2. Supply Chain Planning and Demand Fulfillment

This chapter summarizes the background of the DMC problem. Moving from rather general aspects to more specific topics, Section 2.1 will start with an overview of **supply chain planning** (SCP). This section will be used to introduce the concept of **hierarchical planning** and to characterize the key relations between the individual SCP tasks.

Since the discussion in this thesis is limited to MTS environments, all production processes are driven by forecasts and final items are sold from stock. In such MTS environments, the interrelations between three major SCP tasks need to be characterized to illustrate the background of the DMC problem:

- In Section 2.2, an overview of key aspects of **demand planning** will be presented; in particular, **hierarchical forecasting** will be addressed.
- Then, in Section 2.3, **master planning** will be discussed. Based on the inputs provided by demand planning, master planning determines a mid-term forecast-driven aggregate plan for procurement, production, distribution and sales.
- **Demand fulfillment**, i.e. the order-driven processes in an MTS supply chain will be covered in Section 2.4. In particular, an overview will be provided illustrating how current demand fulfillment systems handle customer heterogeneity in MTS environments, and a comprehensive review of the existing literature contributions will be given.

Finally, Section 2.5 offers a brief summary and concluding remarks. Overall, the analysis will confirm that the main issues associated with the DMC problem have not yet been addressed thoroughly in the literature, preparing the ground for the contributions in the subsequent chapters.

2.1. Supply Chain Planning

The following sections will provide an overview of planning concepts in a supply chain. As a starting point, the objectives and tasks of *supply chain management* will be introduced in Section 2.1.1, allowing for a further characterization of the DMC problem as an intra-organizational channel coordination problem. An appropriate planning concept, both for inter-company and intra-organizational supply chains, is *hierarchical planning*. It takes

S. Vogel, *Demand Fulfillment in Multi-Stage Customer Hierarchies,* Produktion und Logistik, DOI 10.1007/978-3-658-02864-0_2, © Springer Fachmedien Wiesbaden 2014

care of the interrelations between the individual planning tasks at different levels. A brief introduction to hierarchical planning will be provided in Section 2.1.2.

In Section 2.1.3, the interrelations between the individual planning levels will be discussed. An intuitive framework for this is the *supply chain planning matrix* (SCPM). It arranges the SCP tasks along the dimensions *planning level* and *supply chain processes*. Unfortunately, an important characteristic of all supply chains, the customer order decoupling point (CODP) is not reflected in the SCPM. The CODP separates forecast-driven from order-based tasks, and it will be covered in Section 2.1.4.

Almost all planning tasks in today's supply chains can be supported with software modules which are part of APS. APS are powerful implementations of the hierarchical planning logic. Since some of the ideas developed in the subsequent chapters to solve the DMC problem may be integrated into such systems, Section 2.1.5 will provide a short overview of APS.

2.1.1. Supply Chain Management

In a very basic sense, a *supply chain* consists of "two or more parties linked by a flow of goods, information, and funds." (Tsay et al., 1999, p. 301). These parties are typically involved in four types of key activities: *Procuring* necessary raw materials, transforming them into semi-finished and finished products in a series of *production* steps and finally, *distributing* and *selling* these products to the end customers (Lee and Billington, 1993). These activities have to be aligned closely to ensure that individual customer needs can be fulfilled in the best possible manner.

This alignment is usually referred to as *supply chain management*. The **objective of SCM** is to coordinate these aforementioned activities and to manage the relationships between the involved entities. The ultimate goal is to deliver superior customer value at fewer costs to the whole supply chain (see Christopher, 1998, p. 18). The breadth of tasks involved suggests that SCM comprises both a design and an execution perspective.[1]

A key characteristic of SCM is its focus on the collaboration of multiple parties. SCM has risen to prominence as traditional production settings based on vertically integrated companies have been gradually replaced by a sequence—or chain—of multiple parties working together. It is their joint effort which is required in modern industrial production settings.

Many drivers for the establishment of supply chains and of SCM originate from the market environment, for example the globalization of many markets. This qualitative change from traditional production environments to supply chains has been accompanied by a trend towards better customer orientation, resulting in an explosion of product variants, shorter product life cycles and more complex products. Many companies have responded to this challenge by specializing and concentrating on their core competencies (see Prahalad and Hamel, 1990). This has resulted in the participation of more and of many separate economic and legal entities in the overall process of value creation.

[1] For more comprehensive discussions on the term SCM and particularly its relationship to logistics, see Cooper et al. (1997) or Mentzer et al. (2001).

