Changing the demand fulfilment granularity from customer segment to individual customer increases the OTSL between 0.0 and 1.1 percentage points and the short-term profit (real case scenario) between −0.01 and 0.01 percentage points. Therefore, changing the demand fulfilment granularity alone only has a weak effect on demand fulfilment performance. This is because, first, pooling effects within customer segments and, thus, flexibility in order promising is lost. Second, the effect on the OTSL is positive due to the fact that most customers in the dataset display a tendency of inflating their demand forecasts, and the highly profitable customers place their orders earlier than others. Hence, excessive AATP quantities for the highly profitable customers can be used to fulfil orders of the less profitable customers, which are received later.

Considering demand bias data when fulfilling demands at the customer segment level increases the OTSL moderately between 2.2 and 4.0 percentage points. Short-term profits are decreasing slightly between −0.07 and −0.01 percentage points. The reason why the benefits in terms of OTSL are not more significant is that customers within these profit segments are very heterogeneous in terms of their individual forecast accuracy. As a consequence, customers with high demand biases are grouped with customers forecasting their demands truthfully. Raising the α level leads to a prioritisation of customer segments with low average demand bias.

Nevertheless, the demand biases of individual customers within prioritised segments can still be high. The overall short-term profit decreases since, on average, customers with higher forecast accuracy tend to have lower profitability.

Finally, combining both these aspects results in the highest service level for all products except P6. The most distinct effects are achieved for the products P1 and P2, for which the OTSLs of both the DDAP approach on customer segment level and the CAP approach on individual customer level are significantly outperformed. Analogously, a decrease of the overall short-term profits is most pronounced for these products. For the products P3 to P5, the increase of OTSL compared to the CAP approach at the individual customer level is moderate while the advantages compared to the DDAP approach at the customer segment level are low. The overall short-term profits change analogously.

The rationale behind the different magnitudes of leveraging effect is the difference between the predictive qualities of forecast accuracy values. While these values are of high predictive quality for the product datasets P1 and P2, they are of moderate quality for P3 to P5, and give only little indication of the forecast accuracy of customers out-of-sample in the dataset P6. The higher the predictive quality of the forecast accuracy values, the less temporary are the stocks resulting from excess allocation that are built by the DDAP approach. In dataset P6, $\overline{err}^{acc}$ is exceptionally high so that the usage of forecast accuracy data at the individual customer level leads to a negative effect on the OTSL compared to the DDAP approach at the customer segment level.

We, therefore, draw the following conclusions. First, the use of the DDAP approach is only meaningful when additional data at the individual customer level is considered. Second, when looking at OTSL, AP at the individual customer level, considering demand bias data robustly outperforms the conventional allocation planning approaches, allocating supply to customer segments purely based on profitability data. In exceptional cases, when the predictive quality of the historical forecast accuracy is very low, it is, however, more beneficial to fulfil demands at the customer segment level. Third, the benefits of the DDAP approach increase with the predictive quality of the historical data on customer forecast accuracy.

6. Summary and conclusion

The recent advances in big data analysis tools represent an opportunity for companies to further improve their demand fulfilment processes. In particular, the exploitation of data on the ordering behaviour of customers, e.g. their forecast biases, can enable companies to improve the accuracy and robustness of their order promises and increase the performance of their demand fulfilment systems. Here, a first step towards integrating such newly available data into allocation planning and order promising has been taken.

Research has shown that customers systematically inflate their demand forecasts in supply shortage situations to make suppliers increase the supply quantities reserved for them and finally satisfy their total actual demand. This behaviour called the rationing gaming significantly impairs the ability of the supplier to efficiently allocate current and future supply to customers.

We propose an allocation planning methodology called data driven allocation planning (DDAP), which considers the data on individual customers and products by allocating supply on a highly granular level, taking systematic biases in demand forecast into account. The DDAP approach is designed for industrial environments using advanced planning systems for demand fulfilment planning. It is therefore not only of high practical applicability, but also of high practical relevance, as it supports an efficient supply allocation, leading to a significant potential in customer service levels and long-term profitability compared to conventional allocation approaches. In addition, our approach reduces stocks that are created by excess allocation to customers showing low forecast accuracy. By increasing service levels for customers with high forecast accuracy, the approach incentivises customers to provide truthful demand forecasts and, thus, counteracts rationing gaming. To facilitate the implementation of the DDAP approach, we develop an analysis framework for a structured derivation of the information required on long-term potentials, incentive sizes, and short-term impacts to decide on the extent to which the forecast bias is considered.

