5. Model Extensions: Forecast Errors and Enhanced Consumption Planning

In the previous chapter, a basic formal model of the DMC problem has been introduced and different allocation schemes for multi-stage customer hierarchies have been analyzed. More precisely, a partitioned allocation has been performed, assuming that a matching dedicated consumption rule is used which does not overrule the original allocations. It is the purpose of this chapter to present two key model extensions. First, the impact of introducing *forecast errors* on the partitioned allocation will be investigated. Afterwards, by relaxing the assumption of the dedicated consumption, alternative consumption policies in the form of *consumption rules* for individual orders will be discussed. Since two types of such consumption rules will be presented, the discussion in this chapter has been organized into the following three sections:

1. Forecast errors are not negligible if the allocation planning procedure is performed subject to a significant lead time before demand is realized in the form of actual orders. In the previous chapter, all experiments were based on the assumption that quotas can be planned based on a perfect knowledge of demand. This is a reasonable simplification if the objective is to study differences between the many possible allocation schemes. Maintaining the assumption of a partitioned allocation and dedicated consumption, it will be investigated in Section 5.1 to what extent the different allocation schemes are sensitive to forecast errors.

2. Once the allocated quotas per customer segment are subject to forecast errors, some reservations may turn out to be too high compared to the demand of the according customer segments. But other customer segments may have to experience unnecessary stock-outs if the demand has been underestimated. This is particularly disadvantageous if highly profitable orders have to be denied although some quota reservations still exist for other, less profitable customer segments. Hence, it may be preferable to overrule the partitioned allocations when fulfilling individual orders.

   Alternative *consumption rules*[1] other than a dedicated consumption allow fulfilling an individual order by searching for and by consuming the still-available quotas at other *leaf nodes* in the customer hierarchy. Such policies have already been used successfully in models with a flat partitioning of the customer segments. An

[1] Or *consumption policies*, both terms will be used interchangeably in this thesis.

S. Vogel, *Demand Fulfillment in Multi-Stage Customer Hierarchies,* Produktion und Logistik,
DOI 10.1007/978-3-658-02864-0_5, © Springer Fachmedien Wiesbaden 2014

adjustment to the case of multi-stage customer hierarchies will be presented in Section 5.2.

3. Lastly, a second class of consumption rules will be presented by introducing the principle of *decision postponing* into multi-stage customer hierarchies. To mitigate the risk of misallocations in the presence of forecast errors, Kilger and Meyr (2008, p. 193) have suggested retaining some supply quantities at higher hierarchical levels in the form of *virtual safety stocks*. These reservations at intermediate nodes may be consumed on a first-come-first-served basis if incoming orders can no longer be fulfilled from the quota reservations. While similar approaches have already been successfully implemented in multi-echelon inventory systems, neither practical experiences nor actual design recommendations for such retention policies have yet been reported in the context of demand fulfillment. A few starting points for further investigations of this strategy will be presented in the course of Section 5.3.

5.1. Forecast Errors

A particularly strong assumption which has been made in Section 4.1 was that the allocation planning procedure can be performed with accurate information. More precisely, it has been assumed that each sales agent at each of the leaf nodes l is capable of making accurate forecasts of the demand and the unit profit in his customer segment, i.e. $\hat{d}_l = d_l$ and $\hat{p}_l = p_l$. In many practical settings, there will be a lead time between the forecasting and allocation planning step on the one hand and the consumption of the resulting quotas on the other hand. Then, it is unlikely that the sales agents can make accurate forecasts regarding the demand and unit profits. Rather, there will exist forecast errors $d_l - \hat{d}_l = \epsilon_l^d \neq 0$ and $p_l - \hat{p}_l = \epsilon_l^p \neq 0$ at each leaf node. Yet two key assumptions will be maintained in this section:

- The outcome of the allocation planning step consists of partitioned quotas (i.e. separate reservations will be made for each leaf node).
- There is still only a dedicated consumption, i.e. each order may solely consume reservations made for its associated customer segment at the corresponding leaf node.

With these assumptions still in place, but in the presence of forecast errors regarding demand and unit profits, the quotas determined in the allocation planning step will no longer exactly match the actual demand per leaf node even if an optimal allocation scheme is used (e.g. OCA or ODA) and if no shortage occurs ($SR = 0\%$). It is the purpose of this section to study to what extent the different allocation schemes are affected by forecast errors.

Note that a few attempts have already been reported in the literature which go even one step further: Some models utilize information about forecast errors to already improve the allocation step. See Talluri and van Ryzin (1999) for an early example from the revenue management literature and Quante (2009); Quante et al. (2009a) for several approaches from the demand fulfillment literature with a flat customer segmentation. Such ideas are not in scope of this thesis and may be the subject of subsequent research.

Section 5.1.1 contains a description of how forecast errors can be integrated into the simulation framework, whereas Section 5.1.2 will report the results of several numerical experiments.

5.1.1. Forecast Error Model

To accommodate forecast errors, a few adjustments of the simulation environment are required. For the experiments described in the following, it is desirable to control the magnitude of the forecast error directly. The next paragraphs will describe how this can be achieved.

The actual values for demand and unit profit per leaf node are generated as before. However, the forecast values will be obtained by adding a random error term drawn from a specified distribution to these actual values. These error terms will be drawn from a Normal distribution which has a mean of zero and a controlled standard deviation.[2]

Consider demand forecast errors at node l: While the mean value of the forecast error should equal zero, the standard deviation of the forecast error will be modeled as the product of the actual value of demand and a pre-specified factor CV_{d_l}, i.e.

$$\sigma_{d_l} = d_l \cdot CV_{d_l}. \tag{5.1}$$

Rearranging implies $CV_{d_l} = \frac{\sigma_{d_l}}{d_l}$, i.e. the standard deviation of the forecast error—the root of the mean squared error—is normalized by the actual demand. This measure of forecast accuracy has already been introduced in Table 2.3 as the CV-RMSE or simply as the CV (for a shorthand notation).

For the scope of this thesis, the simplifying assumption will be made that the forecast accuracy as measured by CV_{d_l} is identical at all leaf nodes, i.e. $CV_{d_l} = CV_{d_{l'}} = CV_d$ for all $l, l' \in \mathcal{L}$ and $l \neq l'$. This leads to

$$\sigma_{d_l} = d_l \cdot CV_d. \tag{5.2}$$

The resulting demand forecast error ϵ_{d_l} at node l is then distributed according to $\epsilon_{d_l} \sim \mathcal{N}(0, \sigma_{d_l}^2)$. For small values of d_l, or in case of high values of CV_d, this approach entails the risk that the demand forecast error term may become a large negative quantity, i.e. $-\epsilon_{d_l} > d_l$. To keep the simulation framework within a reasonable level of complexity, neither negative demands (which may result from order cancellations) nor negative profits (which may result under certain market environments) will be allowed. To prevent negative demand forecast values, such cases will be truncated and set to zero. As a result, the

[2] Such an approach has also been used in other demand fulfillment models, e.g. in Quante (2009, p. 83).

demand forecasts will be given by

$$\hat{d}_l = \max\{0; d_l + \epsilon_{d_l}\}. \tag{5.3}$$

Profit forecasts will be obtained in a corresponding manner. The standard deviation of the unit profit forecast error at node l is generally given by $\sigma_{p_l} = p_l \cdot CV_{p_l}$. Again, $CV_{p_l} = \frac{\sigma_{p_l}}{p_l}$ is a measure of the forecast accuracy with respect to the unit profits at node l. To simplify the numerical experiments, this value will also be assumed to be identical at all leaf nodes, so $CV_{p_l} = CV_{p_{l'}} = CV_p$ for all $l, l' \in \mathcal{L}$ and $l \neq l'$. Hence, the standard deviation of the unit profit forecast error which is used in the following is given by

$$\sigma_{p_l} = p_l \cdot CV_p. \tag{5.4}$$

The resulting unit profit forecast error ϵ_{p_l} at node l is then distributed according to $\epsilon_{p_l} \sim \mathcal{N}(0, \sigma_{p_l}^2)$. After truncating any possible negative values, the unit profit forecast used in the following corresponds to

$$\hat{p}_l = \max\{0; p_l + \epsilon_{p_l}\}. \tag{5.5}$$

Note that the accuracy of the demand and of the profit forecasts can be controlled independently, i.e. by adjusting either CV_d or CV_p.

This modeling approach with normally distributed error terms allows for a straightforward link to other forecast accuracy measures which are more common in practice. Managers often prefer to express forecast accuracy in terms of percentage errors (MAPE, see Section 2.2.4) rather than in terms of the CV(-RMSE). For normally distributed demand forecast errors, there is a simple relationship between the mean absolute deviation (MAD) and the standard deviation σ of the demand forecast errors (e.g. see Raju and Srinivasan (1996, p. 1460) or Nahmias (2009, p. 112)). It is given by

$$MAD = \frac{1}{\sqrt{\frac{\pi}{2}}} \cdot \sigma = \sqrt{2/\pi} \cdot \sigma \approx 0.8 \cdot \sigma. \tag{5.6}$$

Normalizing both the MAD and σ by the actual demand yields the relationship $MAPE_d \approx 0.8 \cdot CV_d$ for the demand forecast error (the same argument also holds for the unit profit forecast error measures). Hence, a CV of 0.5 roughly translates into a MAPE of 40%. Recall that typical short- and medium-term demand forecast errors at the SKU-level are in the range of 15–20%, but can become as high as 40% (see Table 2.4 in Section 2.2.4). Therefore, the numerical experiments reported in the following will use values for the coefficient of variation CV_d of 0.0, 0.1, 0.3 and 0.5 to cover a broad range of typical values for the MAPE metric found in practice. Since no sources have been found in the literature which give an indication of the typical forecast accuracy with respect to prices or margins, the same range will also be used for CV_p.

The other assumptions of the simulation framework continue to hold. For example, the sales agents are still assumed to report their forecast values truthfully to their superiors. Hence, the demand forecast aggregation in the customer hierarchy again occurs according to (3.16). No additional errors will be introduced in the aggregation process. Note that the aggregation process has a beneficial effect on the mean demand forecast error. As discussed in Section 2.2.5 in the context of hierarchical forecasting, the mean forecast error will *decrease* at higher hierarchy levels.

Profit forecasts at higher hierarchical levels will again be determined according to (3.17). The magnitude of the profit forecast error values at higher levels is determined by two effects:

1. There is an obvious *direct effect* if $CV_p > 0$, i.e. if the individual leaf node profit forecasts are inaccurate.

2. Additionally, there is an *indirect effect.* Assume that there are no profit forecast errors at the leaf nodes (i.e. if $CV_p = 0$). But if demand forecast errors exist ($CV_d > 0$), the demand-weighted aggregation of the profit values via (3.17) will introduce profit forecast errors at higher levels, because the weights in the aggregation formula will be inaccurate.

The calculation of the Theil index at higher hierarchical levels is affected by similar direct and indirect effects. First, lower-level values of T may be biased (direct effect). In addition, there are two indirect effects: Error-prone demand forecast values can enter Equation (3.67) directly. Furthermore, this calculation also involves the (aggregated) profit forecast values. As indicated above, these aggregate profit forecast values may also exhibit a forecast error even if $CV_p = 0$.

In the presence of forecast errors, a major difference compared to the experiments reported in Chapter 4 is that the OCA/ODA schemes no longer constitute the first-best benchmarks. Although both rules per se lead to ‘optimal’ quotas at the leaf nodes of the hierarchy, over- and under-allocations will result since allocation planning is based on inaccurate input data. Both the actual demand and the actual leaf node profitabilities will be different from the forecast. Hence, under the OCA/ODA schemes, excess quantities will remain as leftovers at some nodes while other nodes will suffer a shortage.

For analytical purposes, the *Perfect Central Allocation* (PCA) scheme will be introduced. PCA corresponds to an omniscient central planner who has full transparency across the customer hierarchy and also has a perfect forecasting ability. Essentially, PCA allocates the given supply according to the *actual demands and actual unit profits* at the leaf nodes (and not according to the (error-prone) demand and profit forecasts, as under the OCA scheme). The resulting quotas under PCA will therefore be optimal given the actual values of the demand volumes and profitabilities.

PCA may not be used for actual allocations in a practical context. Rather, PCA-based allocations will only be calculated on an ex-post basis to serve as a first-best benchmark. In particular, all performance assessments of the different allocation schemes will be reported in the form of ARLP with respect to PCA. Since PCA leads to identical results as OCA if no forecast errors exist, the ARLP values reported for the following numerical experiments remain comparable to those discussed in the previous chapter.

5.1.2. Numerical Experiments

The purpose of the following numerical experiments is twofold: to clarify the impact which forecast errors have on the quality of the allocation planning process and to identify differences between the key allocation planning schemes. The sequence of the experiments follows a similar logic as in the previous chapter. First, a base case setting will be presented which illustrates the performance of the different allocation schemes under different settings of the forecast accuracy. Then, changes to both the shortage rate as well as the hierarchy size will be investigated. This is accompanied by a detailed comparison between the performance of the ODA and ADA schemes and between the IDA and PA schemes.

Base Case: Impact of Demand and Profit Forecast Errors

As in Chapter 4, the base case setting again consists of a three-level hierarchy with a span of control of 4 (i.e. 16 leaf nodes and 21 nodes overall). The shortage rate will initially be fixed to a value of $SR = 20\%$ *of the actual demand* at the root note of the hierarchy. This means that the shortage rate will not be affected by the demand forecast errors and can be controlled independently. As before, random values for the demand and unit profit at the leaf nodes will be drawn uniformly from the interval $[0; 100]$ for a total of 100 input data sets.

In contrast to the previous experiments, forecasts for the demand and the profit at the leaf nodes will now be determined via (5.3) and (5.5) for each input data set. In this base case experiment, both CV_d and CV_p will be varied independently and the sixteen combinations of the values 0.0, 0.1, 0.3 and 0.5 for both coefficients of variation will be tested to determine the impact of both types of forecast errors.

The leaf node allocations will result from the application of any of the three profit-based schemes OCA, ADA and IDA or of the proportional, i.e. quantity-based scheme PA. All results will be reported in the form of the average profit loss compared to the theoretical best case scheme (ARLP metric). Recall that in all experiments involving forecast errors, the first-best allocation will be achieved via the PCA scheme and no longer by the OCA/ODA schemes.

Tables 5.1a–5.1d give the resulting ARLP values for each allocation scheme for the 16 different combinations of the forecast error settings for demand (CV_d, horizontal axis of each table) and unit profit (CV_p, vertical axis).

First, note that under all four allocation schemes (with a small exception for PA at $CV = 0.1$), the introduction of a demand forecast error leads to a larger degradation of the ARLP values than a profit forecast error of the same magnitude. As can be seen, the ARLP values in the upper right part of the results matrices are significantly larger than in

		CV_d			
		0.0	0.1	0.3	0.5
CV_p	0.0	0.0	4.0	12.3	20.7
	0.1	0.0	4.0	12.4	20.8
	0.3	0.9	4.9	13.2	21.6
	0.5	4.4	8.2	15.9	23.6

(a) ODA scheme

		CV_d			
		0.0	0.1	0.3	0.5
CV_p	0.0	0.8	4.6	12.5	20.6
	0.1	0.8	4.6	12.6	20.7
	0.3	1.7	5.4	13.3	21.3
	0.5	5.1	8.7	16.1	23.4

(b) ADA scheme

		CV_d			
		0.0	0.1	0.3	0.5
CV_p	0.0	5.0	8.7	16.3	24.3
	0.1	5.2	9.0	16.5	24.4
	0.3	6.1	9.5	17.0	24.5
	0.5	8.6	12.1	19.0	25.5

(c) IDA scheme

		CV_d			
		0.0	0.1	0.3	0.5
CV_p	0.0	15.9	15.8	18.2	23.8
	0.1	15.9	15.8	18.2	23.8
	0.3	15.9	15.8	18.2	23.8
	0.5	15.9	15.8	18.2	23.8

(d) PA scheme

Table 5.1. – ARLP (%) under demand and profit forecast errors ($SR = 20\%$, 3-level hierarchy)

the lower left part. As expected for a quantity-based scheme, there is no influence of the profit forecast error on the performance of the PA scheme. In each column of Table 5.1d, all ARLP values are identical. But even under the profit-based allocation schemes, the impact of profit forecast errors is small. As can be seen by comparing the first two rows in each table, a small profit forecast error ($CV_p = 0.1$) leads to almost the same ARLP values as in the case without any profit forecast errors ($CV_p = 0$).

Proceeding to larger values of CV_p, the impact of profit forecast errors becomes more profound if $CV_p = 0.5$. This disproportionate performance degradation relates to the fact that the (aggregated) unit profit values primarily determine the order in which successor nodes are served. While small profit forecast errors only rarely disturb the optimal sequence according to which individual nodes will be served, once the profit forecast error becomes large enough, these disturbances become more frequent.