Therefore, most SCM initiatives focus on facilitating cooperation and on coordinating decisions at the interfaces to other enterprises along the logistical chain (Zimmer, 2002, p. 1).[2]

At an aggregate level, the key **tasks of SCM** are to reduce costs, particularly with respect to inventory, to gain efficiency in operations and to improve customer service (Lee and Billington, 1995). Essentially, SCM is concerned with determining the trade-offs between these apparently conflicting goals. Improving customer service and operations while at the same time preventing inventory levels from soaring requires a sophisticated and coordinated effort. For this purpose, SCM incorporates a broad spectrum of managerial decisions. Long-term strategic deliberations must be addressed while at the same time important tactical planning activities must be taken care of. This breadth of planning tasks can be handled with the hierarchical planning framework which will be discussed shortly, in Section 2.1.2.

The DMC Problem as an Intra-Organizational Channel Coordination Problem

As stated above, at the heart of most SCM issues lies the problem of coordinating multiple entities fulfilling a variety of tasks. A typical example is the *channel coordination* problem. In a particular sales and distribution channel, a number of independent supply chain entities such as manufacturer, wholesaler and retailer are collectively involved in bringing a particular product to market. The entities are independent since there is often no central authority which can exert discretionary power. Hence coordination is required as the entities differ in terms of their objectives, information endowments or general market power.

The main planning problem consists of incentivizing the independent entities to cooperate for their mutual benefit. A typical phenomenon in uncoordinated sales and distribution channels is *double marginalization* (see Spengler, 1950). Here, independent price setting decisions are made both by the wholesaler and by the retailer. Since each party only focuses on individual profit maximization, this not only jeopardizes overall supply chain profits, but also leads to individually disadvantageous results (e.g. see Corbett and Tang, 1999). Due to the lack of a central coordinating authority, such problems can only be mitigated by proper mutual contracts which align incentives and which encourage information sharing (e.g. see Cachon, 2003). Channel coordination is achieved if all parties involved independently make decisions which maximize joint profits (Barnes-Schuster et al., 2002, p. 173).

A typical by-product of missing channel coordination is the build-up of large and costly inventory positions at each of the involved supply chain entities. These excess inventories result from distorted aggregate demand signals and disproportionate ordering. Such distortions can occur if the demand variability increases from the perspective of more

[2] The related term ‘demand chain management’, while more appropriate to describe the market-related activities (see Selen and Soliman, 2002), was never accepted in literature and practice.

upstream supply chain entities. Forrester (1958, 1961) was the first to observe this phenomenon which is commonly referred to as the *bullwhip effect*. In a seminal paper, Lee et al. (1997a) identified and analyzed four major sources of the bullwhip effect in supply chains:

- **Myopic processing of demand signals** as end customer demand is invisible to intermediate supply chain entities,
- **rationing games** due to proportional allocations and unrestricted return policies,
- **order batching** and
- frequent **price variations**.

The literature on the bullwhip effect and on its prevention has increased beyond measure in recent years. Many theoretical contributions highlight the importance of centralized managerial control to solve such coordination problems. Yet, this is only rarely feasible. In the absence of a central coordination authority, most authors agree that increased transparency, information sharing and an alignment of incentives constitute key measures to prevent the bullwhip effect in supply chains and to ensure channel coordination (e.g. Lee et al., 1997b).

However, this result is not limited to supply chains consisting of separate (legal) entities. Problematic situations which are conceptually similar to the channel coordination problem may also arise in *intra-organizational* settings. If individual entities with private information and selfish behavior make uncoordinated decisions, phenomena which are similar to the bullwhip effect may also occur within organizations. In particular, this is the case for the DMC problem which was introduced in the previous chapter. This problem is characterized by a two aspects which closely mirror the root causes of the bullwhip effect in supply chains:

- Due to the salesforce composite forecasting method, the **true demand signal** from the leaf nodes is typically *not observable* at higher levels in the customer hierarchy.
- **Decentral information** in the customer hierarchy may lead to shortage or **rationing gaming**.[3]

While some sales staff in a customer hierarchy may also possess pricing power, there often exist centrally enforced pricing policies.[4] Furthermore, order batching within a sales organization is typically less of a problem than between independent entities in an inter-organizational supply chain. Hence, the effects order batching and price variations are less likely to be encountered in the DMC problem. Nevertheless, the other two effects

[3] Houlihan (1985) was the first to discuss the rationing or shortage gaming phenomenon also in an intra-organizational context. Note that some authors use the term 'Houlihan effect' when referring to rationing and shortage gaming in general, e.g. Disney and Towill (2003).

[4] Recall that the discussion in this thesis primarily focuses on MTS environments where prices are often set uniformly for all markets.

may result in distorted aggregate demand signals. This can be seen as an analogy to the bullwhip effect in an inter-organizational setting.