We test the DDAP approach in a numerical study using data from a large European semiconductor manufacturer. The analysis proves that allocating supply to individual customers is only valuable when additional data on this granularity level is available and taken into account. Then, DDAP leads to a significant increase of allocation efficiency. The improvements become more distinct with growing scarcity of supply and increased predictive quality of the available data.

Our results show increased service levels for customers that do not perform rationing gaming, i.e. that provide their demand forecasts without a strategic bias. We hereby demonstrate the capability of DDAP to incentivise customers for truthful forecasting. Removing a deliberately added forecast bias does not induce any costs, such as removing a non-intentional forecast error would. The latter may require planning tool and process improvements. The former does not require such efforts. Customers must therefore not trade-off between an increase of service levels and implementation costs for obtaining forecasts with higher accuracy. Also the implementation costs for the supplier are limited. Granular data on individual customer forecasts and orders are readily available in existing databases. However, the adjustment of the

allocation procedure in the advanced planning system does cause some implementation costs.

The data we use in our numerical study indicates that mostly the customers with large demand volumes and high profitability show significant rationing gaming behaviour. Such customers are typically of high importance for the supplier. Applying the DDAP approach in such an environment would lead to a decline of the service levels for these customers. Companies with such a customer structure may therefore be reluctant to implement our methodology in their demand fulfilment processes, because they may fear to lose their most important customers. This may especially be the case, if key account managers have large influence. However, principle decisions on demand fulfilment processes are made on more senior management levels with a more comprehensive perspective on the supply chain. Senior managers are able to account for the long-term profit increase potential, which we have shown to be substantial. They would have to combine the implementation of the proposed approach with an alignment of the incentive structure for the sales organization.

Rationing gaming usually only occurs in (severe) supply shortage situations. In industries like semiconductor manufacturing, these shortage situations are typically of global nature. Customers are hence not able to source from competitors, because other suppliers do not have spare capacity available to fill additional demand on short notice. Moreover, production capacity in the semiconductor industry cannot easily be expanded as it requires substantial investments and installation and ramp-up time. In addition, suppliers often manufacture highly customer-specific products. This makes supplier qualification costly and time-consuming. Hence, customers regularly source from only one or two suppliers.

In consequence, these type of semiconductor manufacturers have a strong market power; also towards their largest and most important customers. This enables semiconductor manufacturers – as well as suppliers in industries with similar characteristics – to apply DDAP in their planning processes and communicate its effects to their customers without losing market share. Furthermore, the increase of forecast accuracy will enable suppliers to serve all of their customers with higher service levels in the long term.

In order to incentivise customers to truthfully forecast their demands, suppliers also need to communicate the effects of implementing DDAP on the individual service levels of customers depending on their forecast accuracy. This might also be part of more collaborative planning initiatives between supplier and customer, such as collaborative planning, forecasting, and replenishment (CPFR). In many cases, such collaborations are common practice in the industry. They are a good means for the supplier to make forecast biases and their effects on service level transparent to the customers. One possibility to conduct such communication is to inform customers about their biases in comparison with an anonymized peer group. This could be combined with information on the service level increase that would have been achievable with an unbiased forecast. However, suppliers always need to consider the potentially negative effects of such communication on the long-term relationship with their customers.

Our work uses data from the semiconductor industry. The generalizability of our results to other sectors has not been proven. The semiconductor industry is characterized by high volatility of demand and particularly short product life cycles. The fact that DDAP is shown to be beneficial in this highly dynamic environment indicates possible applicability in other industries, although further work needs to be done in other sectors or with more stylised models to show or prove its wider applicability.

The findings of our study indicate that an allocation approach additionally accounting for data on customer order lead times could lead to even better results. In our work we assume given supply. However, in the mid-term, the supply mix may be adapted to demand developments, even if the overall supply may remain to be short. Such supply mix changes are frequently implemented in the semiconductor industry, but are out of the scope of our paper that focusses on short-term demand fulfilment. Follow-up work could extend our work by investigating the incorporation of demand fulfilment procedures in a hierarchical advanced planning environment that allows for reactions on multiple planning levels. Furthermore, it would be interesting to integrate order lead-time data on the individual customer level into allocation planning approaches. Additionally, approaches to increase the performance of DDAP in environments with varying predictive quality of historical data have to be developed. Allocation planning methods switching from an individual customer level to a customer segment level when the volatility of demand biases reaches a certain threshold should be investigated. Moreover, considering substitute products in the allocation planning decision would be interesting; especially, taking into account data on the individual willingness of customers to substitute products has so far not been addressed. Finally, the examination of the performance of DDAP under supply uncertainty or demand uncertainty after order arrival indicates possibility of a further extension. Hereby, order rescheduling and cancellation rules, which industrial suppliers and customers agree upon in supply contracts, have to be taken into consideration. Obviously, several of these further research directions can be realised by exploiting additional data.