Another major observation is that with increasing forecast errors the resulting ARLP values for the ADA scheme approach or even undercut those of the ODA scheme (esp. for $CV_d = 0.5$). Hence, there seems to be no additional benefit from using full information (ODA) or central control (OCA)[3] for the allocation decision if (demand) forecast errors are sufficiently large. Rather, the ADA scheme, which only uses aggregated information, leads to competitive or better results.

This behavior can be explained best with the help of a concrete example. For a simpler presentation, the example will be given for the case with only demand forecast errors, i.e. $CV_p = 0$. Now consider Figure 5.1 which depicts a basic example hierarchy with four leaf nodes and given forecast values for the demand and unit profit at each leaf node. Hence, the forecast for the overall demand equals 200 units. Assume that there is a shortage of

[3] Recall that ODA and OCA lead to the same allocation results.

supply of 25%. As a result, only 150 supply units are available and need to be allocated using either the ODA or the ADA scheme. Table 5.2 contains more detailed information on the mechanics of this allocation process to the leaf nodes.

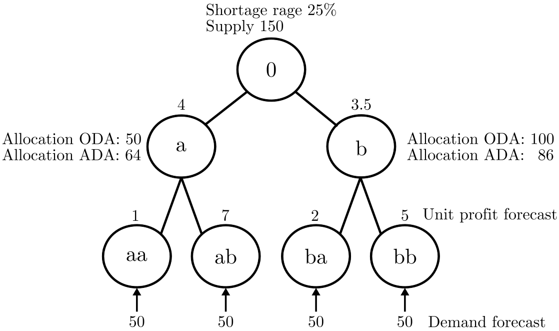


Figure 5.1. – Example hierarchy: Allocation under ODA and ADA with forecast errors

Initially, recall that the ODA scheme serves the leaf nodes strictly in order of decreasing unit profit forecasts (data row 1). In this sequence, the leaf nodes receive an allocation which equals their demand forecast until the available supply has been depleted (in many cases, the last leaf node with a positive allocation may only receive a partial allocation). In the example hierarchy, all leaf nodes except *aa* receive an allocation of 50 units under the ODA scheme. This is reflected in data row 3 of Table 5.2.

Under the ADA scheme, it is the level of customer heterogeneity at the intermediate nodes *a* and *b* which determine the size of the leaf node allocations. For the given data, the allocations to the intermediate nodes correspond to $x_a = 64$ and $x_b = 86$ units.[4] These quantities will be allocated further to the leaf nodes. In this last allocation step, the more profitable nodes will be served first. The resulting allocations per leaf node under ADA are given in line 4 of Table 5.2.

Now consider the impact of demand forecast errors. Assume that the actual demand per leaf node turns out as indicated in row 5 of Table 5.2.[5] If the demand at a leaf node has been overestimated, the allocation is likely to be too high, depending on the allocation scheme. These leftovers cannot be sold and will lead to a profit loss at that leaf node. Equivalently, there may also be a profit loss from too low allocations due to underestimation.

In the example hierarchy, the ODA scheme leads to leftover quantities at nodes *ba* and *bb* where the corresponding demand was overestimated (data row 6). Profit losses from under-allocations occur at nodes *aa* and *ab* (data row 7). Under the ADA scheme, leftovers only occur at the single node *bb* while under-allocations can be noticed at all

[4] This can be confirmed by calculating the parameters θ_a and θ_b via (3.67) and then determining the allocations according to Algorithm 2.

[5] The given values of the forecast and the actual demand imply a MAPE value of 20% if aggregated over all leaf nodes, i.e. $CV_d = 0.25$.

	Data		*aa*	*ab*	*ba*	*bb*	Total
1	Unit profit forecast		1	7	2	5	
2	Demand forecast		50	50	50	50	200
3	Allocation	ODA	0	50	50	50	150
4		ADA	14	50	36	50	150
5	Actual demand		50	70	40	40	200
6	Profit loss leftovers	ODA	0	0	20	50	70
8		ADA	0	0	0	50	50
7	Profit loss under-allocation	ODA	50	140	0	0	190
9		ADA	36	140	8	0	184
10	Total profit loss	ODA	50	140	20	50	260
11		ADA	36	140	8	50	234
12		PCA	50	0	0	0	50

Table 5.2. – Comparison between ODA and ADA under forecast errors: Leaf node data

three other leaf nodes. In the current setup of the simulation environment, which only considers a single period problem (i.e. one allocation planning step and one consumption phase), and with a dedicated consumption policy, leftovers are clearly more harmful than under-allocations in shortage situations: An under-allocation at a highly profitable leaf node only means that additional profits would have been possible had more supply been available at that node (rather than being diverted to a less profitable node). Hence, one only looses the profit differential between a more profitable node with shortage and the less profitable node which is actually served instead. By contrast, a leftover is even worse: Here, the entire unit profit of the most profitable node which still experiences a shortage is lost.

The assumption of a single period problem is certainly valid if the allocated products are perishable or constitute fashion items which will have no or a significantly reduced value in the next period. However, if leftovers do not constitute lost sales, but can be reused again in a subsequent period, a different temporal model is required, e.g. by extending the setup of Meyr (2009) to multi-stage customer hierarchies. Such an approach will also allow accounting for stock-out penalties such as backorder or order denial costs. In the one-period setting studied here, these costs of sub-optimal customer service can only be accounted for partially and indirectly via the unit profits per leaf node (see also Sections 1.2 and 3.3). Overall, in a multi-period setting with inventories, the performance difference between ADA and ODA requires a closer inspection. This constitutes an important avenue for further research.

Returning to the example calculation in Table 5.2, note that under the ADA scheme, the allocation x_b is smaller than the demand forecast. Hence, the less profitable node *ba* only receives a reduced allocation of $x_{ba} = 36$. However, as the demand at this node has been overestimated anyway, this reduced allocation only implies a small profit loss from under-allocation of 8 units (compared to the first-best PCA scheme). Instead, ODA leads to an over-allocation of 20 units.

Now consider the left intermediate node a. The higher allocation to node a allows fulfilling at least some demands at node aa under ADA. Had this quantity been allocated to intermediate node b instead (as under ODA), it would not have resulted in a profitable use. Rather, it would have been allocated to ba and led to lost profit due to an excessive allocation. The net effect is that the total profit difference to the first-best case PCA scheme is smaller under ADA than under ODA in this scenario (see data rows 10–12 in Table 5.2). Put differently, the fuzzier allocation under ADA alleviates the risk of over-allocations and of associated profit losses if forecast errors need to be accounted for. The strict adherence to the latently inaccurate forecasts under ODA often reserves too high quotas for the more profitable nodes.

The above experiment has shown that profit forecast errors have significantly smaller effects than demand forecast errors. To reduce the number of parameters in the following experiments, only the joint effect of both demand and profit forecast errors will be considered in the following experiments. More precisely, it will be assumed that the error terms for both types of forecast will have the same coefficients of variation, i.e. $CV_p = CV_d = CV$. This corresponds to the setting along the first diagonal in Tables 5.1a–5.1d.

Different Shortage Rates and Hierarchy Sizes

As an extension of the above base case experiment, the assumption of a fixed shortage rate will be relaxed. This paragraph will illustrate the joint effect of different shortage rates as well as of different forecast error settings. The same effect will be considered further in the following paragraph, focusing on a detailed comparison between the ODA and ADA schemes as well as between the PA and IDA schemes.

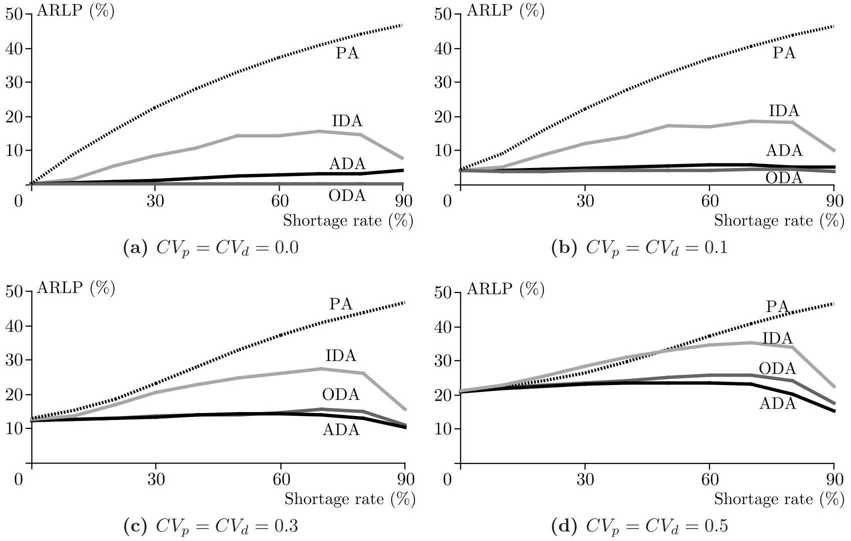
Initially, the 3-level hierarchy will again be used. While the shortage rate will be varied between 0–90%, the forecast error (for demand and unit profit) will be controlled by testing the settings $CV = 0.0$, 0.1, 0.3 and 0.5. As before, the performance of the allocation schemes ODA, ADA, IDA and PA in terms of ARLP will be compared.

The results of the experiment are depicted in Figures 5.2a–5.2d. Each graph corresponds to a different forecast error setting CV and shows the profit loss associated with each of the four allocation schemes as a function of the shortage rate.

In the following, the effect of forecast errors on each of the four allocation schemes will be discussed in more detail:

- ODA (dark gray lines): In case of no forecast errors, ODA is optimal and the corresponding ARLP curve corresponds to the horizontal axis, as shown in Figure 5.2a.[6] The introduction of forecast errors has the effect of moving the ARLP curve upward. This shift of the ARLP curve is almost constant over the entire range of shortage rates tested. Comparing the four graphs, the performance loss in terms of ARLP due to the forecast errors increases with the value of CV. For example, the setting $CV = 0.1$ (Figure 5.2b), corresponding to an average forecast error of about 8%,

[6] Note that Figure 5.2a corresponds to Figure 4.16.



(a) $CV_p = CV_d = 0.0$

(b) $CV_p = CV_d = 0.1$

(c) $CV_p = CV_d = 0.3$

(d) $CV_p = CV_d = 0.5$

Figure 5.2. – ARLP (%) per shortage rate and forecast error setting (3-level hierarchy)

leads to an ARLP of 4.2% on average. Similarly, $CV = 0.3$ (Figure 5.2c), equivalent to a forecast error of 24%, shifts the ARLP curve under ODA to on average 13.8%.

- ADA (black continuous lines): If no forecast errors are present, it has been observed in the previous chapter that the gap to an optimal scheme increases slightly with the shortage rate. At $CV = 0.1$, this increase is significantly smaller and ARLP is almost constant over the entire range of shortage rates analyzed. At high ($CV = 0.3$) and very high ($CV = 0.5$) forecast error settings, the ARLP values even become smaller again at higher levels of shortage. Furthermore, ADA leads to strictly better allocations than ODA from a particular threshold value of the shortage rate onwards.

- IDA (light gray lines): For the case without forecast error (see Section 4.5.3), it was already noted that the difference to an optimal allocation scheme increases with the shortage rate until a shortage rate of about 80%. For higher values of SR, the ARLP values decrease again. This pattern can also be observed after the introduction of forecast errors and the same explanation as before applies (see page 252). As with the ODA scheme, positive values of CV have the result of shifting the ARLP curve upwards.

- PA (black dotted lines): For this scheme, a different phenomenon can be observed. While the introduction of increasingly larger forecast errors clearly has a negative impact on ARLP for small shortage rates, there are hardly any differences in ARLP between the depicted forecast error settings for shortage rates > 40%. This can be explained as follows:

In case of no or only small shortage rates, the demand forecast errors will lead to situations where some leaf nodes will receive too high allocations. Given the dedicated consumption policy, the surplus supply units will constitute lost sales in the one-period setting studied here. Clearly, the higher the forecast error setting, the stronger is this effect. It shifts the leftmost part of the ARLP curves in Figures 5.2a–5.2d upwards. However, with rising shortage rates, these surplus allocations cease to occur, almost all allocated quantities will be sold and the total profit over all nodes will approach on average a value of $1 - SR$ also for high forecast error settings.

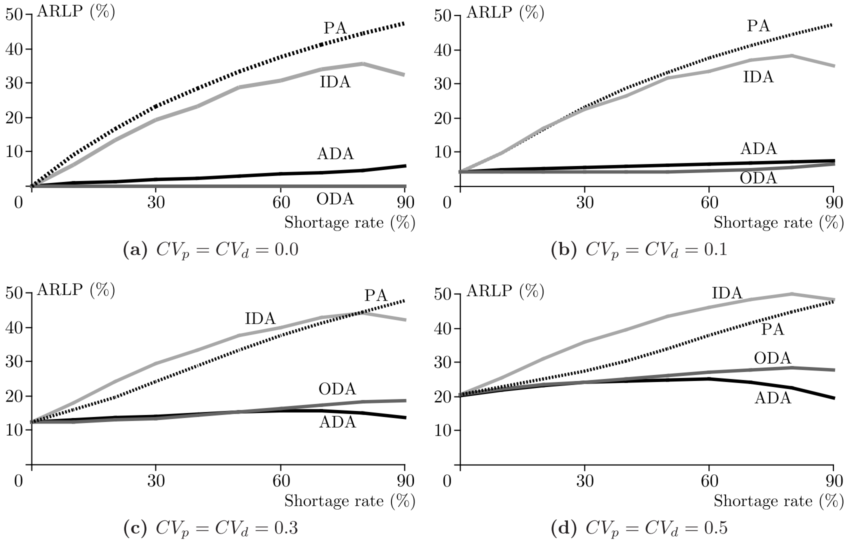
In sum, the results in Figure 5.2 indicate that the ADA scheme is the superior scheme once demand and profit uncertainty are incorporated. While ODA and ADA lead to similar ARLP values if the forecast error values remain small, ADA outperforms ODA at higher forecast error levels if the shortage rate is large enough. It appears that the organizational and technical efforts to allocate quotas based on the ODA scheme (or equivalently, based on the OCA scheme) may often not be justified if forecast errors matter. The use of ADA, which only requires aggregate data, will lead to equivalent or even better allocations.

To further illustrate the consequences which arise when introducing forecast errors, the test bed settings described above for a 3-level hierarchy have also been applied to a 5-level hierarchy. Varying again both the shortage rate and the coefficient of variation to control the forecast errors, the corresponding ARLP values in this larger hierarchy have been depicted in Figures 5.3a–5.3d for the four allocation schemes studied here.

The qualitative results are similar to those which have been observed for the 3-level hierarchy. One noticeable difference refers to the performance of the IDA scheme. It has already been observed before (see Section 4.5.4) that the gap between the IDA and the PA scheme in terms of ARLP values is small in the larger hierarchy. With forecast errors, the picture changes for the worse. This is particularly obvious in Figure 5.3d. At a forecast error setting of $CV = 0.5$, the PA scheme strictly outperforms the IDA scheme over the entire range of shortage rates tested. This is an important result for practical applications. Including aggregated profit forecasts into the allocation decision at higher hierarchy levels (without considering the heterogeneity of the corresponding sub-trees) entails a significant loss of profit. A simple proportional scheme is a far better choice in this setting.

Detailed Comparison: ODA vs. ADA and IDA vs. PA

Two aspects of the above results will be investigated further in the following. The first paragraph will focus on a direct comparison of the performance under the ADA and the ODA scheme in the presence of forecast errors. Figures 5.4a and 5.4b visualize the absolute difference of the corresponding ARLP values (in % points) for each hierarchy type, at different levels of forecast errors and over a wide range of shortage rates.



(a) $CV_p = CV_d = 0.0$

(b) $CV_p = CV_d = 0.1$

(c) $CV_p = CV_d = 0.3$

(d) $CV_p = CV_d = 0.5$

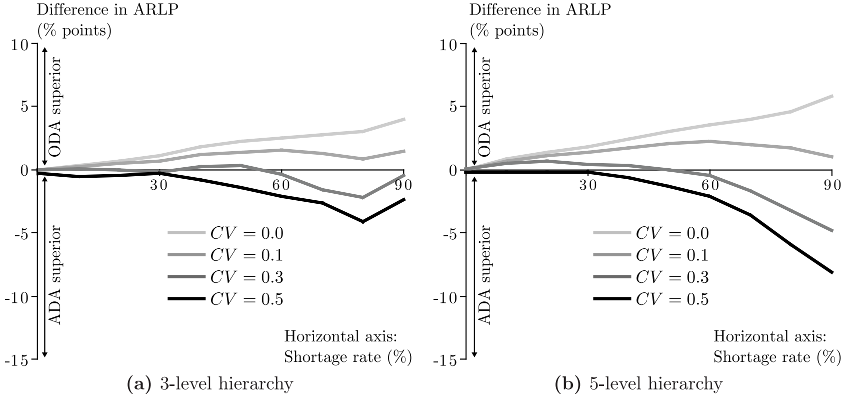
Figure 5.3. – ARLP (%) per shortage rate and forecast error setting (5-level hierarchy)

The second paragraph will address the relative performance of the PA and the IDA scheme. Correspondingly, Figures 5.5a and 5.5b depict the absolute difference of the ARLP values (in % points) between the PA and the IDA schemes, again for different levels of the shortage rate and for different settings of CV.