Countermeasures against this 'internal bullwhip effect' are similar to the inter-company case: The activities of the individual agents in the customer hierarchy have to be coordinated, information transparency needs to be improved and incentives must be aligned. In contrast to the traditional channel coordination problem, however, the individual agents in a customer hierarchy already have an established, hierarchical relationship. To some extent, this may allow for a tighter control of the resulting superior-subordinate relationships. But important information asymmetries remain and need to be dealt with. Overall, the DMC problem can therefore be interpreted as an *intra-organizational channel coordination problem*.

As can be seen, SCM is not limited to an inter-company setting. Many similar coordination problems also occur within larger firms with distributed decision-making. While the analogy discussed above only addresses the sales and demand fulfillment tasks in customer hierarchies, also the other SCM tasks performed by the different entities in a multi-divisional firm need to be aligned. An adequate planning concept to solve such coordination problems in inter-company and intra-organizational supply chains is hierarchical planning. In the following section, this concept will be introduced briefly.

2.1.2. Hierarchical Planning

According to Ijiri et al. (1968), planning can be understood "as the process of developing a strategy for changing or responding to changes in one's environment" by identifying and evaluating alternatives. In an SCM context, Fleischmann and Meyr (2003) defined supply chain planning (SCP) "as a generic term for the whole range of those decisions on the design of the supply chain, on the mid-term coordination and on the short-term scheduling of the processes in the supply chain." This definition exhibits two key characteristics: First, the large problem of planning an entire supply chain actually consists of many individual, but closely related subproblems. These subproblems are referred to as *planning tasks*. Second, this definition illustrates that several of these planning tasks can be grouped at certain *planning levels*.

The term planning level requires a definition. Mesarovic et al. (1970, p. 52) observed that 'level' is a rather generic term, and they distinguished between three different notions of levels in a planning context:

- **Strata**: Levels in the sense of strata refer to different *degrees of abstraction*. Strata may be used to differentiate between the extents to which certain features are included in a planning model.

- **Layers**: Layers refer to different *degrees of decision complexity* which result from vertically decomposing a comprehensive decision problem into one or multiple usually simpler subproblems.

- **Echelons**: Echelons refer to different *organizational levels*, i.e. the mutual relationships between different decision units in larger organizations.

While these three aspects are inextricably linked in most practical problems, the process perspective inherent in planning suggests putting a strong focus on the notion of layers in defining a planning level. Hence, the following definition of a planning level will be adopted. It is based on Emery (1964, p. 20), who summarized earlier work:

Definition 2. *A planning level is a particular vertical partitioning of a larger problem. A certain plan, addressing the entire large problem or parts of it, lies at a lower planning level if it partitions the behavior described by plans at higher levels into finer details.*

In many cases, such a partitioning may simply reflect the decisions that need to be made at different points in time.[5] Often, lower planning levels consist of multiple plans which collectively address the entire problem at the higher level. Emery (1964, p. 20) pointed out that such an apportionment corresponds to a consistent "one-to-many transformation" between the high-level plan and its associated lower-level plans. Consistency implies that the different lower-level plans are indistinguishable in terms of the variables which have been used in defining the high-level plan.

An early differentiation between different types of planning levels was introduced by Anthony (1965). He suggested partitioning a larger planning problem by grouping individual planning tasks according to the time during which these decisions have an effect. In particular, he observed that some decisions are more concerned with the broader aspects of the overall system behavior than others. The related decision periods are longer. The result is the familiar differentiation between *long-term*, *mid-term* and *short-term* planning levels which is typically used in SCP. With each of these three major planning levels, a number of key SCP tasks are associated (see Miller (2002, Ch. 1.1) and Voß and Woodruff (2006, pp. 4–5)):

- **Long-term** or **strategic planning** is concerned with setting the long-term objectives of a company or of an entire supply chain and with defining a strategy which allows meeting these objectives. Such decisions have major implications over a long period of time and are thus associated with high risk and many uncertainties. Typical strategic supply chain decisions pertain to the potential markets to serve and to finding ways to differentiate from competitors. From a design point of view, strategic planning requires making choices regarding the structure of the supply chain network and its key links. Such decisions typically have an impact over several years and are made by senior management, usually based on aggregated internal and also external data (Miller, 2002, p. 2). Decisions at a strategic planning level are the least structured ones, are associated with high levels of uncertainty and are often difficult to formalize in quantitative terms (Steven, 1994, pp. 54–55).

- **Mid-term planning** focuses on the efficient allocation and utilization of the resources which were established by long-term planning. At a mid-term level, SCP

[5] This perspective will be referred as a *decision-time hierarchy*, see Section 3.2.

tasks can be split along the four main functional areas procurement, production, distribution and sales planning. The time frame of mid-term planning covers at least one full seasonal cycle, i.e. usually at the minimum one year. Production planning—often the most important mid-term supply chain planning task—is typically split into two sub-tasks, particularly in the case of multi-site production environments. While **master planning** focuses on aligning and optimizing production plans across multiple sites, **production planning and scheduling** has a more limited scope and addresses lot-sizing, machine assignment, scheduling and sequencing decisions at the level of a single plant (Fleischmann and Meyr, 2003, p. 481).