CRediT authorship contribution statement

Alexander Seitz: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Martin Grunow:** Conceptualization, Writing - review & editing, Supervision. **Renzo Akkerman:** Conceptualization, Writing - review & editing, Supervision.

Acknowledgement

We thank the three anonymous referees for their constructive comments, which helped substantially improve this paper.

Appendix A

Allocation planning at customer segment level

In our numerical study, allocation planning at the customer segment level is done with the following linear programming model, which is an adaption from Meyr (2009).

Maximise

$$z = \sum_{k}\sum_{\tau}\left[\sum_{t}\left(prof_{k}^{norm}\cdot aatp_{kt\tau}\right) - \sum_{t<\tau}\left(c_{t\tau}^{e}\cdot aatp_{kt\tau}\right) - \sum_{t>\tau}\left(c_{t\tau}^{l}\cdot aatp_{kt\tau}\right)\right] \tag{A.1}$$

subject to

$$\sum_{t} aatp_{kt\tau} \leq d_{k\tau} \qquad \forall k \in K, \tau \in T \tag{A.2}$$

$$\sum_{k}\sum_{\tau} aatp_{kt\tau} = atp_{t} \qquad \forall t \in T \tag{A.3}$$

$$aatp_{kt\tau} \geq 0 \qquad \forall k \in K, t \in T,\ \tau \in T \tag{A.4}$$

The objective function (A.1) maximises the segment-score-weighted supply allocation $aatp_{kt\tau}$ to customer segment k using the supply becoming available in period t to satisfy demand $d_{k\tau}$ being due in period τ. The factors $c^{e}_{t\tau}$ and $c^{l}_{t\tau}$ penalise early and late fulfilment of demand, respectively. Constraints (A.2) ensure that the sum of allocated supply quantities does not exceed customer demands, while constraints (A.3) state that the sum of allocated quantities must equal the total available supply atp_t.

Order promising model

The order promising model we use for our numerical study is adapted from Meyr (2009). It decides on the portions of allocated supply $x_{it'}$ which are available in period t' and are employed to fulfil an order of $o_{i^*\tau'}$ product units from customer i^* which is due in period τ'.

Maximise

$$z = \sum_{i \in I_{i^*}} \left[\sum_{t'} (PAS_i x_{it'}) - \sum_{t' \leq \tau'} \left(c^{e}_{t'\tau'} \cdot x_{it'} \right) - \sum_{t' > \tau'} \left(c^{l}_{t'\tau'} \cdot x_{it'} \right) \right] \tag{B.1}$$

subject to

$$\sum_{i \in I_{i^*}} \sum_{t'} x_{it'} \leq o_{i^*\tau'} \tag{B.2}$$

$$0 \leq x_{it'} \leq \sum_{\tau} aatp_{it'\tau} \qquad \forall i \in I_{i^*}, t' \in T \tag{B.3}$$

The objective function (B.1) maximises the amount of allocated supply promised to the customer. The factors $c^{e}_{t\tau}$ and $c^{l}_{t\tau}$ penalise early and late fulfilment of demand, respectively. Constraints (B.2) state that the sum of consumed supply for the incoming order must not exceed the ordered quantity. Constraints (B.3) ensure that the allocations $aatp_{it'\tau}$ set at the allocation planning step are not exceeded.

The model allows nesting of supply allocated to customers. The customers from which customer i^* is allowed to consume allocated supply are represented in the set I_{i^*}, which contains all customers for which the Inequality (B.4) holds.

$$PAS_{i^*} \geq PAS_i \tag{B.4}$$

In case allocation planning is done at the customer segment level, we replace I with K, i^* with k^*, i with k and PAS_i with $prof^{norm}_{k}$ in Equations (B.1) to (B.3). k and k^* represent a customer segment and the customer segment of the ordering customer, respectively. Furthermore, we replace the sets I_{i^*} with K_{k^*}, which represent the set of customer segments from which customer segment k^* is allowed to consume the allocated supply. It contains all customer segments for which Inequality (B.5) holds.

$$prof^{norm}_{k^*} \geq prof^{norm}_{k} \tag{B.5}$$

In our numerical study, early fulfilment of orders is not allowed. Therefore, the promises $p_{it'\tau'}$ are calculated with Equation (B.6).

$$p_{it'\tau'} = \begin{cases} \sum_{t' \leq \tau'} x_{it'} \\ x_{it'} \quad , \forall t' > \tau' \end{cases} \tag{B.6}$$

References

Alarcón, F., Alemany, M.M.E., Ortiz, A., 2009. Conceptual framework for the characterization of the order promising process in a collaborative selling network context. Int. J. Prod. Econ. 120 (1), 100–114.