All four figures allow for a straightforward interpretation: For those settings of SR and CV where the curves lie in the positive half-plane, the ODA and the IDA scheme, respectively, lead to a better allocation of quotas (in terms of profit loss compared to the first-best scheme PCA). For those settings of SR and CV where the corresponding curves lie in the negative half, the ADA and PA schemes result in better allocations.

Comparison Between ADA and ODA: Focus initially on Figure 5.4a, which shows the difference in ARLP between the ODA and the ADA scheme in the 3-level hierarchy. Once particular threshold values for CV and SR have been reached, ADA clearly outperforms ODA. For $CV = 0.3$, both schemes have similar ARLP values, but for values of $SR > 60\%$, there is a clear advantage to be gained by using ADA. For $CV = 0.5$, ADA leads to better overall allocations over the entire range of shortage rates tested.

Similar conclusions can be drawn by inspecting Figure 5.4b, which illustrates the gap between the ARLP for ADA and ODA in the 5-level hierarchy. ADA is superior to ODA for a shortage rate of about 50% or larger if $CV = 0.3$. If $CV = 0.5$, ADA outperforms ODA again for all shortage rates.



(a) 3-level hierarchy **(b)** 5-level hierarchy

Figure 5.4. – Difference between ADA and ODA per shortage rate and forecast error

Comparing the graphs for both hierarchy sizes, there is one puzzling difference to be observed: While the curves in Figures 5.4a and 5.4b are very similar for small and medium levels of shortage, the advantage of ADA over ODA starts to decrease again in the 3-level hierarchy at very high levels of shortage. Interestingly, this observation cannot be made in the 5-level hierarchy.

The root cause of this phenomenon is that the performance differential between ODA and the first-best solution PCA decreases significantly in small hierarchies at high levels of shortage. The reason for this effect will be given in the following: Recall that the values of the demands at the leaf nodes are determined by randomly drawing uniformly distributed figures from the interval $[0; 100]$. Hence, the average demand per leaf node is approximately 50 units. In the 3-level hierarchy, total demand over all leaf nodes is approximately in the range of $16 \cdot 50 = 800$ units. At very high shortage rates, say 90%, only 80 supply units will be available on average. As the ODA scheme serves the leaf nodes in order of decreasing profitability, these 80 supply units will only be sufficient to fulfill the actual demand of *less than two* leaf nodes (80/50=1.6 leaf nodes) on average. Now consider the impact of the demand forecast errors.[7] One of two situations will occur at the most profitable leaf node. Assume that this most profitable leaf node is l while the second most profitable node is l':

1. If $\hat{d}_l > d_l$, the forecast was exaggerated, resulting in a surplus allocation $\hat{d}_l - d_l$ (recall that the allocation corresponds to $x_l = \hat{d}_l$ as it is the most profitable node). This surplus exceeds the customer demand and a potential profit amount of $p_l(\hat{d}_l - d_l)$ will be lost. Note that the allocation to l' will be less likely to exceed the actual

[7] Profit forecast errors have a minor influence. As stated before, their presence primarily influences the order in which the leaf nodes will be served. The following explanation will focus on demand forecast errors.

demand $d_{l'}$ because on average, only an amount of $0.6 \cdot \hat{d}_{l'}$ will be allocated to l' in the first place. Hence, under ODA, no profit will be lost at node l' *compared to the PCA scheme.*

Seen over all leaf nodes, there will be a significant difference in terms of ARLP between the ODA and PCA scheme due to the leftovers. As stated, this gap amounts to $p_l(\hat{d}_l - d_l)$.

2. If $\hat{d}_l < d_l$, actual demand was underestimated. Not all demands at the most profitable node l will be fulfilled. However, the shortfall $d_l - \hat{d}_l$ will lead to an increased allocation to the second node l'. But despite this additional quantity $d_l - \hat{d}_l$, the total allocation to node l' will on average still be less than the actual demand $d_{l'}$. Put differently, the risk of any surplus allocation at node l' remains small even in case the demand at node l has been underestimated.

   The overall loss in profits over all leaf nodes under the ODA scheme compared to the PCA scheme will therefore only amount to $(p_l - p_{l'})(d_l - \hat{d}_l)$. This loss is significantly smaller than in the first case discussed above. In particular, no leftover units will remain at any of the leaf nodes and also the difference between p_l and $p_{l'}$ is likely to be small. In sum, the resulting difference in ARLP between ODA and PCA will be rather limited.

A main insight is that both of the above two cases, i.e. leftovers and no leftovers, will occur *with equal probability* in the 3-level hierarchy at a shortage rate of 90%. This is due to the fact that only one leaf node will be affected by leftovers. Furthermore, over- and underestimation are equally likely because forecast errors have been drawn from the (symmetric) Normal distribution.

Against this background, the different performance gaps in the 3-level and in the 5-level hierarchy between ODA and PCA at a shortage rate of 90% can be explained. The main reason is that a situation without leftover quantities *at any leaf node* will occur with a significantly lower probability in the larger hierarchy. Therefore, the gap in terms of ARLP between ODA and PCA and between ODA and ADA is much larger in the 5-level hierarchy at high levels of shortage.

A numerical example will help to highlight the mechanics. Recall that the 5-level hierarchy which is used for the experiments has 256 leaf nodes. Despite an average total demand of $256 \cdot 50 = 12,800$ units across all leaf nodes, only $1,280$ units of supply will be available if the shortage rate equals 90%. On average, this supply quantity is only sufficient to satisfy the total demands of 25.6 of the 256 leaf nodes. Hence, in the 5-level hierarchy, approximately 26 nodes can be expected to receive a positive allocation at this level of the shortage rate, in contrast to only 2 nodes in the 3-level hierarchy. *At each* of these 26 nodes, demand may be both over- and underestimated. The difference to the three-level hierarchy is that *simultaneous under-forecasting* of the demand at all of these 26 leaf nodes is highly unlikely. Rather, for almost every input data set, surplus allocations from overestimation will result at least at some leaf nodes. On average, these surpluses will affect every other of the 26 leaf nodes. Now recall that the performance of

ODA will only come close to PCA if there are no leftover quantities *at any* of the leaf nodes. While this case occurs in the 3-level hierarchy with about 50% probability, it will happen almost never in the 5-level hierarchy.

Hence, the resulting gap between ODA and PCA in terms of ARLP will remain high for a shortage rate of 90%. To see the same improvement as in the 3-level hierarchy, the shortage rate needs to be increased further, until the supply is sufficient to only serve the demands of 1.6 leaf nodes. This requires a shortage rate of $SR = 1 - 1.6/256 = 99.375\%$. Only in this extreme case, there is also a 50% probability that no leftover quantities will remain at any leaf node in the 5-level hierarchy. This explains why there is an improvement of the ARLP values under ODA in the 3-level hierarchy at a shortage rate of 90%, but not in the 5-level hierarchy.

Comparison Between PA and IDA: It is similarly instructive to directly compare the performance of the IDA and of the PA schemes. Focus first on Figure 5.5a which gives the difference in terms of ARLP between both allocation schemes in the 3-level hierarchy, again for different values of CV and SR. While the IDA scheme leads to superior results at small and medium levels of forecast error, it can be observed that the simpler PA scheme actually performs better for shortage rates of up to 55% at high forecast errors ($CV = 0.5$). Put differently, there appears to be no benefit from employing a simple profit-based scheme such as IDA if forecast errors are large, unless there is a significant level of shortage in the customer hierarchy.

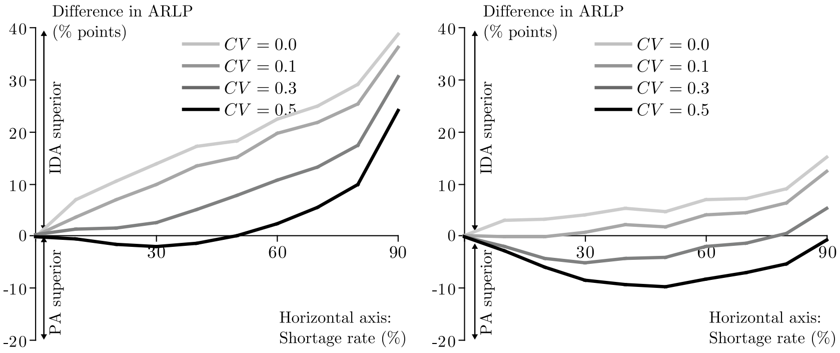
If the hierarchy is larger, the advantages of the PA scheme are more apparent. As depicted in Figure 5.5b, the PA scheme leads to a strictly lower ARLP at all levels of shortage tested if $CV = 0.5$. Even if forecasts are more accurate ($CV = 0.3$), PA is still superior to IDA at shortage rates of up to 80%. Also note that there are hardly any differences in performance between both schemes if forecast errors are small and if there is only a mild shortage ($CV = 0.1$, $SR \leq 25\%$). The superiority of PA at low levels of shortage and high forecast errors can be explained by the superposition of two effects:

- On the one hand, IDA clearly favors more profitable nodes through its ‘all-or-nothing’ property (see also Section 4.5.3). This is particularly helpful if supplies are very tight so that only a small number of leaf nodes will receive any allocation at all. By design, these are the more profitable leaf nodes while the less profitable nodes will receive no allocations at all. As was discussed earlier, this effect is diluted by more hierarchy levels and if many leaf nodes exist. By contrast, PA will serve all leaf nodes equally.

- On the other hand, however, this ‘all-or-nothing’ property of IDA may lead to significant surplus allocations in case of forecast errors (both demand and unit profit forecasts may play a role here). Moreover, on average every other leaf node *with a positive allocation* under IDA will have leftover quantities since forecast errors are normally distributed; and these nodes with a positive allocation belong to the more profitable nodes.

By contrast, leftovers due to forecast errors will occur on average at every other of *all* leaf nodes under the PA scheme if there is no shortage. But with increasing shortage rate, the number of leaf nodes with leftovers will decrease under PA.

At small and medium levels of shortage, the second effect prevails, leading to a superiority of the PA scheme. However, at high levels of shortage, the first effect will dominate, as can be inferred from Figure 5.5a and especially from Figure 5.5b.



(a) 3-level hierarchy

(b) 5-level hierarchy

Figure 5.5. – Difference between PA and IDA per shortage rate and forecast error

The above experiments indicate that forecast errors have a significant impact on the performance of the individual allocation schemes. Two results stand out:

1. The new ADA scheme has been shown to perform particularly well against all other allocation schemes under forecast errors. Once forecast errors are as high as in many practical situations ($CV \geq 0.3$, i.e. MAPE $\geq 24\%$), the ODA/OCA schemes cease to yield the best allocation results.

2. For the PA and IDA schemes which are more widely used in practice, the above experiments have shown that a basic profit-based scheme like IDA rarely offers any advantage over a simple quantity-based scheme like PA if forecast errors need to be considered. In the presence of realistic forecast errors ($CV \geq 0.3$, i.e. MAPE $\geq 24\%$ or larger) and for small levels of shortage ($SR \leq 30\%$), a simple proportional allocation leads to similar results already for small customer hierarchies.

5.2. Enhanced Quota Consumption Rules

In all previous experiments, the focus was placed on the allocation planning step and on methods to allocate scarce supply quantities to the leaf nodes in multi-stage customer hierarchies. The resulting quotas correspond to reservations for particular customer segments which were subsequently consumed by the incoming orders in a real-time process, as discussed in Section 2.4. So far, the key assumption has been that the allocation planning step leads to a *partitioned allocation* which is not overruled by the consumption policy, i.e. there is a *dedicated consumption*.

However, when deciding whether to accept an arriving customer order from a particular customer segment, companies may not only check the corresponding quota reservation. Rather, if demand for a particularly important, i.e. highly profitable customer segment has been underestimated, companies may at times prefer fulfilling such important orders by diverting quotas which have originally been reserved for other order arrivals, thus overruling the original quota allocation.

From the perspective of the simulation model used in this thesis, the key difference between a dedicated consumption and more enhanced consumption rules is that in the latter cases, each *individual order* and the *sequence of order arrivals* need to be considered. By contrast, with dedicated consumption, it is sufficient to rely on the fact that the total demand d_l at node l corresponds to the sum of the quantities requested per individual order. Once the quota has been depleted, orders in excess of x_l will be denied. Neither the size of each individual order nor the arrival of the orders at all other nodes need to be modeled explicitly to determine the total profit earned under the different allocation schemes in the setup used in Chapter 4.

However, if the consumption policy permits a consumption from an alternative, less profitable node $l' \neq l$ (as considered in this and the next section), the sequence of order arrivals (and the order size) matters. It is instructive to briefly illustrate the more important point regarding the order arrival sequence: Assume that the quotas and the actual total demand at both nodes l and l' are equal, i.e. $x_l = x_{l'}$ and $d_l = d_{l'}$, and that there is a shortage, i.e. $x_l < d_l$ and correspondingly $x_{l'} < d_{l'}$. For simplicity, consider only unit size orders. If most of the orders at the other node l' arrive later than those at node l, some of the orders arriving at node l after x_l has been depleted may still be fulfilled by 'stealing' from node l'. However, if there is a truly mixed order arrival, both quotas x_l and $x_{l'}$ will be depleted almost at the same time and no stealing will occur.

Such enhanced consumption policies which grant highly profitable orders access to quotas reserved for lower-profit orders are conceptually similar to quantity-based revenue management.[8] With the consideration of consumption policies other than a simple dedicated consumption, e.g. nested ones which are used in revenue management, the quota reservations made in the allocation planning step no longer determine the outcome of the demand fulfillment process (i.e. the maximum possible profit). Rather, such alternative consumption rules may effectively override some of the previously fixed quotas. Cleverly designed consumption rules thus have the potential to partially offset deficiencies caused by improper allocation planning or by forecast errors.

[8] See also the brief overview of revenue management-based demand fulfillment which was given in Section 2.4.4, especially Quante et al. (2009b).

The following discussion of consumption-related aspects of the DMC problem consists of two parts. Initially, in Section 5.2.1, an overview will be given concerning the different consumption rules which may be employed in customer hierarchies. In Section 5.2.2, results will be reported from numerical experiments to evaluate the promising rule types.

5.2.1. Enhanced Consumption Rules in Customer Hierarchies

As long as the allocation planning procedure does not involve a physical transportation of supply quantities to other locations, it only results in a virtual earmarking of the individual quantities for particular customer segments. A number of alternative fulfillment options can easily be implemented. *Enhanced consumption rules* establish a prioritized order of the alternative fulfillment options. Searching for fulfillment alternatives may be particularly useful for highly profitable orders. As discussed in Section 2.4.3, there are three typical search dimensions in the case of a flat partitioning of the customer segments:

Customer segment If the demand for a particular customer segment has been underestimated, consumption rules may grant access to quota reservations which have been made for other customer segments.

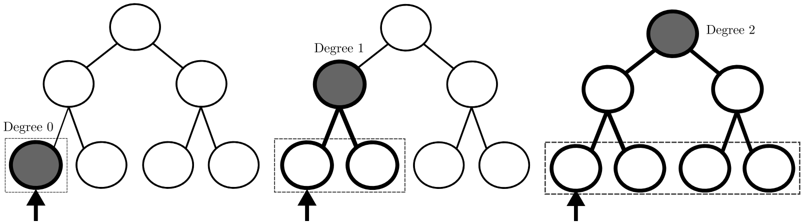
Time If quota reservations per customer segment have been made for particular consumption periods, consumption rules may permit also consuming quotas which have originally been reserved for the same customer segment for earlier or for later periods.

Substitute product If a requested product is out of stock, a company may at times choose to fulfill the order request by delivering a similar product of equal or higher value, provided the customer accepts such substitutions.

Since the focus of the models discussed in this thesis is limited to a single-period problem (consisting of one allocation planning step and one subsequent consumption period) and since only a single product is considered, the following presentation will only address the first search dimension. A key idea here is *nesting*, i.e. using a system of hierarchically linked quotas rather than strictly separated, i.e. partitioned quotas. As highlighted in Section 2.4.4, nested quotas have primarily been discussed in the context of quantity-based revenue management, but the application to demand fulfillment with a flat partitioning of the customer segments has already received much attention. The following paragraph will present an extension of the nesting concept to multi-stage customer hierarchies. Afterwards, the order arrival model will be explained.

Nesting in Customer Hierarchies

Before discussing nesting strategies in customer hierarchies, a few definitions and concepts need to be introduced. While in case of a flat partitioning of the customer segments, a central planner oversees the current quota levels and knows the sequence of the customer segments in terms of profitability, this level of data transparency only rarely exists in customer hierarchies. Instead, it is helpful to characterize the customer segments which can be nested with the help of the *degree of kinship* (DK) metric. This concept is illustrated in Figure 5.6.