Mid-term decisions are usually made by middle managers and lower-level senior executives (Miller, 2002, p. 4). An important decision which already has to be made at a tactical planning level is the development of specific inventory allocation policies. In case of foreseeable shortages during the mid-term planning horizon, these policies are used to determine which customers will be served with priority (Miller, 2002, p. 183).

- **Short-term planning** ensures that individual tasks per functional area are performed efficiently and effectively (Miller, 2002, p. 5). In most supply chains, this includes routine sequencing and lot-sizing decisions, but also distribution and transportation planning to deliver goods or to pick up material. In contrast to mid-term and long-term planning, the horizontal interrelationships between individual planning tasks at the short-term level are less crucial and the use of integrated decision models is less common. Instead, there is typically a close vertical relationship between short-term planning and **execution**. Operational short-term plans have a short planning horizon in the range of days, up to several weeks.

The above assignment of individual planning tasks to planning levels represents an ideal planning situation. In practice, the actual assignment is rather fuzzy and strongly depends on the particular supply chain (type) considered (Fleischmann and Meyr, 2003, p. 471).

Given the many interdependencies between the individual planning tasks at all planning levels, all decision problems should be considered simultaneously to find a solution which is optimal from a global perspective. However, designing and solving a monolithic model covering all major supply chain planning tasks is typically not feasible. Such a *simultaneous planning* model requires significant amounts of data and thus will have enormous memory requirements. Moreover, it will exhibit a high computational complexity, rendering it impossible in most practical cases to actually determine the optimal solution.

Another major problem of simultaneous planning is the uncertainty which is associated with the required long-term and mid-term forecasts. For example, production decisions for all individual final items have to be made for several years in advance. Since the accuracy of forecasts typically improves with shorter lead times, such a monolithic model could theoretically be executed again at later points in time with updated data. However, this is highly problematic. Most updated decisions can no longer be implemented in the

short run as they will be inconsistent with prior decisions. Furthermore, higher-level planning tasks have longer re-planning frequencies than short-term tasks. For example, supply network adjustments will be revised at most annually whereas lot-sizing decisions will usually be updated daily or weekly. A common schedule to revise all planning tasks at all planning levels will introduce undesirable nervousness in the planning system. Overall, simultaneous planning approaches are no feasible option in practice.

An alternative of the other extreme is *successive planning*. In a successive planning approach, the entire problem is clustered into several smaller subproblems with the objective of minimizing the interdependences between them. These subproblems will then be solved sequentially. Usually, this sequential planning approach will come at the cost of over-simplifying the interrelationship between the individual subproblems. In practice, only a one-dimensional (forward) flow of information is assumed between the subproblems while the impact of other subproblems is either estimated or ignored altogether (Steven, 1994, p. 12). This simplifies the planning situation considerably and usually permits determining feasible and often optimal solutions to each subproblem. However, the successive planning approach leads to a suboptimal overall solution.

A compromise between the simultaneous and successive planning approach is the so-called *hierarchical planning* concept (Fleischmann and Meyr, 2003, p. 457). In hierarchical planning, a larger planning problem is broken along the lines of hierarchically linked planning levels. At each planning level, only certain subproblems of the overall problem are solved. Moving down the planning hierarchy, one obtains a more detailed explanation of a complex planning problem. Contrariwise, moving up in the hierarchy leads to a deeper understanding of the overall problem and its significance (Mesarovic et al., 1970, p. 42). Lower planning levels are associated with a high degree of detail, a high re-planning frequency as well as a short planning horizon whereas the opposite applies to higher planning levels.

The key strength of a hierarchical planning concept lies in its ability to allow for *decision postponing*. While long-term and aggregate decisions with a long time horizon such as supply network planning have to be made early (i.e. at higher planning levels), decisions affecting more detailed issues may be moved to lower planning levels. These detailed decisions (e.g. lot-sizing or transportation planning) are thus postponed to later points in time when better decisions based on updated and more accurate information can be made.[6] However, it is important to account for interdependencies between these planning levels and to ensure that decisions made at a lower planning level are not in contradiction with prior decisions at higher planning levels (this is referred to as *consistency*). Decisions at higher planning levels should only restrict the decision space at the lower levels, but not pre-determine a particular decision for the short-term problems. The key challenge lies in ensuring that the decision spaces conceded to the lower planning levels always permit the generation of feasible detailed plans. While this splitting of the overall problem into multiple hierarchically aligned partial solutions usually does not necessarily lead to an

[6] In Section 3.2, this approach will be characterized as a *decision time hierarchy*.

optimum, it provides at least a feasible, consistent and in many cases quite good overall solution (Steven, 1994, p. 1).