Alemany, M., Grillo, H., Ortiz, A., Fuertes-Miquel, V., 2015. A fuzzy model for shortage planning under uncertainty due to lack of homogeneity in planned production lots. Appl. Math. Model. 39 (15), 1–35.

Ali, M., Gaudreault, J., D'Amours, S., Carle, M.-A., 2014. A multi-level framework for demand fulfillment in a make-to-stock environment - a case study in Canadian softwood lumber industry. MOSIM 2014, 10ème Conférence Francophone de Modélisation, Optimisation et Simulation, November 2014. Nancy, France.

Armstrong, J., 1985. Long Range Forecasting: from Crystal Ball to Computer. Wiley-Interscience, New York.

Babarogić, S., Makajić-Nikolić, D., Lečić-Cvetković, D., Atanasov, N., 2012. Multi-period customer service level maximization under limited production capacity. Int. J. Comput. Commun. Contr. 7 (5), 798–806.

Bakal, I., Erkip, N., Güllü, R., 2011. Value of supplier's capacity information in a two-echelon supply chain. Ann. Oper. Res. 191 (1), 115–135.

Ball, M.O., Chen, C.-Y., Zhao, Z.-Y., 2003. Material compatibility constraints for make-to-order production planning. Oper. Res. Lett. 31 (6), 420–428.

Ball, M.O., Chen, C.-Y., Zhao, Z.-Y., 2004. Available to promise. In: Simchi-Levi, D., Wu, D.D., Shen, Z.-J. (Eds.), Handbook of Quantitative Supply Chain Analysis. Springer, New York, pp. 447–483.

Cano-Belmán, J., Meyr, H., 2019. Deterministic allocation models for multi-period demand fulfillment in multi-stage customer hierarchies. Comput. Oper. Res. 101, 76–92.

Cederborg, O., Rudberg, M., 2009. Customer segmentation and capable-to-promise in a capacity constrained manufacturing environment. In: Proceedings of the 16th International Annual EurOMA Conference, June 2009, Göteborg, Sweden.

Chen, J., Dong, M., 2014. Available-to-promise-based flexible order allocation in ATO supply chains. Int. J. Prod. Res. 52 (22), 6717–6738.

Chevalier, P., Lamas, A., Lu, L., Mlinar, T., 2015. Revenue management for operations with urgent orders. Eur. J. Oper. Res. 240 (2), 476–487.