(a) Degree 0 **(b)** Degree 1 **(c)** Degree 2

Figure 5.6. – Enhanced consumption rules: Different degrees of kinship

Consider Figure 5.6a first, which represents the base case without nesting ($DK = 0$), i.e. a dedicated consumption. An order $\hat{\imath}_l$ from customer segment l with a size of $q_{\hat{\imath}_l}$ may only consume the quota x_l. If $q_{\hat{\imath}_l} > x_l$, the order must be denied if a partial fulfillment is not permitted. In Figure 5.6b, the degree of kinship equals $DK = 1$. Order arrivals from a particular customer segment l may also be fulfilled by using the reservations at the immediate sibling nodes, i.e. from the nodes which have the same parent node as node l. In Figure 5.6c, $DK = 2$. An order arrival at node l may also be fulfilled by using the quotas at the nodes which have the same grand-parent node as node l. This is equivalent to a central planner.

Hence, the setting of the DK value represents the degree of data transparency in a customer hierarchy. For example, if $DK = 1$, it is the local sales manager, being responsible for several local sales territories, who can decide whether the quotas at her subordinate nodes can be nested. The necessary information for such a decision is often not available at higher hierarchical levels, e.g. at the level of corporate management.

Given a certain degree of kinship $DK = n > 0$, the following notation will be introduced. First, denote the parent node i of a particular node $k \in \mathcal{D}_i$ as $\rho(k) = i$. Then $\rho(\rho(k))$ denotes the grand-parent of node k. For simplicity, write $\rho(\rho(k)) = \rho^2(k)$ and similarly $\rho^n(k)$ for the n-th parent of node k which is positioned n hierarchy levels above node k.

Consider again the arrival of order $\hat{\imath}_l$ with size $q_{\hat{\imath}_l}$. Given a degree of kinship setting $DK = n$ all leaf nodes in the set $\mathcal{L}_i$ may be searched for available quotas to serve $\hat{\imath}_l$. Here, the set $\mathcal{L}_i$ comprises all leaf nodes in the sub-tree which descends below node i, and node i corresponds to $i = \rho^n(l)$. Assume now that there are L' leaf nodes in this set $\mathcal{L}_i = \mathcal{L}_{\rho^n(l)}$, and that these leaf nodes have have been ordered in decreasing order of profitability. The resulting order $1, \ldots, L'$ means that leaf node 1 corresponds to the most profitable customer segment while node L' contains the reservations for the least profitable segment, with associated quotas $x_{l'}$, $l' = 1, \ldots, L'$. Denote the cumulative quota at all nodes (or the capacity) by $\bar{x} = \sum_{l'=1}^{L'} x_{l'}$.

The key idea of nested quotas is that order $\hat{\imath}_l$ may not only access quota x_l, but also the cumulative quota at all other leaf nodes associated with equal or less profitability compared to node l. This means that an order will be accepted if $q_{\hat{\imath}_l}$ is less than or equal to the cumulative quota $b_l = \sum_{l'=l}^{L'} x_{l'}$. Conversely, this implies that an amount $y_{l-1} = \bar{x} - b_l = \sum_{l'=1}^{l-1} x_{l'}$ remains reserved for orders of the higher segments $1, \ldots, l-1$.[9] In the revenue management literature, b_l is referred to as the *booking limit* for segments l and lower (i.e. with higher indices $l, l+1, \ldots, L'$), while y_{l-1} is the *protection limit* of the segments $1, \ldots, l-1$ with a higher unit profit (see, e.g. Talluri and van Ryzin, 2004, pp. 28–29).

Generally, with nested quotas, an order $\hat{\imath}_l$ will be accepted if $q_{\hat{\imath}_l} \leq b_l$, or equivalently, if $q_{\hat{\imath}_l} \leq \bar{x} - y_{l-1}$. This leads to a remaining overall capacity of $\bar{x} - q_{\hat{\imath}_l}$. Two types of nesting are typically discussed in the revenue management literature (see Talluri and van Ryzin, 2004, p. 30). They can be distinguished by focusing on the protection limit y_l (i.e. the quantities reserved for orders from segments $1, \ldots, l$):

- In *standard nesting*, the protection limit for segment l will be reduced after accepting an order of segment l. This reflects an updating (i.e. decrease) of the amount of demand which is still expected to arrive for segment l.

- An alternative approach is *theft nesting* which is associated with a 'memorylessness' property (Talluri and van Ryzin, 2004, p. 31). Even after accepting an order, it will be continued to protect an amount of y_l for later arriving demand of segment l.

Theft nesting is equivalent to standard nesting if the orders arrive in successive order from the segments $L', L'-1, \ldots, 1$. However, in applications with mixed arrival orders, theft nesting protects a comparably large share of supply for the more profitable customer segments (Talluri and van Ryzin, 2004, p. 31). This *over-protection* is often undesired; hence theft nesting is less commonly used.

Furthermore, note that the use of nested quotas in the consumption phase could already be anticipated in the allocation planning step. For example, in case of theft nesting, it may be useful to reserve a larger quantity for the least profitable customer segment compared to a simple partitioned allocation (and corresponding dedicated consumption policy), knowing that all other segments may also access the quota at this lowest-profitability segment. In other words, it also appears promising to already anticipate in the allocation planning step the consumption policy which will be used afterwards. However, such an extension is beyond the scope of this thesis. Here, the focus remains limited to studying alternative consumption policies to mitigate improper allocation quotas which may result from the presence of demand and profit forecast errors in the allocation planning step. An anticipation of the consumption rule already in the allocation step may be the subject of follow-up research.

[9] Note that $y_l = \bar{x} - b_l + x_l$.

In multi-stage customer hierarchies, consumption policies based on both standard and theft nesting can be implemented with the help of search rules. In addition, it is suggested here to combine elements of both of these established types of nesting to arrive at a third strategy for nested consumption. Recall the definition of the search space as a function of the DK parameter and consider the following rules. They are applied once an order $\hat{\imath}_l$ of customer segment l with a size $q_{\hat{\imath}_l}$ arrives:

- **Standard nesting:** First, check the remaining allocation x_l at node l. If $x_l > 0$, fulfill $f_l = \min(q_{\hat{\imath}_l}; x_l)$ and reduce x_l accordingly. For the residual order volume $q_{\hat{\imath}_l} - f_l$ check the next node $l+1$ (with slightly lower profitability) and fulfill an amount of $f_{l+1} = \min(q_{\hat{\imath}_l} - f_l; x_{l+1})$. Afterwards, continue in order of decreasing node profitability, i.e. to higher node indices. If also the quota at the least profitable node L' is insufficient to fulfill the remaining residual order volume $q_{\hat{\imath}_l} - \sum_{i=l}^{L'} f_i$, this residual quantity cannot be fulfilled at all and will be lost. If the customer does not allow partial order fulfillment, the entire order will be lost. In that case, the quantities $f_l, \ldots, f_{L'}$ can be added back to the corresponding quotas.

  This strategy effectively leads to a gradual reduction of the quantities which remain protected for orders arriving at node l (and of those arriving at the nodes with a lower profitability $l+1, \ldots, L'$).

- **Theft nesting:** First, check the quota at the least profitable node L'. Afterwards, proceed by checking the reservations at the other permitted nodes in order of increasing node profitability. The last node to be checked is node l.

  Note that this strategy generally preserves the protection limit y_l at node l (unless node l ultimately also has to be searched; since the quotas at all leaf nodes with lower profitability have been insufficient to fulfill the order).

- **Combined nesting:** First, check node l for available quota, as in standard nesting. Afterwards, adopt the logic of theft nesting and search the remaining quotas in increasing order of profitability, i.e. checking the quota at the least profitable node L' second and proceeding to node $l+1$ (i.e. by checking in decreasing order of the node indices L', $L'-1$, $L'-2$, etc).

  This strategy may avoid over-protection at node l often seen with theft nesting. Furthermore, an over-ruling of the original partitioned allocation is done in a manner that any ‘stealing’ from other nodes first affects the leaf nodes with the lowest profitability, rather than in decreasing order of profitability, as seen in standard nesting.

The implications of the search rules for these nesting strategies can be illustrated with the help of a simple example.[10] Assume a particular intermediate node k at level $m = M-2$ has three successor leaf nodes $\mathcal{D}_k = \mathcal{L}_k = \{1, 2, 3\}$ at which orders arrive. All three

[10] The example is based on Klein and Steinhardt (2008, p. 133–134). Rather than booking limits, protection limits have been used and an adjustment to the case of multi-unit orders has been made.

leaf nodes $1, 2, 3$ may be searched for available quotas ($DK = 1$). Node 1 has the highest while node 3 has the lowest unit profit.

The four Tables 5.3a–5.3d illustrate the development of the quota reservations per node after the arrival of six order requests. While Table 5.3a represents the base case with dedicated consumption and no nesting, the results for standard, theft and the new combined nesting strategies are shown in Table 5.3b, 5.3c and 5.3d, respectively.

For each incoming order, the *order size q*, the *arrival node* and the *acceptance decision* are stated in the tables. Furthermore, the tables give the *quota* at each node after each order acceptance decision. Note that in Tables 5.3b and 5.3c the last six columns also state the effective nested quotas, i.e. the *booking limits*, and the corresponding *protection limits*. The booking limit takes into account the still available quantities at the other nodes which may be used to make order acceptance decisions. For example, the booking limit for orders at node 1 initially equals 15 units for all types of nesting since orders from this node may access the reservations at all nodes. Accepting the first order of size 3 at node 2 leads to a decrease of the booking limit at nodes 1 and 2 under all three nesting strategies. In case of standard and combined nesting, the order will be fulfilled by consuming the quota at node 2. This also affects the booking limit at node 1, but not the quota at node 3. In case of theft nesting, the order will be fulfilled by consuming the quota at node 3. As a result, also the booking limit at nodes 1 and 2 will be reduced. Similarly, the protection limit gives the amount of quantities still preserved for orders arriving from the same or from a higher priority segment (i.e. with a lower index).

In the base case, dedicated consumption strategy without nesting, the last two orders 5 and 6 can only be fulfilled partially, although there are still 2 units available at node 3. Standard nesting and the new combined nesting strategy allow for an efficient use of the available quantities and manage to fulfill all orders. The only difference between both strategies in the above examples can be seen in the penultimate line of Tables 5.3b and 5.3d. In both cases, the order of 5 units arriving at node 1 will primarily be served with 4 units from this node. Under standard nesting, the fifth unit of the order, which cannot be fulfilled from node 1, is being consumed from the neighboring node 2. Under the combined nesting, it will be taken from the least profitable node 3 instead.

However, larger differences are apparent between these two nesting strategies and theft nesting. While the two former nesting strategies consume quantities from other nodes only if the current quota is insufficient, theft nesting protects the quotas of a particular node (and of all higher-ranked nodes) as long as possible, as illustrated in the protection limit columns. Theft nesting guarantees a particularly high service level to the most profitable customer segments represented by node 1. As can be seen by the leftover quantity of 4 units at node 1, this strategy has led to over-protection in the example. Orders of lower profitability were lost or only satisfied partially, although sufficient overall supplies were available to fulfill all orders.

From a technical point of view, also a number of other nesting strategies besides standard, theft and the new combined nesting can be defined. For example, also those nodes may be searched which have higher unit profitabilities than the node where the original

Order				Quota		
No.	Size	Node	Decision	1	2	3
0	—	—	—	5	5	5
1	3	2	ok	5	2	5
2	1	1	ok	4	2	5
3	3	3	ok	4	2	2
4	1	2	ok	4	1	2
5	5	1	partial	0	1	2
6	2	2	partial	0	0	2

(a) Dedicated consumption, no nesting

Order				Quota			Booking limit			Protection limit		
No.	Size	Node	Decision	1	2	3	1	2	3	1	2	3
0	—	—	—	5	5	5	15	10	5	5	10	15
1	3	2	ok	5	2	5	12	7	5	5	7	12
2	1	1	ok	4	2	5	11	7	5	4	6	11
3	3	3	ok	4	2	2	8	4	2	4	6	8
4	1	2	ok	4	1	2	7	3	2	4	5	7
5	5	1	ok	0	0	2	2	2	2	0	0	2
6	2	2	ok	0	0	0	0	0	0	0	0	0

(b) Standard nesting

Order				Quota			Booking limit			Protection limit		
No.	Size	Node	Decision	1	2	3	1	2	3	1	2	3
0	—	—	—	5	5	5	15	10	5	5	10	15
1	3	2	ok	5	5	2	12	7	2	5	10	12
2	1	1	ok	5	5	1	11	6	1	5	10	11
3	3	3	partial	5	5	0	10	5	0	5	10	10
4	1	2	ok	5	4	0	9	4	0	5	9	9
5	5	1	ok	4	0	0	4	0	0	4	4	4
6	2	2	denied	4	0	0	4	0	0	4	4	4

(c) Theft nesting

Order				Quota			Booking limit			Protection limit		
No.	Size	Node	Decision	1	2	3	1	2	3	1	2	3
0	—	—	—	5	5	5	15	10	5	5	10	15
1	3	2	ok	5	2	5	12	7	5	5	7	12
2	1	1	ok	4	2	5	11	7	5	4	6	11
3	3	3	ok	4	2	2	8	4	2	4	6	8
4	1	2	ok	4	1	2	7	3	2	4	5	7
5	5	1	ok	0	1	1	2	2	1	0	1	2
6	2	2	ok	0	0	0	0	0	0	0	0	0

(d) Combined nesting

Table 5.3. – Examples of order consumption using different nesting strategies

order has arrived. Nevertheless, there is usually no rationale for most of these possible alternatives. As already indicated in Section 2.4.3, Meyr (2009) has tested a number of such alternative consumption strategies for demand fulfillment with a flat partitioning of the customer segments. Since none of the alternative strategies was found to be particularly helpful, they will be disregarded in the following.

Order Arrival Model

As argued above, the actual realization of demand needs to be modeled in more detail to determine the impact of the different enhanced consumption strategies. In the previous experiments with a dedicated consumption policy without nesting, no information about individual order arrivals was required. The analyses focused on the share of the overall demand at each node which could be satisfied under a particular allocation policy (see equation (4.56) which summarized the profit realization).

In the following, a more detailed perspective is required. After an allocation procedure has taken place, the individual order arrivals need to be considered. Order arrivals in customer hierarchies for physical goods differ in two respects from the simple assumptions used in many revenue management models for service industries:

- There is no defined order arrival sequence. When considering the stream of incoming order requests, each base customer segment is equally likely to receive the next order.
- Orders may comprise several units.

In this thesis, deterministic models are used both for the allocation planning and the consumption step. This allows taking care of orders easily which differ in terms of the *order size*. Nevertheless, a variation of the order size would introduce another parameter, further adding to the complexity of the simulation environment. Therefore, a detailed study of this aspect will be left for future research. Instead, the simplifying assumption will be made that all orders have the size of one unit.

No further modifications to the simulation environment are required to accommodate individual orders. To maintain consistency with previous experiments, the actual demand between two allocation procedures will simply be broken into individual orders of size 1. The resulting orders from all leaf nodes will then be shuffled to obtain an order stream for the hierarchy. This order stream is then characterized by a random order arrival sequence. As the model environment only considers a single cycle between two allocation procedures, neither the actual arrival of each order nor its due date need to be modeled explicitly. This situation is depicted in Figure 5.7. As a consequence, no quota reservations will be made for individual periods between two allocation instants. Hence, it is only the sequence of the order arrival which needs to be considered in the experiments. A natural question is whether the effort of allocation planning and the subsequent use of

consumption search rules is justified at all. Instead, a very simple alternative is to serve all order requests in a first-come-first-served (FCFS) manner. Under FCFS, orders consume the available supply without giving any priority to a particular customer segment. Thus,

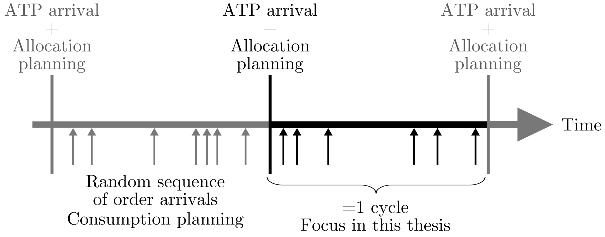

ATP arrival + Allocation planning
ATP arrival + Allocation planning
ATP arrival + Allocation planning
Time
Random sequence of order arrivals Consumption planning
=1 cycle Focus in this thesis

Figure 5.7. – Random sequence of order arrivals within one cycle between two allocation planning instances

FCFS constitutes a (worst-case) benchmark. With the extended scope of the DMC model and the inclusion of individual orders, this aspect will also be considered in the following numerical experiments.

5.2.2. Numerical Experiments

The purpose of the following three numerical experiments is to test the effectiveness of the different consumption strategies which use nesting as a means to mitigate adverse effects of quotas which have been set under the influence of forecast errors.