The hierarchical planning concept was originally proposed by Hax and Meal (1975) as hierarchical production planning (HPP) for a tire manufacturer. This initial publication has spurred an enormous amount of follow-up work, was subsequently extended to various other industries and broadened in scope to include other supply chain processes. Nevertheless, all hierarchical planning systems are still built upon five major principles (Stadtler and Fleischmann, 2012): Decomposition, coordination, aggregation, model building and model solving.[7] Each principle will now be characterized in more detail.

Decomposition: As illustrated, monolithic models are usually difficult to solve in practice. Furthermore, neither is such a model readily accepted by managers in charge of specific SC tasks. Hence, hierarchical planning always entails a decomposition of the overall problem into a set of interrelated subproblems and corresponding smaller models.[8] In contrast to successive planning approaches, this decomposition leads to a hierarchical structure which typically exploits existing responsibilities and information channels (Steven, 1994, p. 1). Decomposition—or hierarchization—is thus closely linked to the existing organizational structure of a company or of an entire supply chain. Therefore, hierarchical planning facilitates the split of a larger planning problem into multiple decision areas along the lines of responsibility of individual departments or of separate legal entities.

Coordination: In contrast to successive planning approaches, the interrelations between individual subproblems in hierarchical planning are closely coordinated. Each subproblem belongs to a specific planning level and is linked to the next lower planning level in a series of top-down *instructions*. The subproblem at the higher planning level controls and restricts the decision space of the problem at the lower level by these instructions. This way, a high level of integration can be enforced, contributing to the consistency of the overall plan. Two types of instructions can be differentiated (see Stadtler, 1988):

- Primal instructions (e.g. target production quantities, available capacities or inventory levels) primarily limit the solution space and thus guarantee the solvability of the lower-level subproblem.
- Dual instructions (e.g. lot-sizing costs, inventory costs, more generally: transfer prices) directly affect the objective functions of the lower-level subproblems.

A number of other types of links may exist between individual subproblems besides simple unidirectional top-down instructions (see Steven, 1994, pp. 36–37). For example, a higher level of reciprocity can be ensured by an asymmetrically bidirectional link (e.g. one-way instructions with a feedback mechanism) or a truly symmetric, mutual link. The

[7] Steven (1994) and Mesarovic (1970) discuss similar principles.

[8] As will be discussed later in Section 3.2, such a decomposition is a prime example for a so-called *constructional hierarchy*.

latter is often the case between subproblems at the same planning level, e.g. between master planning and mid-term distribution planning.

Aggregation: Loosely speaking, aggregation refers to the grouping of similar objects into one (Steven, 1994, p. 43), usually with the objective of reducing complexity. The reverse operation to aggregation is referred to as disaggregation. A more thorough definition of these operators will be provided later in Section 3.1.2.

At higher planning levels, aggregation significantly reduces the complexity of the plan and the uncertainties of input data, e.g. by balancing lower-level demand forecast fluctuations. Typical dimensions are the aggregation of time, geographies, products and capacities:

- *Aggregation of time:* In mid-term planning, typically weekly or monthly figures of the expected demand are used rather than considering data at the level of days. As will be shown in Section 2.2.5, demand figures aggregated over time are less volatile and easier to forecast.

- *Aggregation of geographies:* Production planning can often be facilitated by combining the demands from several smaller regions. The actual geographical origin of the demands only needs to be considered at the later stages of distribution and transport planning.

- *Aggregation of products:* Several similar end items are combined to *product families* which in turn may be aggregated to *product types*. Often, there is no need for setup changes when producing items from the same product family. Items from the same product type can often be produced on the same production line. This facilitates aggregate production planning, and more detailed product data will only be considered when making scheduling decisions.

- *Aggregation of capacities:* Similarly, rather than considering the individual machine capacities per production line per minute, an aggregate figure is the amount of production capacity at a particular plant per month. The latter figure is often appropriate when deciding at which plant in a network production should occur.

Aggregation may be accomplished in a number of different ways:[9]

- *Perfect* (or consistent) *aggregation* is a fully commutative operation, i.e. it may be reversed without loss of information (Switalski, 1988, p. 384).

- Since perfect aggregation usually involves significant practical problems, Axsäter (1979) and Axsäter and Jönsson (1984) proposed an alternative approach termed *approximative aggregation*. Acknowledging that some loss of information is often unavoidable in practice, approximative aggregation merely requires that the results of both the aggregated and detailed model should coincide as much as possible.

[9] For a general framework on aggregation and disaggregation methodology and a literature overview, see Rogers et al. (1991).