Chiang, C., Hsu, H.-L., 2014. An order fulfillment model with periodic allocation review mechanism in semiconductor foundry plants. IEEE Trans. Semicond. Manuf. 27 (4), 489–500.
Chiang, D.M.-H., Wu, A.W.-D., 2011. Discrete-order admission ATP model with joint effect of margin and order size in a MTO environment. Int. J. Prod. Econ. 133 (2), 761–775.
Dickersbach, J.T., 2009. Supply Chain Management with APO: Structures, Modelling Approaches and Implementation of SAP SCM 2008. Springer, Berlin.
Ehm, H., Ponsignon, T., Kaufmann, T., 2011. The global supply chain is our new fab: integration and automation challenges. In: Advanced Semiconductor Manufacturing Conference (ASMC), 2011 22nd Annual IEEE/SEMI, pp. 1–6.
Ervolina, T.R., Ettl, M., Lee, Y.M., Peters, D.J., 2009. Managing product availability in an assemble-to-order supply chain with multiple customer segments. In: Günther, H.-O., Meyr, H. (Eds.), Supply Chain Planning. Springer, Berlin, pp. 1–24.
Framinan, J.M., Perez-Gonzalez, P., 2016. Available-To-Promise systems in the semiconductor industry: a review of contributions and a preliminary experiment. In: 2016 Winter Simulation Conference (WSC). IEEE, pp. 2652–2663.
Gao, L., Xu, S.H., Ball, M.O., 2012. Managing an available-to-promise assembly system with dynamic short-term pseudo-order forecast. Manag. Sci. 58 (4), 770–790.
Gordon, V., Strusevich, V., Dolgui, A., 2012. Scheduling with due date assignment under special conditions on job processing. J. Sched. 15 (4), 447–456.
Gössinger, R., Kalkowski, S., 2015. Order promising - a robust customer-oriented approach. In: Dethloff, J., Haasis, H.-D., Kopfer, H., Kotzab, H., Schönberger, J. (Eds.), Logistics Management. Springer, Cham, pp. 135–149.
Jeong, B., Sim, S.-B., Jeong, H.-S., Kim, S.-W., 2002. An available-to-promise system for TFT LCD manufacturing in supply chain. Comput. Ind. Eng. 43 (1–2), 191–212.
Jung, H., 2010. An available-to-promise model considering customer priority and variance of penalty costs. Int. J. Adv. Manuf. Technol. 49 (1–4), 369–377.
Kilger, C., Meyr, H., 2015. Demand fulfilment and ATP. In: Stadtler, H., Kilger, C., Meyr, H. (Eds.), Supply Chain Management and Advanced Planning. Springer, Berlin, pp. 177–194.
Kloos, K., Pibernik, R., 2020. Allocation planning under service-level contracts. Eur. J. Oper. Res. 280 (1), 203–218.
Kloos, K., Pibernik, R., Schulte, B., 2018. Allocation planning in sales hierarchies with stochastic demand and service-level targets. OR Spectrum 1–44.
Lebreton, B., 2015. Integrated campaign planning, scheduling and order confirmation in the specialty chemicals industry. In: Stadtler, H., Kilger, C., Meyr, H. (Eds.), Supply Chain Management and Advanced Planning. Springer, Berlin, pp. 475–485.
Lee, H.L., Padmanabhan, V., Whang, S., 2004. Information distortion in a supply chain: the bullwhip effect. Manag. Sci. 50 (12), 1875–1886.
Lin, J., Hong, I.-H., Wu, C.-H., Wang, K.-S., 2010. A model for batch available-to-promise in order fulfillment processes for TFT-LCD production chains. Comput. Ind. Eng. 59 (4), 720–729.
Liu, S., Song, M., Tan, K., Zhang, C., 2015. Multi-class dynamic inventory rationing with stochastic demands and backordering. Eur. J. Oper. Res. 244 (1), 153–163.
Meyr, H., 2009. Customer segmentation, allocation planning and order promising in make-to-stock production. OR Spectrum 31 (1), 229–256.
Mönch, L., Uzsoy, R., Fowler, J.W., 2018. A survey of semiconductor supply chain models part III: master planning, production planning, and demand fulfilment. Int. J. Prod. Res. 56 (13), 4565–4584.
Ott, H.C., Heilmayer, S., Sng, C.S.Y., 2013. Granularity dependency of forecast accuracy in semiconductor industry. Res. Log. Prod. 3 (1), 49–58.
Pibernik, R., 2005. Advanced available-to-promise: classification, selected methods and requirements for operations and inventory management. Int. J. Prod. Econ. 93 (1), 239–252.
Pibernik, R., 2006. Managing stock-outs effectively with order fulfilment systems. J. Manuf. Technol. Manag. 17 (6), 721–736.
Pibernik, R., Yadav, P., 2009. Inventory reservation and real-time order promising in a make-to-stock system. OR Spectrum 31 (1), 281–307.
Rabbani, M., Monshi, M., Rafiei, H., 2014. A new AATP model with considering supply chain lead-times and resources and scheduling of the orders in flowshop production systems: a graph-theoretic view. Appl. Math. Model. 38 (24), 6098–6107.
Seitz, A., Grunow, M., 2017. Increasing accuracy and robustness of order promises. Int. J. Prod. Res. 55 (3), 656–670.
Tsai, K.-M., Wang, S.-C., 2009. Multi-site available-to-promise modeling for assemble-to-order manufacturing: an illustration on TFT-LCD manufacturing. Int. J. Prod. Econ. 117 (1), 174–184.
Venkatadri, U., Srinivasan, A., Montreuil, B., Saraswat, A., 2006. Optimization-based decision support for order promising in supply chain networks. Int. J. Prod. Econ. 103 (1), 117–130.
Vogel, S., 2014. Demand Fulfillment in Multi-Stage Customer Hierarchies. Springer Gabler, Wiesbaden.
WSTS Inc, 2015. World Semiconductor Trade Statistics. Accessed: 08.04.2015. https://www.wsts.org/.
Yang, W., Fung, R., 2014. An available-to-promise decision support system for a multi-site make-to-order production system. Int. J. Prod. Res. 52 (14), 4253–4266.
Zhao, Z.-Y., Ball, M.O., Kotake, M., 2005. Optimization-based available-to-promise with multi-stage resource availability. Ann. Oper. Res. 135 (1), 65–85.