- In the first experiment, the impact of using consumption rules based on standard, theft and combined nesting will be evaluated for the 3-level and the 5-level hierarchies using different settings of the DK parameter. These enhanced consumption rules will be compared, on the one hand, to the base case consumption policy without nesting (dedicated consumption) and, on the other hand, to a simple FCFS order acceptance rule. Here, only two settings of the shortage rate will be considered.

- In the second experiment, a wider range of shortage rates will be considered for the most powerful combination of allocation scheme and consumption policy identified in the first experiment. Using the ADA allocation scheme and the new *combined nesting* consumption strategy, different sizes of the search space in terms of the DK setting will be compared to a simple FCFS fulfillment strategy.

- One result of this experiment is that even for the simplest consumption strategy without nesting, there appear to be particular conditions under which an allocation planning-based approach (with subsequent dedicated consumption) clearly outperforms a simple FCFS fulfillment strategy. These conditions will be evaluated further in a third experiment which includes a consideration of different levels of customer heterogeneity. Approximate worst-case threshold values for the shortage rate, forecast accuracy and the level of customer heterogeneity will be determined for the 3- and the 5-level hierarchy. Below these threshold values, an approach based on the ADA scheme in connection with a dedicated consumption leads to worse results than a simple FCFS order acceptance strategy. These values can indeed constitute worst case settings, as these thresholds will generally be lower if more enhanced consumption rules are used.

Consumption Rules Using Nested Quotas

As shown in Section 5.1, forecast errors typically found in practice may lead to a severe deterioration of the ARLP metric for the allocation schemes ODA, ADA, IDA and PA, although to a different degree. The following experimental results will show how enhanced consumption rules based on nested quotas can lead to more profitable order acceptance decisions in the presence of forecast errors. A high forecast error setting of $CV = 0.5$ will be used to generate the demand and unit profit forecast values. Two settings for the shortage rate will be tested, $SR = 20\%$ and $SR = 90\%$.

3-level Hierarchy: Initially, the focus lies on the 3-level hierarchy. ARLP values per allocation scheme for the base case with a dedicated consumption and no nested quotas have already been determined in Section 5.1 and were reported in Tables 5.1a–5.1d (lower-right entries in each of the four tables).

The main question is whether the introduction of more enhanced consumption rules can lead to significantly lower ARLP values. By varying the degree of kinship and the nesting strategy, seven different consumption policy settings can be defined for the 3-level hierarchy. Each of the policies specifies a particular sequence in which the leaf nodes of the customer hierarchy will be searched for available quota reservations to fulfill an incoming order:

1. DK = 0, i.e. no nesting at all
2. DK = 1, standard nesting
3. DK = 1, theft nesting
4. DK = 1, combined nesting
5. DK = 2, standard nesting
6. DK = 2, theft nesting
7. DK = 2, combined nesting

The first policy with degree 0 implies that quota consumption is restricted to the node where the order has arrived. This dedicated consumption was the default setting in all previous experiments. The main motivation behind using any of the other consumption policies 2–7 is to increase the number of profitable orders which will be accepted and fulfilled if the allocation planning process is subject to forecast errors.

As the PCA scheme corresponds to an ex-post perspective, forecast errors have no impact under the PCA scheme. The allocations determined by PCA already equal the optimal reservation quantities. As a consequence, orders arriving at each leaf node shall only be served up the quota level. The use of any consumption rule which grants access to reservations at other nodes is counterproductive and will lead to lower total profits. Hence, PCA in connection with consumption policy 1 constitutes the first-best benchmark against which all other settings will be evaluated.

Initially, focus on the case with $SR = 20\%$. The ARLP values per allocation scheme and per consumption strategy are shown in Figure 5.8a for standard nesting (policies 1,2,5), in Figure 5.8c for theft nesting (policies 1,3,6) and in Figure 5.8e for combined nesting (policies 1,4,7).[11] Note that the base case policy 1 (no nested consumption) is repeated in each figure to allow for an easier comparison.

In the base case setting, there are hardly any differences between the allocation schemes ODA, ADA and PA (all result in ARLP values between 22–23%). As already reported in Section 5.1, the IDA scheme seems to suffer to a larger extent from the forecast errors, leading to an ARLP value of 25%. Compare these values to the result for the simplest order acceptance strategy FCFS (dashed horizontal line). For the given values of $SR = 20\%$ for the shortage and $CV = 0.5$ of the forecast error setting, the additional efforts of using allocation and consumption planning (policy 1, i.e. without nesting) do not pay off. Simply accepting the orders based on FCFS appears to be a far superior strategy from the perspective of profit maximization, resulting in an ARLP of only 16%.

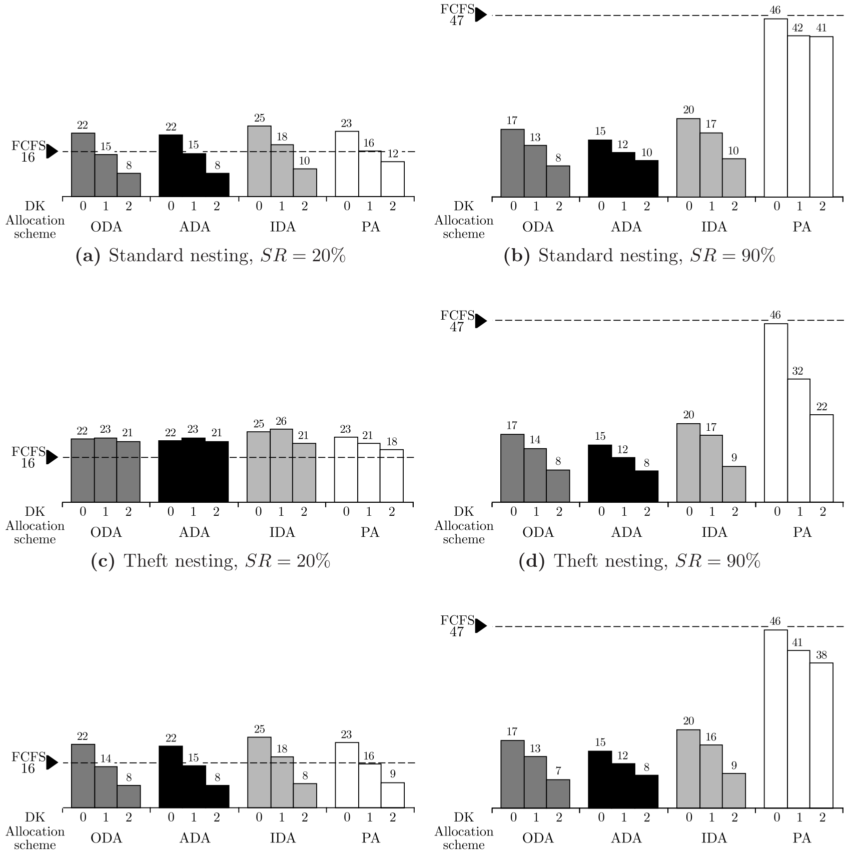
However, allowing consumption from the immediate sibling leaf nodes ($DK = 1$) with standard nesting (policy 1) leads to ARLP values for ODA, ADA and PA which are comparable to the FCFS results. Increasing the degree of kinship setting further to $DK = 2$ (policy 3), all leaf nodes of the customer hierarchy will be searched for available quotas. This requires a planner with full oversight over all leaf nodes. Should this be feasible, an additional decrease of the ARLP values can be observed for all allocation schemes, leading to ARLP values of only 8% under ODA and ADA and to values of 10% and 12% under IDA and PA, respectively.

While the use of standard nesting improves overall profits considerably, theft nesting either does not lead to any improvements at all (ADA, ODA) or only results in slightly better ARLP values (IDA, PA), at least for policy 5. Nevertheless, all theft nesting simulations lead to worse results than FCFS if the shortage rate is low with SR=20%. The third nesting strategy, combined nesting, performs best, being on par or even outperforming standard nesting. For example, using the large search space DK=2, all four allocation schemes result in ARLP values of 8–9%.

In Figures 5.8b, 5.8d and 5.8f, the above experiment has been repeated with a high shortage rate of 90%. At this high level of shortage, the use of a profit-based allocation scheme with dedicated consumption (ODA, ADA or IDA with base case consumption policy 1) leads to significantly better results than a quantity-based scheme (PA) or a simple FCFS strategy. As shown in the fourth set of data bars in all three figures,

[11] Note that for the same settings of SR and CV, there are small differences (+/− 1%-point) between the values reported here for the base case with consumption rule 1 (first data row) and those established above in Section 5.1. These are due to the stochastic nature of the input data. Recall that random input data is used for the unit profit and demand per leaf node; furthermore, the sequence in which the indivdiual orders arrive at the leaf nodes is different in each of the 100 input data sets.

the latter two approaches result in extraordinarily high ARLP values of 46% and 47%, respectively. Among the profit-based schemes, ADA is more effective than ODA (and more effective than IDA).



(a) Standard nesting, $SR = 20\%$

(b) Standard nesting, $SR = 90\%$

(c) Theft nesting, $SR = 20\%$

(d) Theft nesting, $SR = 90\%$

(e) Combined nesting, $SR = 20\%$

(f) Combined nesting, $SR = 90\%$

Figure 5.8. – ARLP (%) for different nesting strategies (3-level hierarchy, $CV = 0.5$)

As under the low shortage rate setting, the ARLP values obtained under any of the four allocation schemes can be improved further by allowing for a nested consumption. An interesting finding is that all three types of nesting have roughly the same impact under the profit-based schemes ADA, ODA and IDA. This result holds for all degrees of kinship tested and thus appears to be independent of the size of the search space. Upon close inspection, the new combined nesting is again the best strategy, leading to slightly better overall results.

The similar performance of all three nesting types can be explained with the high shortage rate present in this experiment. The high degree of supply shortage will lead to quota values of zero for most of the low-profit leaf nodes.[12] The unit profit differential among the few nodes which receive a positive allocation is comparably small, so the actual sequence in which nodes are searched only has a small impact on overall profits. Furthermore, the risk of overprotection is small as well. As a result, standard, theft and combined nesting will lead to similar results. Differences between the three nesting strategies at high shortage rates can be expected to be more significant in larger hierarchies, since more nodes will receive a positive allocation at a shortage of 90%.

It remains to briefly comment on the performance of the quantity-based PA scheme at the high level of shortage. By comparing Figures 5.8b and 5.8f against 5.8d, one can conclude that a significant improvement in terms of ARLP can be achieved by allowing theft nesting rather than standard or combined nesting. Recall from above that one of the differences between these three types of nesting lies in the node which is checked first for an available quota. While under standard and combined nesting, this first node is the node at which the order arrives, it is the least profitable node under theft nesting.

The superiority of theft nesting for the PA scheme in Figure 5.8d can be explained as follows: At the very high level of shortage considered here, *all* leaf nodes will receive roughly 10% of the actual demand under the PA scheme. This is the main difference to the profit-based schemes where only a small share of the leaf nodes will receive a positive allocation at high levels of shortage. Assume a highly profitable order arrives first. Under theft nesting, it will be served by consuming the less profitable quotas at the other nodes, leaving the reservations at the original node unchanged for later-arriving orders. Some of the less-profitable orders which arrive early will also be served, but on average, theft nesting will protect more quantities for the highly profitable orders than standard or combined nesting if the shortage is severe.

5-level Hierarchy: The above experiment has also been conducted for a 5-level hierarchy to better study the impact of a gradually increased search space for available quotas from $DK = 0$ (no nested consumption) to $DK = 4$ (all leaf nodes may be searched). Figure 5.9 shows the results in terms of ARLP values for each allocation scheme, nesting strategy and the two shortage rate settings ($CV = 0.5$, as before). Figures 5.9a, 5.9c and 5.9e depict the case with $SR = 20\%$ for standard, theft and combined nesting. The results for the higher shortage rate $SR = 90\%$ are represented in Figures 5.9b, 5.9d and 5.9f. The dashed horizontal lines again give the ARLP values which are realized under a simple FCFS order acceptance scheme.

The results for the 5-level hierarchy are similar to those obtained for the 3-level hierarchy: Standard nesting or combined nesting involving only the immediate sibling nodes

[12] Recall from the discussion in Section 5.1.2 that the ODA scheme will lead to positive allocations at no more than two nodes at this level of shortage.

($DK = 1$) leads to similar ARLP values as a simple FCFS approach in case of the lower shortage rate. This holds if either the ADA, ODA or PA scheme is used in the allocation planning step. If even higher values for DK are chosen, the ARLP values improve further, with similar results for these three allocation schemes. As already established before, IDA performs worse than ODA, ADA and PA in the larger hierarchy, although noticeable improvements can be realized for this scheme by increasing the search space in the order consumption, especially for $DK = 4$. Nevertheless, the conceptually much simpler PA scheme leads to better results than IDA for all values of DK if there is only a mild level of shortage. For all four allocation schemes in scope of this experiment and for all five settings of DK tested, theft nesting did not turn out to be an attractive alternative. The resulting ARLP values amounted to at least 19%, i.e. were higher than under FCFS.

Turn now to the results obtained under a high level of shortage in the 5-level hierarchy (Tables 5.9b, 5.9d and 5.9f). Here, the superiority of the ADA scheme is apparent for all types of nesting. ADA outperforms all other schemes, usually by a large margin, except for very large search spaces (if $DK > 2$, ODA is slightly better). IDA leads to worse allocations than PA unless the value of DK is sufficiently high. However, consumption rules with a larger search space can lead to significant improvements for IDA whereas PA hardly benefits from a larger search space.

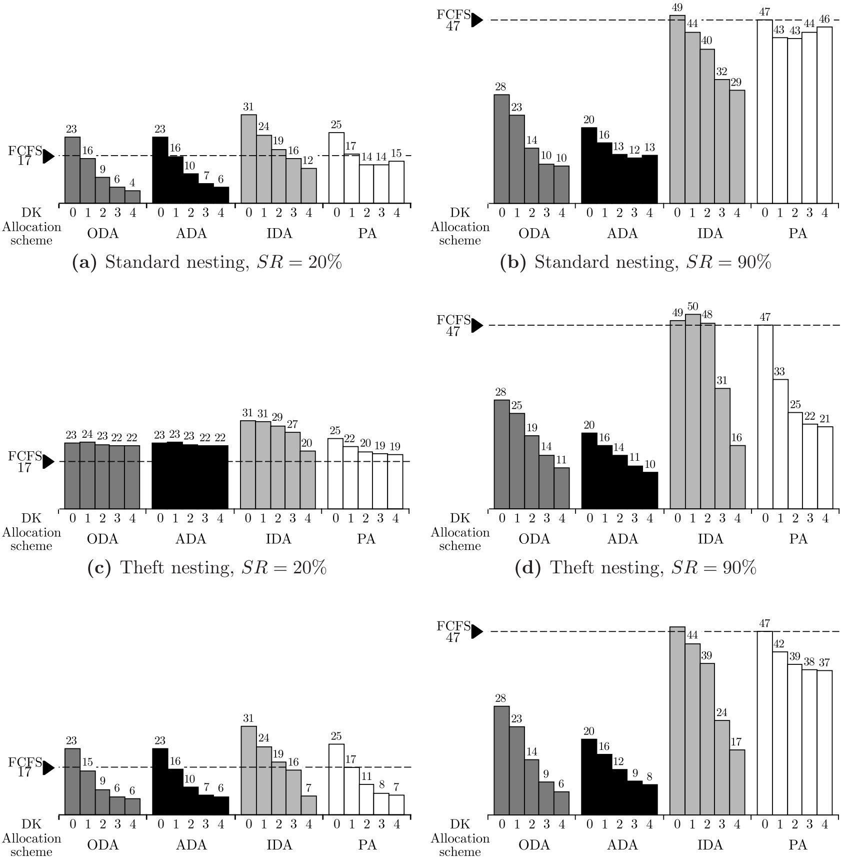
In a similar manner as in the 3-level hierarchy, combined nesting performs best from an overall perspective at the high shortage rate $SR = 90\%$, followed by standard nesting. As before, the only exception is the PA scheme where theft nesting strongly outperforms standard and combined nesting. Nevertheless, the use of a good profit-based allocation scheme such as ADA will always lead to lower ARLP values than any combination of PA and theft nesting.

Overall, the introduction of nested consumption is an effective means to make more profitable order acceptance decisions in a customer hierarchy. The larger the search space, the higher the improvement in terms of ARLP. Yet, realizing a consumption policy in practice with a high setting of DK requires a significant level of data transparency within the customer hierarchy.

The above experiments have indicated that at a high level of the shortage rate ($SR = 90\%$), a demand fulfillment approach based on a prior allocation planning step using the ADA scheme with dedicated consumption of the quotas (i.e. no nested consumption) already outperforms a simple FCFS scheme for the order acceptance decision. At a lower level of the shortage rate ($SR = 20\%$), an improvement over FCFS requires the use of enhanced consumption rules which permit to utilize quotas from other nodes, possibly at a large distance as measured by the degree of kinship metric. In the following paragraph, this relationship will be investigated in more detail for a wider range of shortage rates. Afterwards, also the impact of the level of customer heterogeneity will be explored.