- Rather than relying on a formal aggregation method, many practical applications make use of an existing "natural hierarchical structure" (Axsäter, 1979, p. 79) and thus perform a *heuristic aggregation*. Both the object and the aggregation method are picked conveniently to facilitate the overall planning problem. Typical applications of heuristic aggregation date back to the original HPP concept by Hax and Meal (1975). Heuristic aggregation is often applied with respect to the aggregation of products (see Miller (2002, pp. 25–26)). At a lower planning level, product families are formed by grouping several end items which use similar tooling or which have similar setup costs. Items from the same product family are often produced together to limit changeovers and to facilitate lot-sizing. At a more aggregate level, several product families are often combined into product types which have a similar seasonal demand and which can be produced at the same production rate on the same production line (albeit incurring changeover costs).

Independent of the actual aggregation method employed, aggregate plans are less detailed to allow for efficient planning over a longer time horizon. Some loss of information is justified at higher planning levels as aggregation unburdens the higher levels from irrelevant details (Rohde and Wagner, 2008, p. 172). However, at later (lower-level) planning stages, these aggregate plans need to be disaggregated and detailed information has to be amended again. Since it is mandatory that feasibility is preserved throughout the planning process, the disaggregation steps need to ensure that feasibility of the lower-level subproblems is maintained in all subsequent periods. Put differently, the disaggregation needs to be *consistent*. Most research on feasibility and consistency in hierarchical planning assumes a deterministic planning setting. For this, Axsäter (1986) has given conditions which ensure feasibility at an aggregate level. However, an unsuitable disaggregation at a particular instant in time may destroy feasibility for the remaining part of the planning horizon of the (originally feasible) aggregate plan. Methods to ensure consistent disaggregation have been suggested, among others, by Gabbay (1979) and by Bitran et al. (1981). A first characterization of the disaggregation problem in stochastic planning situations has been given in Ari and Axsäter (1988).

In many practical hierarchical planning situations, there is actually no need for the disaggregation step to exactly reverse the original aggregation operation. Rather, it is usually sufficient if the disaggregated problem fulfills the constraints imposed by the aggregate problem. Hence, any consistent *allocation* of a top-level object to multiple lower-level objects is usually a feasible and thus perfectly acceptable disaggregation. Schneeweiß and Kleindienst (2004, p. 270) observed that most HPP problems usually constitute such simpler aggregation-allocation problems.

Model Building: The various subproblems in an SCP problem are usually solved with the help of planning models. These have the form of a mathematical model, consisting of an objective function and several sets of constraints. The constraints represent the essential restrictions which must be satisfied by the decision variables. Many simple planning models have the form of a *linear program* (LP). Both the objective function as

well as the constraints correspond to linear expressions and the decision variables can take on any continuous value within a certain range. More complex model formulations usually lead to *mixed-integer programs* (MIP) with continuous as well as binary and other integer decision variables.

A necessary requirement of model building is abstraction. Planners need to trade-off model simplicity against the desire to account for numerous behavioral aspects of the complex systems in real-world supply chains. The choice of a particular level of abstraction (stratum) to describe a given SCP problem is often subjective and depends significantly on the observer, his knowledge and interest in the operation of the system (Mesarovic et al., 1970, p. 40).

The modeling principles at different strata are generally not related (Mesarovic et al., 1970, p. 41): For example, at a strategic planning level, a supply chain is usually modeled as a set of nodes—representing supply and customer demands—which are linked by input-output relations and which need to be optimally balanced over a longer period of time. At this aggregate level, the focus of the model lies on the (homogeneous) flow of goods and information between the nodes. By contrast, at a disaggregate, operational level, order promising models are represented at the level of individual supply units and individual customer orders. Both the supply units and the individual orders are no longer seen as homogeneous since individual cost and value functions may be associated with each supply unit and with each customer order. An entirely new quality of the planning problem has been introduced into the model at the disaggregate level.

Model Solving: Lastly, the individual subproblems in supply chain planning need to be solved. Model solving implies finding an extremal value for a given objective with efficient algorithms. Standard algorithms for mathematical programs permit solving almost all LP and many MIP to optimality. In situations where a true optimization is either impossible or the computation has a too long run time (e.g. very complex MIP in lot-sizing), heuristics are applied. Heuristics do not always find the optimal solution but rather help to find an acceptable solution in a reasonable amount of time.

In contrast to a monolithic model, each subproblem in a hierarchical planning framework is solved individually and may be associated with a different solution method. This permits tailoring the solution method to the problem type.