Search Space for Nested Quotas

The following experiment will focus on the ADA scheme. It has been the best-performing allocation scheme in the previous two experiments if the simulations were run at levels



(a) Standard nesting, $SR = 20\%$

(b) Standard nesting, $SR = 90\%$

(c) Theft nesting, $SR = 20\%$

(d) Theft nesting, $SR = 90\%$

(e) Combined nesting, $SR = 20\%$

(f) Combined nesting, $SR = 90\%$

Figure 5.9. – ARLP (%) for different nesting strategies (5-level hierarchy, $CV = 0.5$)

of forecast error which are typically found in practice. Furthermore, in the previous experiment, a consumption planning approach based on the combined nesting strategy has been shown to lead to higher profits than both standard and theft nesting if used in connection with ADA. The combination of both choices constitutes a promising overall demand fulfillment strategy which warrants a more detailed investigation, both for the 3-level and the 5-level hierarchy. The focus in the following will be on these two aspects:

- Different settings of the search space in the consumption planning in terms of the DK parameter (using the range $DK = 0, 1, 2$ for the 3-level and testing the range $DK = 0, \ldots, 4$ for the 5-level hierarchy) and
- a wider range of the shortage rate setting (0–90%).

The combination of ADA-based allocation and consumption planning with combined nesting will be evaluated against a simple FCFS strategy for two different settings for the forecast errors ($CV = 0.3$ and $CV = 0.5$). It is the objective of the following experiment to find settings for the consumption strategy to be used in connection with ADA-based allocation which will lead to an order acceptance performance which is superior to FCFS.

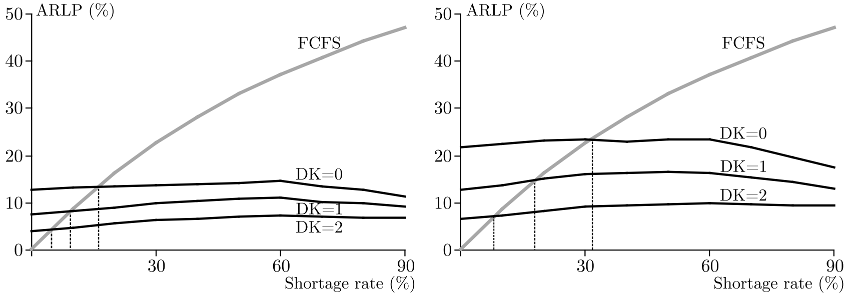
A graphical representation of the resulting ARLP values in the 3-level hierarchy can be found in Figure 5.10a for $CV = 0.3$ and in Figure 5.10b for $CV = 0.5$. For the 5-level hierarchy, corresponding graphs are given in Figures 5.11a and 5.11b. In each chart, the FCFS benchmark is represented by the gray curve whereas the combinations of the ADA scheme with the different consumption rule settings are depicted in black.

For each setting of CV and DK, there is a particular threshold value of the shortage rate: Below the threshold, FCFS leads to higher profits. This threshold value is marked with black dashed vertical lines. For example, in the 3-level hierarchy, if no nested consumption is used (DK=0), the threshold value roughly corresponds to a shortage of 18% for the lower forecast error setting ($CV = 0.3$). If the forecast error setting equals $CV = 0.5$, the threshold lies at a shortage rate of slightly above 30%. Similar observations can be made for the 5-level hierarchy. When increasing the search space in terms of DK, the additional efforts of allocation planning and nested consumption already pay off at lower levels of the shortage rate. However, as can be seen especially in Figures 5.11a and 5.11b for the 5-level hierarchy, the marginal effect of a higher DK setting is decreasing. Increasing the search space from $DK = 2$ to $DK = 3$ only leads to a modest improvement of the threshold value, and searching the entire customer hierarchy at a setting of $DK = 4$ barely leads to any further improvement.

The above experiment has shown that it may at times be necessary to use comprehensive consumption rules and a large search space to achieve similar levels of performance like a simple FCFS strategy. However, in customer hierarchies with limited information transparency, consumption rules with settings of $DK > 1$ may often not be feasible. For managers, this raises the question under which worst-case environmental conditions an allocation planning-based approach is clearly superior to a simple FCFS approach. This will be addressed next.

Different Levels of Heterogeneity

The previous experiments reported in this chapter have analyzed the impact of two important dimensions which influence the performance of allocation planning-based demand fulfillment strategies, i.e. the level of the shortage rate and the forecast accuracy (or equivalently: forecast error). The higher the values along any dimension (or both), the more



(a) $CV = 0.3$ **(b)** $CV = 0.5$

Figure 5.10. – ARLP (%): Allocation planning-based demand fulfillment (ADA with consumption planning based on combined nesting) vs. FCFS, for different shortage rates (3-level hierarchy)

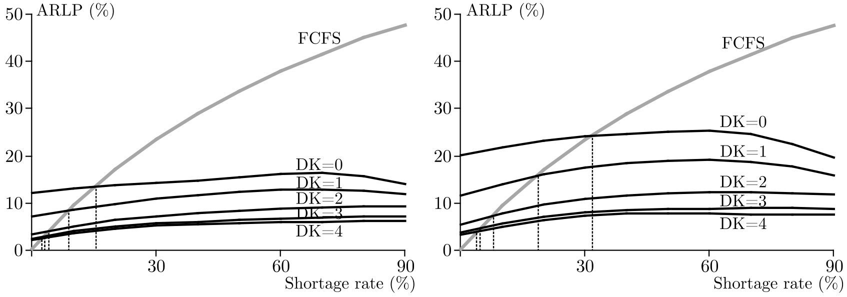
profitable is an allocation planning-based demand fulfillment approach in comparison to simple FCFS.

It is the purpose of this section to use the simulation framework to illustrate that also the level of customer heterogeneity plays an important role. As with the last experiment, the focus will be on identifying worst-case conditions under which it is clearly preferable to incur the additional efforts of using allocation and consumption planning with decentralized decisions, rather than employing a simple FCFS order acceptance strategy. Therefore, the following setup will be used:

- Allocation planning will be performed using the new ADA scheme. Other schemes like OCA or ODA may often not be available in practice, and IDA and PA have not performed well against FCFS, particularly at high levels of shortage. Recall that the information about potential forecast errors is not yet exploited in this allocation planning step. This is in line with the objective of primarily identifying worst-case conditions under which ADA-based allocation is superior to FCFS.

- Consumption planning relies on a simple dedicated consumption without nesting ($DK = 0$), in line with the assumption that a partitioned allocation is performed. The reason for this choice is again the focus on identifying worst-case conditions under which an allocation planning-based approach remains superior to FCFS. The last experiments have shown that enhanced consumption rules (especially combined and standard nesting) will result in the ADA-based approach being superior to FCFS already at lower levels of shortage and also in the presence of less accurate forecasts.

Overall, all combinations of the shortage rate SR, forecast accuracy CV and customer heterogeneity T have been determined for which FCFS and the ADA scheme with dedicated consumption lead to similar ARLP values. This search was conducted both for



(a) $CV = 0.3$

(b) $CV = 0.5$

Figure 5.11. – ARLP (%): Allocation planning-based demand fulfillment (ADA with consumption planning based on combined nesting) vs. FCFS, for different shortage rates (5-level hierarchy)

the 3-level and the 5-level hierarchy. The results are depicted in Figures 5.12a and 5.12b. Each of these graphs has been derived as follows:

- Four batches, each consisting of 100 different input data sets, have been generated by varying the setting for the unit profit interval from which the unit profit values are drawn randomly. In line with the simulations reported in Section 4.5.5, the four different interval settings correspond to a relative range of $RR_p = 200\%$, 100%, 50% and 25%. This approach leads to different levels of average customer heterogeneity per batch.[13] These averages have been denoted by $\bar{T}$, and the corresponding coefficient of variation of the level of customer heterogeneity per batch is CV_T. Note that the four values $\bar{T}$ are almost identical for the 3- and the 5-level hierarchy, but that the dispersion of the individual T values as measured by CV_T per batch is much wider for the smaller hierarchy.

- The 100 individual input data sets of each batch have been applied to the corresponding customer hierarchy (3- and 5-level, respectively) under different settings of the shortage rate and the forecast error parameter. While the shortage rate has been varied between 0 and 60% with a step size of 5%, the forecast error as measured by CV(-RMSE) was varied between 0 and 0.75 at a step size of 0.025. For each combination of input data set, SR and CV, the total profit values under the ADA scheme (with dedicated consumption and $DK = 0$) and under a simple FCFS[14] strategy have been determined.

[13] Recall from the description of the test environment in Section 4.5.1 that the level of customer heterogeneity per input data set is a random value. It is based on the random draws for the demand and unit profit at each leaf node of the customer hierarchy. As a result, T cannot be controlled directly, and its values tend to be normally distributed with a rather small variance. More precisely, T is a weighted sum of multiple uniformly distributed random variables. Hence, T is approximately normally distributed due to the Central Limit Theorem (e.g. see Bertsekas and Tsitsiklis, 2008, p. 274).

[14] The variation of CV was not necessary for FCFS as this scheme does not require any forecasts.

- Afterwards, the average profit over the 100 input data sets per batch has been calculated for the different combinations of the shortage rate setting and forecast error level, both for the ADA and the FCFS scheme.[15] Then, for each of the (discrete) values of the shortage rate, the (discrete) forecast error level has been identified at which the squared difference between the average profits per batch under ADA and FCFS was minimal. These pairs of the shortage rate and forecast error level have been used to draw one curve per batch in Figures 5.12a and 5.12b.

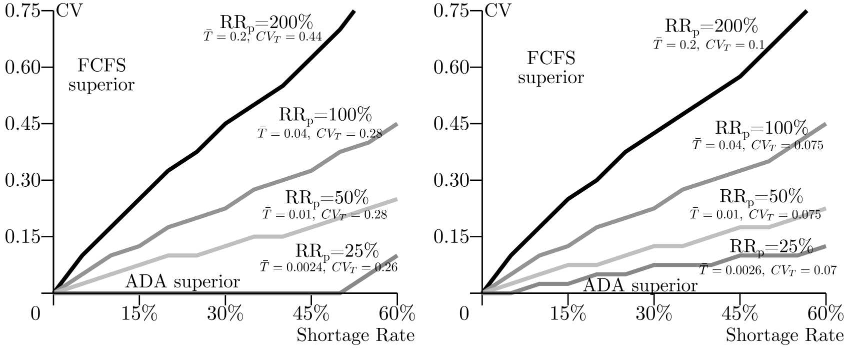
Each curve represents a different level of average customer heterogeneity $\bar{T}$ and marks the values of SR and CV at which both ADA and FCFS result in identical total profits on average. To the left of each curve, FCFS leads to superior results while to the right ADA performs better. For example, in case of a heterogeneous hierarchy with $\bar{T} = 0.2$ (black bold line) and a shortage rate of 30%, the allocation planning-based approach is superior to FCFS if the forecast error is roughly below a value of 0.45 as measured by the CV. If the forecast error is higher, FCFS performs better. This holds for both the 3- and the 5-level hierarchy. In case of less heterogenous hierarchies, e.g. with $\bar{T} = 0.04$, at the same level of shortage rate of 30%, the critical value for the forecast error is slightly above a CV value of 0.2. By comparing Figures 5.12a and 5.12b for rather homogenous hierarchies with $\bar{T} \approx 0.0025$ (lowest gray line), one difference becomes apparent between the two hierarchy sizes considered in this study: The critical value for the forecast error is higher for the 5-level hierarchy than for the 3-level hierarchy at all levels of shortage below 60%. In case of the 3-level hierarchy, the lowest gray line lies at the horizontal axis for almost all values of the shortage rate, meaning that FCFS is superior once forecasts are no longer exact. In case of the 5-level hierarchy, the allocation-based ADA approach already performs better in these difficult situations with rather smaller shortage rates, almost homogeneous hierarchies and few forecast errors. Hence, the more hierarchy levels, the more allocation decisions are required which are then less dependent on very accuarte forecasts.

The representation in Figures 5.12a and 5.12b is a different way to state a key result established before: If forecasts are comparably accurate, ADA outperforms FCFS already for very low values of the shortage rate. Two new aspects are highlighted in these graphs: The influence of the level of customer heterogeneity and the size of the customer hierarchy.

Nevertheless, it is important to keep in mind that the relationships depicted in Figure 5.12 only represent worst case situations. Still, these values can give helpful guidelines to practitioners to assess whether profit-based allocation planning approaches are worthwhile for a company. As discussed before, the performance of allocation planning-based schemes such as ADA can be improved further:

- **Allocation planning step:** The information about potential forecast errors may already be exploited when determining quotas. Furthermore, the consumption strategy can be anticipated, for example if nested quotas are used and if yes, which type of nesting is employed.

[15] Rather than considering average profits, also ARLP values could have been calculated. As the latter is only a normalized version of the former, the same overall results would have emerged.



(a) 3-level hierarchy

(b) 5-level hierarchy

Figure 5.12. – FCFS vs. ADA: Shortage rate, forecast error and heterogeneity threshold values for different hierarchy sizes

- **Consumption planning:** More enhanced consumption rules which rely on nested quotas can lead to significantly lower ARLP values for given settings of the forecast error and shortage rate. In particular, higher settings of the DK parameter representing a larger search space will allow for better results of ADA-based approaches, provided there is a sufficient level of information transparency in the customer hierarchy.

  Furthermore, the type of search rule used within the nested quotas has been shown to have a significant impact on the resulting ARLP values. An influencing factor here is the distribution of order arrivals per customer segment. If arrivals are truly mixed, the new combined nesting strategy was superior to theft nesting and even superior to standard nesting in the simulation experiments studied above.

Using one or several of these ideas will lead to lower threshold values along the dimensions shortage rate, forecast accuracy or customer heterogeneity at which allocation planning-based approaches have clear advantages over FCFS. However, for particularly low values along these three dimensions, it is likely that a simple FCFS strategy may remain better. Therefore, it may be beneficial to complement the allocation and consumption planning demand fulfillment strategy with some FCFS behavior. A rather simple extension of the existing simulation framework is to retain some quota reservations at higher hierarchy levels in the form of virtual safety stocks which may be consumed on a FCFS basis. This strategy will be investigated in the following section.

5.3. Quota Retention and Virtual Safety Stocks

Serving orders on a FCFS basis has been shown to be a simple yet highly beneficial approach for making order acceptance decisions in the presence of forecast errors if there is only a mild level of shortage. Once the shortage rate becomes larger, an approach based on a prior allocation planning step appears to be a better strategy. However, both strategies may be combined in multi-stage customer hierarchies. Acknowledging that the quota reservations at the leaf nodes are subject to forecast errors, Kilger and Meyr (2008, p. 193) have suggested retaining a certain share of the available supply at intermediate levels of the customer hierarchy in the form of virtual safety stocks (VSS). Once the quota reservations at the leaf nodes have been depleted, the VSS at the higher hierarchical levels may be consumed in a FCFS manner. This decision postponement bears strong similarity with the idea of postponement with respect to place, as introduced in Section 2.1.4. The term VSS refers to the concept of centralized safety stocks in multi-level inventory systems. Their 'virtual' nature stems from the fact that the retention in a customer hierarchy does not involve any physical inventory movements. Only changes in the reservations will be made.

Given this close relationship with the inventory literature, Section 5.3.1 will provide a brief summary of selected literature contributions which discuss allocation procedures for centralized safety stocks in inventory systems. Afterwards, it will be shown in Section 5.3.2 how the VSS concept can be included in the simulation environment for the multi-stage customer hierarchies. A few experimental results will be presented in Section 5.3.3.

5.3.1. Stock Retention in Multi-Echelon Inventory Systems

In the inventory management literature, the positioning of safety stocks in multi-echelon inventory systems has received much attention. In multi-echelon inventory systems, a central warehouse is used to supply a number of local warehouses (see also Figure 4.3). These local warehouses are often equivalent to retailers in practice, they are positioned at the second stage of the system. The central warehouse is replenished by an outside supplier, usually subject to a lead time. The primary role of the central warehouse is to forward any external replenishments to the retailers. The resulting divergent system has several advantages. For example, it allows exploiting quantity discounts at the supplier. Furthermore, the lead time within the system is usually small compared to the lead time of the outside supplier.[16]

The demand at each retailer between the arrival of two successive replenishments from the central warehouse (one cycle) follows a stochastic process. Some retailers will experience higher cycle demand than predicted and will run short on inventory before the next external replenishment arrives while others may still have plenty of stock. One option to resolve such inventory imbalances within each cycle is to conduct transshipments between the retailers (e.g. see Herer et al. (2002) or Tiacci and Saetta (2011)). An alternative is to invest in additional safety stocks at the retailers. For stationary, stochastic demands, an additional safety stock at each retailer inventory leads to a higher service level. However,

[16] Therefore, a centralized ordering process can hedge against demand fluctuations at the retailers *during the external lead time* (this effect is often termed "statistical economies of scale", see Eppen and Schrage (1981)).

as the internal lead times are comparably small, such safety stocks can also be held at the central warehouse. It is a well-known result in inventory theory that centralization of stocks reduces the amount of total inventory required to provide a particular service level (e.g. see the classical reference Maister (1976)).

This is the typical understanding of safety stocks which implies that several units of stock are held *in addition* to the mean cycle demand. For such settings, many contributions in the inventory literature address two main questions: How much total safety stock should be held and where should these stocks be placed? The answers to these questions depend strongly on the modeling assumptions. Overviews of the comprehensive literature in this area can be found, e.g., in Inderfurth (1994) or in Graves and Willems (2003).

A part of the literature addresses the case where forecasts regarding the demand at the retailers until the next external replenishment are highly unreliable, but improve over time. In this situation, it may be useful to initially retain a certain amount of each external replenishment quantity at the central warehouse. The remainder of the external replenishment quantity will be immediately forwarded to the retailers to raise their stock levels. Following the concept of decision postponing, the retained central stock may be allocated to the retailers at a later point in time. Once more accurate forecasts regarding the demand at each retailer are available, and before the next external replenishment arrives at the central warehouse, the retained quantities may be used to resupply any retailer which has experienced higher than expected demand.

Several models have been discussed in the literature addressing this stock retention and allocation problem. The typical setup can be summarized as follows (e.g. see Cao and Silver, 2005): The central warehouse forecasts the aggregate demand at all retailers and regularly places orders at the uncapacitated outside supplier. While the bulk of the arriving supplies is forwarded immediately to the retailers upon arrival, a certain amount of stock remains at the central facility. It will be used in between two outside replenishments cycles to re-balance the local inventories. There are w time periods between any two external replenishments at the central warehouse. For analytical tractability, the retailers are usually assumed to be identical, i.e. the demand distribution at each retailer has the same mean and the same variance. Most models for this problem setting aim at answering one or several of the following key questions:

- Which quantity should be delivered directly to each retailer?
- Which amount of stock should be retained centrally?
- At what point in time before the next replenishment should the central stock be allocated?
- How should the central stock be split among the retailers?

One of the earliest contributions was provided by Jackson (1988).[17] In his model, internal replenishments are shipped to all retailers at the end of each time period $t =$

[17] A working paper version of the model has been circulated at least since 1983.

$1, \ldots, w-1$ until the centrally held stock has been depleted. The last period in which these internal replenishments are sufficient to bring all retailer inventories to their maximum level is referred to as the *pooled-risk period*. In the subsequent time period, usually not all retailers can be replenished fully. In Jackson's model, the remaining central quantities will be allocated among all retailers with lower-than-average inventory levels. In other words, the objective of this last allocation is to ensure that the lowest inventory level among all retailers after the final allocation of central stock will be as high as possible. Overall, this scheme has been shown to lead to lower system-wide backorder costs than a procedure without any central safety stock.

If there is a cost for the internal replenishment, separate shipments at the end of each time period $t = 1, \ldots, w-1$ are expensive. It is preferable to only re-supply the local retailers once before the arrival of the next external replenishment order. Jönsson and Silver (1986, 1987) fix this single internal replenishment instant to the end of period $w-2$. This means that two periods remain before the next external replenishment will arrive. The authors argued that shortages in earlier periods of the cycle can be expected to be small and that the number of expected backorders will be highest towards the end of each cycle. Hence, they focused on maximizing the service rate in periods $w-1$ and w. Fixing the replenishment instant allows for a number of analytical simplifications. Assuming identical, normally distributed demands at each retailer, Jönsson and Silver showed an efficient way to determine the amount of stock to retain at the central warehouse for a given overall supply from the external supplier. This policy has been found to be helpful in a practical application at the Swiss pharmaceuticals and chemicals company Ciba Geigy, today Novartis (see Fincke and Vaessen, 1988).

This setting with only a single reallocation instant has also been analyzed by a number of other authors. Erkip (1984) attempted determining both the optimal amount of central safety stock to retain and the optimal allocation instant. His results are computationally expensive and he unfortunately did not provide any numerical examples. McGavin et al. (1993) derived a '50–25' heuristic for this problem. The allocation instant of the central safety stock was set to the mid-point of the cycle length and the amount of central safety stock was set to 25% of the mean cycle demand. Cao and Silver (2005) presented a heuristic for the case when the retailers are replenished directly by the outside supplier and only the retained stock is kept at the central warehouse.

While the contributions mentioned so far supported centrally held stocks, Schwarz et al. (1985) and Badinelli and Schwarz (1988) arrived at a contradictory conclusion. They found little beneficial effects of centrally retained stocks and argued that the average on-hand depot inventory should be close to zero. Zipkin (1980) sided with this view and showed that the risk-pooling benefit of the central stock decreases with positively correlated demands at the retailers. This different result has been explained with the fact that these latter models employ a FCFS rule for the internal replenishment of the retailers from the central safety stock (see McGavin et al., 1993). As a result, inventory imbalances among the retailers will occur once the central safety stock has been depleted. By contrast, the models by Jönsson and Silver (1987) and Jackson (1988) conducted an *allocation over all retailers* to bring all local inventories to comparable levels.

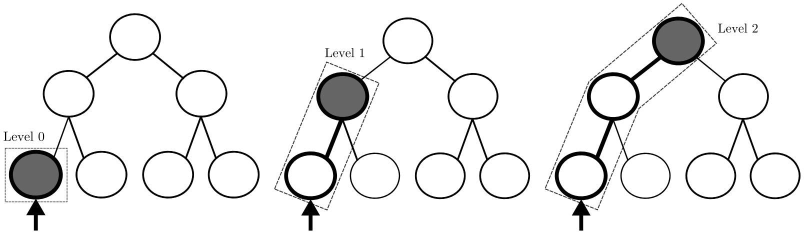
In sum, the idea of VSS in customer hierarchies bears more similarity with the concept of the *retained central stock* than with actual *safety stock planning*. The main analogy between the former and VSS refers to the time horizon of the decisions. While safety stock investment choices are usually made on a mid-term basis and aim at preventing stock-out situations at all retailers in the first place (see Section 2.1), stock retention and allocation decisions are more of an operational nature. They are made in the short-term and aim at using the scarce supplies as efficiently as possible. Therefore, the application of this stock retention concept to the DMC problem appears promising. Some initial ideas will be presented in the next section.

5.3.2. Virtual Safety Stocks in Multi-Stage Customer Hierarchies

In the course of this section, an extension to the simulation framework will be suggested to allow testing the impact of VSS on the order acceptance decision in multi-stage customer hierarchies.

First, realize that VSS can be held at each hierarchical level $0, \ldots, M-2$ in a customer hierarchy, except for the leaf node level $m = M-1$. The variable *parent level* (PL) will be used to denote the maximum number of hierarchy levels above a leaf node which may be searched for retained VSS. PL has a similar interpretation as the DK variable which has been introduced in Section 5.2 to characterize consumption rules for nested quotas.

Consider the three cases represented in Figure 5.13 for the 3-level hierarchy. Figure 5.13a with $PL = 0$ corresponds to the base case where no quota is retained at a higher hierarchical level (dedicated consumption). In Figure 5.13b with $PL = 1$, a quota will be retained at the immediate parent node. This quota can be accessed by all leaf nodes positioned in the sub-tree below this parent node, but not from the leaf nodes in the right subtree. Lastly, in Figure 5.13c with $PL = 2$, a quota will be retained both at the intermediate nodes and at the top node. Note that the quota at the root node is accessible to all leaf nodes in the hierarchy.



(a) $PL = 0$ (Base case) **(b)** $PL = 1$ **(c)** $PL = 2$

Figure 5.13. – Virtual safety stocks: Retaining supplies at higher hierarchy levels

In the experiments presented in the next section, the focus will be limited to illustrating the impact of particular quota retention settings, rather than determining the optimal VSS policy for a particular setting. In addition to PL, a second parameter *retention share* (RS) will be introduced. Provided a certain non-leaf node i (i.e. intermediate or root node) will hold VSS—this is specified by the setting PL—, an amount $RS \cdot x_i$ is retained at node i while the amount $(1 - RS) \cdot x_i$ is allocated to the successor nodes in $\mathcal{D}_i$.

In a hierarchy with $PL = 1$ and $RS = 10\%$, a ten percent quota will be retained only at the first intermediate node above the leaf nodes. By contrast, consider a 3-level hierarchy with $PL = 2$ and $RS = 10\%$. Assume that the PA scheme is used for the allocation. Ten percent of the overall supply, i.e. $(1 - SR) \cdot d_0 \cdot RS = (1 - SR) \cdot d_0 \cdot 10\%$, will be retained at the root node. An amount of $x_a \cdot RS$ will be held back at the intermediate node a. The allocation at each leaf node l corresponds to $(1 - RS) \cdot (1 - RS) \cdot (1 - SR) \cdot d_l = 90\% \cdot 90\% \cdot (1 - SR) \cdot d_l = 81\% \cdot (1 - SR) \cdot d_l$. If other allocation schemes different to PA are used, the reduction of the leaf node allocations will not be uniform among the leaf nodes. However, the same amount of overall VSS will be held.

Once a certain amount $RS \cdot x_a$ of the available supply at an intermediate node a shall be held back as VSS, two types of policies can be distinguished:

Direct Retention (DR) The VSS is retained directly from the entire available supply x_a. The remaining amount of $(1 - RS) \cdot x_a$ is allocated and each of the successor nodes x_l receives a share in accordance with the chosen allocation scheme. If a profit-based allocation scheme is used, it is primarily the less profitable successor nodes which will experience a significantly reduced allocation in comparison to a situation without retention.

Proportional Retention (PR) First, for each successor node $l \in \mathcal{D}_a$, the regular allocation quantities x_l are determined. However, only a reduced quantity $x'_l = (1 - RS) \cdot x_l$ is actually allocated to each node l, the remaining quantities $RS \cdot x_l$ are held back as VSS. As a result, all successor nodes will experience an equal proportional reduction.

Both policies are equivalent in case the quantity-based PA scheme is used for the allocation.[18] Under the profit-based schemes and direct retention, the least profitable successor nodes will contribute most to the VSS stock. The most profitable nodes will usually receive about the same allocation as without VSS. As all successor nodes can access the VSS on a FCFS basis, the direct retention policy will effectively reserve higher quotas for the more profitable customer segments.

[18] This holds for all levels of shortage if the overall supply quantity S corresponds to a certain fraction of the *demand forecast*. However, if the shortage rate is related to the *actual demand* (as in the simulations reported in this thesis), situations may arise in the simulation environment where there appears to be an oversupply if the actual shortage rate is very low or zero. Such an apparent oversupply during the allocation planning step occurs if the actual demand in the entire hierarchy has been underestimated ($\hat{d}_0 < d_0$).

In other words: the available supply $(1 - SR) \cdot d_0$ may be larger than the demand expectation $\hat{d}_0$ during the allocation planning step. In this situation, the DR policy with $PL = 2$ reserves a higher VSS quantity at the root node of the 3-level hierarchy than the PR policy. Since there is in fact no oversupply (or even a mild level of shortage), more quantities will be consumed on an FCFS basis under DR than under PR, leading to a slight performance disadvantage for the DR scheme. This effect is only small and vanishes with higher shortage rates.

The motivation behind the alternative proportional retention is to avoid potential overprotection of the most profitable customer segments. In case the demand forecast for the most profitable segments has been overestimated, surplus allocations would normally remain as leftovers at these nodes if no VSS are held. By truncating the allocations at all successor nodes by a common percentage, too high reservations may be prevented. The retained VSS quantity is available for consumption on a FCFS basis by all successor nodes. This way, less profitable nodes can be fulfilled by consuming quantities which would otherwise have been lost as surplus allocations at the most profitable nodes. However, if the demand at these most profitable nodes has been underestimated, the proportional retention policy option will lead to a lower average profit than without retention.

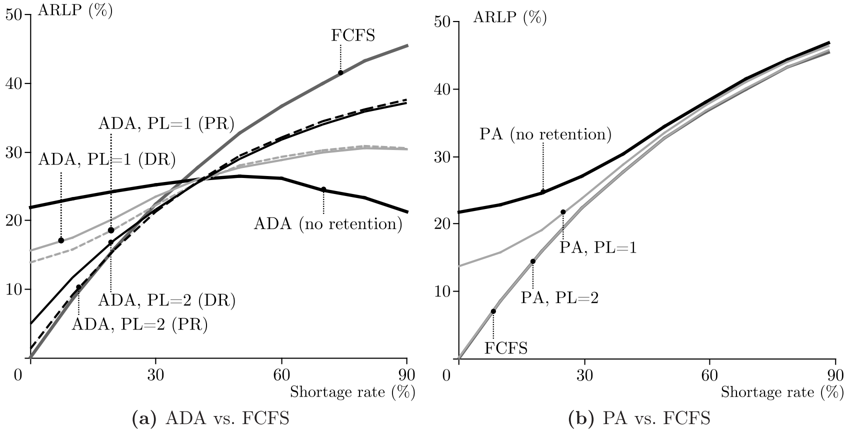
5.3.3. Numerical Experiments

The impact of placing VSS will be illustrated with the help of four simple numerical examples for the 3-level hierarchy and a forecast error setting of $CV = 0.5$:

- First, the impact of different parent level settings will be investigated. In addition, also the impact of the two types of retention settings will be simulated. Essentially, it will be investigated to what extent VSS policies based on different settings of the PL parameter and the retention setting (DR or PR) can lead to an improvement over plain ADA- and PA-based allocation with dedicated consumption, i.e. without VSS.
- Then, two different strategies for the positioning of the VSS in the customer hierarchy will be analyzed.
- Using the best overall VSS policy settings as established in the previous experiments, the impact of different sizes of the VSS volume will be simulated.
- The last experiment will briefly illustrate the effect of a comprehensive consumption policy in which both a nested quota (using the combined nesting search rule) and VSS are employed in combination.

Different Parent Level Settings: The main focus of the following simulations lies on the new ADA scheme, as it has already been shown to be the best allocation scheme in the presence of forecast errors. Figure 5.14a represents the ARLP of the ADA scheme in combination with different VSS policies over a wide range of the shortage rate (0–90%). Two benchmarks are worth considering. First, the bold black line represents the base case, i.e. the ADA scheme with dedicated consumption (and thus without any VSS). Second, the bold dark gray curve corresponds to a simple FCFS order acceptance strategy. As already discussed in Section 5.2, FCFS leads to higher profits than this plain ADA strategy (without VSS or nested quota consumption) as long as the shortage rate is below $\approx 32\%$ in the 3-level hierarchy.

The two light gray lines correspond to the two types of VSS strategies (DR or PR) for the settings $PL = 1$ and $RS = 30\%$, i.e. 30% of the allocation to each intermediate node is retained for FCFS consumption by the respective leaf nodes. As can be seen, the PR policy (dashed line) leads to lower ARLP values than the DR policy (continuous line) for lower levels of the shortage rate up to about 45%.



(a) ADA vs. FCFS

(b) PA vs. FCFS

Figure 5.14. – ARLP (%): Virtual safety stocks at different hierarchy levels (3-level hierarchy, $CV = 0.5$)

Focus now on the relative performance with respect to the two benchmark strategies. Both VSS strategies with $PL = 1$ lead to an improvement over FCFS once the shortage rate becomes larger than about 32%. In comparison to a plain ADA-based strategy, the introduction of VSS leads to lower ARLP values under both retention policies (DF and PR) up to a shortage rate of about 45%.

Furthermore, Figure 5.14a also contains the simulation results for VSS policies with $PL = 2$, i.e. quantities are retained both at the root node and at the intermediate nodes at level 1. The total VSS stock in the customer hierarchy amounts to $1-(1-0.3)(1-0.3) \approx 50\%$. As before, the continuous line represents the DR strategy while the dashed line corresponds to the PR strategy. Again, the PR strategy is superior for shortage rates below approximately 45%. At higher shortage levels, the use of VSS is generally not justified, as the plain ADA scheme results in significantly lower ARLP values.

Overall, the retention of VSS will introduce the advantages and disadvantages of FCFS consumption into the customer hierarchy. For low shortage rates, the use of VSS appears to ensure a more efficient use of the available quantities by reducing overprotection at the more profitable leaf nodes. Once higher shortage rates need to be coped with, many of the retained quantities will be used to fulfill less profitable orders, resulting in an overall performance deterioration. The more quantities are retained at higher nodes in the hierarchy, especially at the root node, the more the resulting ARLP curve approaches that of the FCFS strategy.

This behavior becomes more obvious if the PA scheme is considered instead of ADA. Using the same settings as before, the corresponding ARLP curves are represented in Figure 5.14b. As discussed above, both the DR and the PR strategy are roughly equivalent under the PA scheme. Therefore, only two curves for the VSS schemes with $PL = 1$ and $PL = 2$ are represented in addition to the plain PA-based and FCFS schemes. This graph shows clearly how larger values for the PL setting bring the ARLP values closer to those obtained under a plain FCFS strategy. The curves for PA with $PL = 2$ and FCFS are practically indistinguishable.