Overall, it can be stated that hierarchical planning is a flexible and pragmatic approach for many practical SCP situations. Its popularity is often summarized by three major advantages over other planning concepts (see Dempster et al., 1981, p. 708):

- Hierarchical planning reduces complexity: Smaller subproblems are usually easier to solve, and information flows and mutual dependencies can be minimized by a clever decomposition of the overall problem.

- Hierarchical planning allows coping with uncertainty by postponing decisions until reliable and disaggregate forecasts are available.

- Hierarchical planning embraces the existing hierarchical structure found in most companies: The individual sub-models and planning tasks are often aligned in a natural way with existing organizational structures and decision-making authorities. This is usually a key factor to ensure acceptance of the resulting plans by its key users.

After this general outline of the hierarchical planning concept, the discussion in the next section will focus in more detail on the interrelations between the many planning tasks.

2.1.3. Interrelations between Planning Tasks

Clearly, decomposing a complex problem into smaller and more manageable subproblems offers significant benefits. In hierarchical planning, these subproblems do not rank equally, but exhibit a hierarchical relationship. One subproblem may exercise more power or simply has to make decisions at an earlier point in time (Schneeweiß, 1998, p. 547). In the following, first a framework will be presented which allows structuring the various vertical relationships in hierarchical planning. Then, also horizontal relationships at the same level of planning will be considered. Both perspectives are captured in the Supply Chain Planning Matrix.

Hierarchical Interrelations - The Schneeweiß Framework

Schneeweiß proposed a general framework to describe and analyze the interrelations between different subproblems at different levels in hierarchical planning. At the very minimum, a hierarchical planning situation consists of two interrelated levels. A hierarchically superior planning level, termed *top level*, interacts with a lower *base level*,[10] as shown in Figure 2.1. The plans generated by both levels will ultimately be implemented in a concrete *object system*. For example, the top level may correspond to a mid-term aggregate planning level whereas the base level may correspond to a short-term production scheduling level. Plans from both levels will ultimately be implemented in a production environment. Schneeweiß suggested differentiating between three different types of interrelationships which govern the link between the top and the base planning levels: instructions, reactions and anticipations (Schneeweiß, 2003, pp. 17–18).

Instruction: As outlined before, the top level may exercise a direct top-down influence on the base level via a set of *instructions*. These instructions limit the decision space of the base level and thus influence its subsequent decisions and plans. If the top level corresponds to a mid-term planning level in a production environment (master planning, see also Section 2.3), its output consists—among other things—of required production quantities and of end-of-period inventory levels which are passed downwards as instructions. Assume that the base level is equivalent to short-term production scheduling and has to come up with a detailed schedule determining when to produce which quantities.

[10] The terms base level and lower level will be used interchangeably.

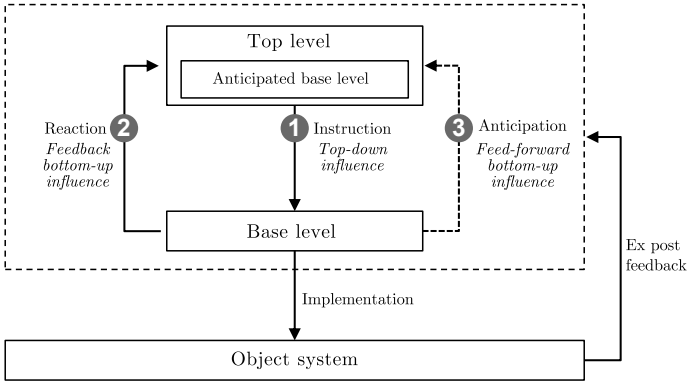


Figure 2.1. – Interrelations between hierarchical planning levels (slightly adapted from Schneeweiß, 2003, Fig. 1.10, p. 17)

The production quantities and inventory levels from the top level then correspond to constraints which the plan at the base level has to meet.

Both the plans at the top and the base level are ultimately executed and implemented within the object system, i.e. in the actual production environment. After the plans from both planning levels have been implemented in the production environment, *ex-post feedback* information from the object system can be obtained and returned to both planning levels. This information may help to improve subsequent planning iterations at both planning levels. For example, if the actual production output turned out to be less than planned, e.g. due to quality rejects, this information needs to be taken into account in the next planning and production cycle.

Reaction: In many planning situations, the plan derived at the base level (before being implemented in the object system) already contains information which is helpful for the top level, e.g. if the instructions provided by the top level do not permit the base level to determine a feasible plan. This bottom-up feedback is referred to as a *reaction*. Such a reaction may be used to trigger a recalculation of the plan at the top level to remove the infeasibility. If a reaction function is present between both planning levels, the coupling between both levels is no longer of an unidirectional nature, but can better be described as being asymmetrically bidirectional (see Steven, 1994, p. 37).