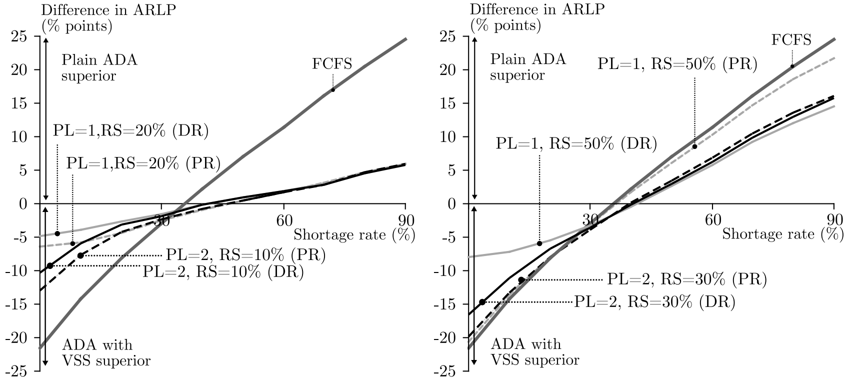
Positioning of the VSS: In this second experiment, different safety stock positioning strategies will be investigated. Two different cases will be considered, with a low (20%) and a high (50%) total VSS volume. Both will be compared to a plain ADA strategy with dedicated consumption and without VSS. As before, a 3-level hierarchy will be used, the forecast error setting remains at $CV = 0.5$ and the shortage rate will be varied between 0–90%.

In the 3-level hierarchy, there are two options regarding the positioning of the VSS. First, VSS can only be held at the intermediate nodes, corresponding to the setting $PL = 1$ with either $RS = 20\%$ or $RS = 50\%$. This will also be referred to as *single-level VSS*. Second, VSS can be held both at the root node and at the intermediate nodes in the form of a *distributed VSS*. With $PL = 2$, a setting of $RS = 10\%$ implies that only an average amount of $(1 - 0.1)(1 - 0.1) = 81\%$ of the original allocation remains at each leaf node. The total VSS at the top two hierarchy levels thus comprises almost 20% of the overall supply. Similarly, in the case $PL = 2$, a setting $RS = 30\%$ implies a total VSS volume of $1 - (1 - 0.3)(1 - 0.3) \approx 50\%$. This permits direct comparisons between different positioning strategies using the same amount of VSS.

The results of the numerical experiments are depicted in Figure 5.15. They have been calculated in the following manner: Each curve represents the difference in percentage points between the ARLP values for a particular VSS policy and the plain ADA strategy per shortage rate. If this difference is positive, a plain ADA strategy leads to lower ARLP values whereas a negative value implies that the use of that VSS policy is more beneficial. In addition, also the difference between the ARLP values of a FCFS strategy and a plain ADA-based strategy has been depicted for comparisons. This type of presentation based on the percentage point differences between different policies has been used before in Figures 5.4 and 5.5.

Figure 5.15a corresponds to the case with the low overall VSS stock of 20% of the overall supply whereas the total VSS stock in Figure 5.15b amounts to 50%. Each figure contains four curves: Two light gray lines for the single-level VSS and two black lines for the distributed VSS. As in Figure 5.14, the dashed lines indicate a PR policy whereas the continuous lines stand for a DR policy.

Figures 5.15a and 5.15b convey similar messages: All eight different VSS policies (four per figure) lead to a better performance than the plain ADA scheme for low shortage rates. The reverse holds for high shortage rates. Generally, the higher the total amount of VSS stock, the larger the improvement over a plain ADA policy at low shortage rates.



(a) Total VSS stock: 20%

(b) Total VSS stock: 50%

Figure 5.15. – Different positioning strategies for virtual safety stocks (3-level hierarchy, $CV = 0.5$, ADA scheme)

But high VSS stocks lead to a severe deterioration of performance at the high range of the shortage rate. Two further conclusions can be drawn:

- PR is a better strategy than DR: For the low total VSS volume (20%), the two dashed lines lie below the continuous lines for low and medium levels of shortage. At high levels of shortage, hardly any performance differences can be noted. For the high total VSS volume (50%), essentially the same observation can be made regarding the distributed VSS policy ($PL = 2$). In case of the single-level policy ($PL = 1$) with RS=50%, the DR strategy performs significantly better at very high levels of shortage. But it has already been concluded that the use of VSS is not recommended in these cases with high shortage levels, hence PR is the preferred overall retention policy.

- The distributed policy appears to be advantageous: For the low total VSS volume (20%), the performance of both types of policies for the distributed VSS (see the continuous and dashed black lines for the DR and PR policies) is clearly better than the performance of the two single-level policies (continuous and dashed light gray lines) at low levels of the shortage rate <20%. At higher levels of shortage, both positioning policies lead to a similar results.

  For the large size of the VSS (50%), this only holds if the DR policy is employed. In case of the PR policy, both the single-level and the distributed VSS lead to very similar results for shortage rates of up to 40%. As before, the use of VSS is not recommended at higher levels of shortage as neither VSS strategy will lead to an improvement over a plain ADA strategy.

Size of the VSS: While the previous simulation has already shown the impact of two different sizes of the overall VSS stock, this parameter will now be varied over a wider range. Again, the 3-level hierarchy will be used, allocation is performed using the ADA scheme, consumption does not involve nesting (DK=0), but VSS are employed. As a consequence of the previous simulation results, only the proportional retention strategy will be analyzed, as this setting has been found to be superior to DR for lower levels of shortage. Furthermore, only the case $PL = 2$ with a distributed VSS will be considered, i.e. VSS is held both at the nodes of the intermediate level as well as at the root node. To cover a wide range of the overall VSS size, the parameter RS will be varied with settings of 2.5%, 10%, 20%, 30% and 50%. This means that the size of the overall VSS ranges between 5 and ≈ 75%.

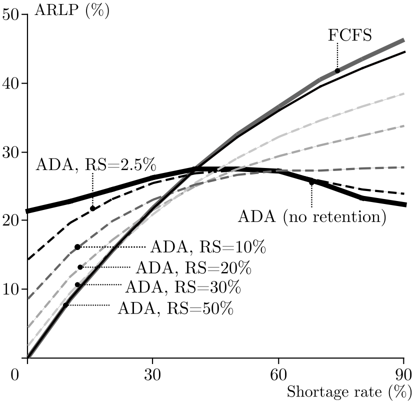


Figure 5.16. – Different sizes of the VSS as measured by the RS parameter(3-level hierarchy, $CV = 0.5$, ADA scheme, proportional retention, $PL = 2$)

Figure 5.16 shows the corresponding ARLP values for a simple FCFS order acceptance scheme (continuous dark gray line) and different ADA-based demand fulfillment strategies (thin dashed and continuous lines). Note that the curves for FCFS, for the basic ADA-based allocation without any VSS (no retention) as well as the case with $RS = 30\%$ have already been shown in Figure 5.14a.

Generally, the higher RS, the more units will be served in a FCFS fashion. At a level of $RS = 30\%$, about 50% of the overall supply is served in a FCFS manner due to the $PL = 2$ setting. The resulting ARLP curve is already very close to that of a simple FCFS scheme, with very low ARLP values at low levels of shortage, but there is a significant performance deterioration if shortages are severe. At $RS = 50\%$, the corresponding ARLP curve (thin continuous black line) is almost indistinguishable from that depicting the plain FCFS strategy, except for very high levels of shortage.

Overall, these simulation results confirm the intuition. A higher volume of VSS leads to a performance which is closer to FCFS. Yet, the above simulation does not yet answer the question regarding the best level of VSS to retain, it only illustrates the implications of different settings. While this certainly requires further research, the following guideline can be inferred from the above results: Retaining small amounts of VSS appears useful if the shortage rate fluctuates significantly. At low levels of shortage, a small amount of VSS (e.g. $RS = 2.5\%$ together with $PL = 2$) will lead to better ARLP values than a plain ADA-based allocation with dedicated consumption and no VSS. Still, the allocation planning-based approach has its merits if from time to time the shortage rate reaches higher levels.[19] Maintaining this small VSS volume also in situations with higher shortages only leads to a small decrease in performance. But if situations with significant shortage rates can be anticipated, it is advisable to discontinue reserving VSS in the allocation planning step.

Combining Nested Quota Consumption and VSS: To conclude this introduction into VSS, joint policies will be illustrated briefly which involve consumption planning based on a combination of nested quotas and VSS. The setup of this simulation is as follows: Again, the 3-level hierarchy is used, the forecast error setting remains at $CV = 0.5$ and the shortage rate will again be varied between 0–90%. The objective is to find a set of consumption rules for an ADA-based allocation which leads to low ARLP values for all levels of shortage. First, only results for 'pure' policies will be reported for an easier comparison. More precisely, the following cases have been considered and their corresponding ARLP values have been depicted in Figure 5.17a:

- FCFS: As before, a FCFS order acceptance can serve as a simple benchmark (bold, dark gray line)

- Base case: A demand fulfillment strategy consisting of a plain ADA-based allocation and a corresponding dedicated consumption constitutes the default case (bold black line, $DK = 0$, $PL = 0$).

- Nested quotas: The previous experiments in Section 5.2 have shown that the ADA scheme performs well in combination with the combined nesting strategy. Simulations with only nested quotas have been run for the two different search space settings with $DK = 1$ and $DK = 2$ (thin black lines). Recall that especially the last setting requires a high level of data transparency within the customer hierarchy which may not always be given in practice.

- VSS: Last, ADA-based allocation can also be combined with VSS instead of nested quotas. To simplify the analysis, the size of the overall VSS will be limited to 20% of the overall supply. Hence, it can either be placed only at the intermediate level ($PL = 1$, $RS = 20\%$, i.e. single-level policy) or also at the parent level ($PL = 2$, $RS = 10\%$, i.e. distributed policy). These two settings correspond to the continuous and the dashed light gray lines, respectively, in Figure 5.17a.

[19] If the shortage rate is always low and does not fluctuate, a simple FCFS strategy may be preferred if the worst case threshold values for ADA-based allocation given in Section 5.2.2 and Figure 5.12 are not surpassed.

The comparison of the simulations shows that the use of nested quotas leads to more consistent reductions of the ARLP values over the entire range of the shortage rate tested than the VSS-based strategies. As already discussed before, the distributed VSS policy leads to improved ARLP values especially for low values of the shortage rate ($< 20\%$), compared to retaining VSS only at the intermediate level. With this setting of $PL = 2$, the VSS-based consumption rule even leads to superior results for shortage rates lower than 10%, compared to a consumption rule with nested quotas and $DK = 1$. Had an even higher VSS size been used (compare the previous simulation), a further improvement would have been possible, but only for small shortage rates $< 10\%$. From an overall perspective, a plain FCFS order acceptance policy performs best for these very low levels of shortage. But considering the performance over the entire range of shortage rates, a consumption policy based on nested quotas with $DK = 2$ is the most competitive one which can be used combination with an ADA-based allocation.

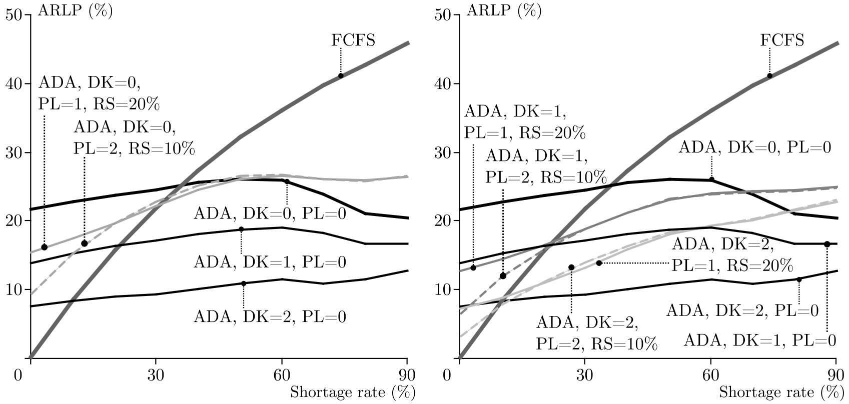
It will now be analyzed to what extent a combination of nested quotas and VSS can constitute an attractive alternative consumption policy. This joint policy will handle individual orders in the following manner: Upon arrival of an order, initially the permitted nested quotas will be checked according to the combined nesting strategy. If the order cannot be fulfilled, the VSS will be checked next. In case of $PL = 2$, the VSS check begins at the parent level of the leaf node receiving the order. The VSS at the top level will be checked last. The order will be lost if no supply quantity was found during this search.[20]

In Figure 5.17b, the ARLP values for the following demand fulfillment strategies have been depicted for the shortage rates in the range 0–90%:

- FCFS: This remains the simplest benchmark (bold dark gray line).
- Base case (for comparison): ADA-based allocation with corresponding dedicated consumption ($DK = 0$, $PL = 0$, bold black line).
- Nested quotas (for comparison): ADA-based allocation and nested consumption with $DK = 1$ and $DK = 2$ and no VSS (thin black lines).
- VSS and nesting over the immediate sibling nodes: ADA-based allocation, nested consumption with $DK = 1$ as well as VSS with a total size of 20%. The VSS can either be placed only at the intermediate level ($PL = 1$, $RS = 20\%$, continuous dark gray line) or at both the intermediate and the top level ($PL = 2$, $RS = 10\%$, dashed dark gray line).
- VSS and nesting over all leaf nodes: ADA-based allocation, nested consumption with $DK = 2$ and VSS. Again, either single-level or distributed VSS can be used, indicated by the continuous and dashed light gray lines, respectively.

[20] Recall that the order size is assumed to equal one unit in all simulation experiments reported in this chapter.

Observe first that also in case of these joint consumption policies (nested quotas plus VSS), the distributed VSS ($PL = 2$, $RS = 10\%$) generally performs better than a single-level VSS with the same overall amount of VSS, as illustrated by the dashed gray lines in comparison to the continuous gray lines. Nevertheless, the more important observations from Figure 5.17b relate to the performance of these joint policies: On the one hand, the combination of VSS with a consumption policy based on nested quotas leads to a further improvement of the ARLP values at lower levels of the shortage rate compared to a consumption policy which uses nesting, but no VSS. On the other hand, however, such a joint policy also leads to higher ARLP values at medium and higher levels of shortage. More precisely, in the case of $DK = 1$, the best joint policy (with distributed VSS) leads to an improvement over a policy which only uses nested quotas up to shortage rates of about 20%. If $DK = 2$, the best joint policy is only superior up to a shortage rate of about 10%. From a shortage rate of 60% onwards, such a joint policy performs even worse than consumption planning based only on nested quotas with $DK = 1$. Furthermore, note that there seems to be no real advantage from combining a local VSS (with $PL = 1$ and $RS = 20\%$) with a nested consumption policy which only searches the immediate sibling nodes ($DK = 1$). This can be seen by comparing the middle black line with the thin dark grey line in Figure 5.17b. The additional VSS only leads to a minor improvement up to a shortage rate of 20%, and the performance under this joint policy degrades significantly at higher levels of shortage.



(a) Nested quotas and VSS separately **(b)** Nested quotas and joint policies with VSS

Figure 5.17. – ARLP (in %) for consumption policies based on nested quotas, based on VSS and joint policies (3-level hierarchy, $CV = 0.5$, ADA scheme)

To summarize, retaining some quantities at higher hierarchy levels is a conceptually simple strategy to improve the performance of allocation planning-based demand fulfillment in multi-stage customer hierarchies for low shortage rates. For such settings, VSS can at best lead to a similar performance as achieved under a plain FCFS scheme. Nevertheless, this has to be traded off against a severe performance degradation at higher levels of shortage. Overall, the following conclusions can be drawn from the simulations reported above:

- A proportional retention has been found to lead to better results than a direct retention. Under the former policy, all successor nodes contribute in equal relative terms to the size of the VSS. By contrast, the latter policy implies that the VSS is formed primarily by reducing the allocation to the least profitable successor nodes.

- Furthermore, for a given overall size of the VSS, slightly better results have been noticed if the VSS is retained not only at the intermediate level, but also at the top level of the 3-level customer hierarchy studied here, i.e. a distributed VSS performs better than a single-level policy. Further experiments may analyze such positioning policies also in larger hierarchies.

- The appropriate size of the VSS appears to be dependent on the typical range of the shortage rate, but this question certainly warrants further investigation.

Conceptually, VSS can also be used in combination with consumption rules based on nested quotas. Such joint consumption policies will result in a superposition of the effects observed when using either nested quotas or VSS, i.e. a further performance improvement for lower levels of shortage at the cost of worse ARLP values at medium and higher levels of shortage. Overall, the results reported above only constitute some first simulation results, leaving sufficient room for further analyses of the VSS concept in multi-stage customer hierarchies.