Anticipation: The third type of interrelation or coupling between both planning levels is the most sophisticated one: Usually, planners at the top level are aware of the hierarchical planning situation and know that the top level plan will be refined at the base level. In deriving their own plans, the top level planners may have means to anticipate the impact of their top level decisions on the lower level. For example, if the top level planners can anticipate minimum lot-size requirements which have to be respected in the operational

base level plans for sequencing and scheduling, better overall plans may result. Top level plans may then be designed appropriately in the first place to prevent infeasibilities at the lower level.

The extent of such an improvement obviously depends on the form and quality of the anticipation model (cf. Schneeweiß, 2003, pp. 42–44): On the one extreme, a *perfect anticipation* will require the top level to be in possession of an embedded, fully specified base-level planning model. On the other extreme, the reaction of the base level may not be taken into account at all and only some general features of the lower level may be considered by the top level. This other extreme form of an anticipation function is referred to as a *non-reactive anticipation*. It is similar to the coupling between subproblems in successive planning environments. In many practical cases, only some aspects of the base level reaction can be anticipated by the top level, either explicitly or only implicitly. Following Schneeweiß, these types of coupling may be termed *approximate reactive anticipation*.

The Supply Chain Planning Matrix

The concept of hierarchical planning—and in particular the framework of Schneeweiß (2003)—primarily stressed the vertical relationships between individual decisions which are made at different planning levels in an SCP environment. In general, the individual planning tasks will have further mutual relationships. Steven (1994, pp. 9–10) introduced a classification into three different types of interdependencies between individual planning tasks in a supply chain:

- **Vertical or temporal interdependencies**: Vertical relationships reflect the hierarchical decomposition of the overall supply chain planning problem and have been discussed above. Since planning levels typically result from clustering planning tasks with a similar temporal scope, most vertical dependencies between planning tasks at different levels have a strong temporal component.

  Two types of temporal relationships prevail: On the one hand, current plans will depend on past decisions made at higher planning levels. For example, the production environment built according to previously established strategic and long-term plans limits the current production possibilities. On the other hand, current plans also set precedence for future plans. The inventory levels which have been determined in today's aggregate plan determine the amount of orders which may be accepted in one of the following periods by the short-term demand fulfillment planning task.

  As already highlighted in the introduction, these temporal interdependencies and the resulting lead times are at the heart of many allocation problems in supply chains, including the DMC problem.

- **Horizontal or objective interdependencies**: Horizontal or objective interdependencies reflect the relationship between planning tasks at the same planning level, i.e. decisions with a similar temporal scope. Such objective relationships may

exist *between several instances of the same planning task*. Production plans for different products have to compete for the same set of limited common resources such as the available manufacturing capacity or available raw materials. For example, the decision to produce a particular item i usually cannot be made without considering the remaining operations schedule for the same period. It is the entire operations schedule for all products which determines whether sufficient resources remain for the production of i. Furthermore, the sequence of all production jobs in a particular period has an influence on the associated costs of the production of i (e.g. via required production line setup operations).

Other horizontal interdependencies exist *between different planning tasks*: For example, purchasing decisions can only be made once the actual production program is known. However, once purchasing plans have been implemented, the production possibilities have been fixed for a certain period of time and during the purchasing lead time, production decisions are difficult to alter.

- **Level of integration**: The level of integration characterizes to which extent related subproblems are handled separately. In case of strong objective interdependencies, a comprehensive model will ensure a high level of simultaneous coordination between the constituent subproblems. Partial models with a reduced scope, by contrast, have been stripped off most interdependencies with other subproblems. This allows planners solving them successively.

  For example, if procurement planning can be neglected (e.g. because supplier lead times are short and there are no constraints on quantities) and if sales and distribution activities are handled by an external third party (with whom detailed service level agreements exist), the core planning problem in this supply chain for the mid- and short-term essentially reduces to a sequence of manufacturing planning steps. As outlined before, master planning and detailed scheduling are hierarchically linked and may usually be solved successively, with one-directional top-down links via instructions.

  High levels of integration can often be found at more aggregate planning levels, for example in strategic planning. Given the close interrelationships, procurement options, production facilities as well as the geographical scope of distribution and sales areas need to be planned simultaneously.

The multi-faceted nature of the interdependencies between the various supply chain planning tasks is conceptually captured by the *supply chain planning matrix*. It depicts the major SCP tasks as modules in a two-dimensional framework. A typical representation of the SCPM is given in Figure 2.2. The SCPM complements the hierarchical planning perspective (vertical axis) with a functional view (horizontal axis). The SCPM representation decomposes the overall SCP problem at each vertical planning level into the individual planning tasks associated with the four major supply chain processes: procurement, production, distribution and sales (Fleischmann et al., 2008). This horizontal

sequence of processes roughly corresponds to the flow of material in a typical supply chain (see also Section 2.1.1).

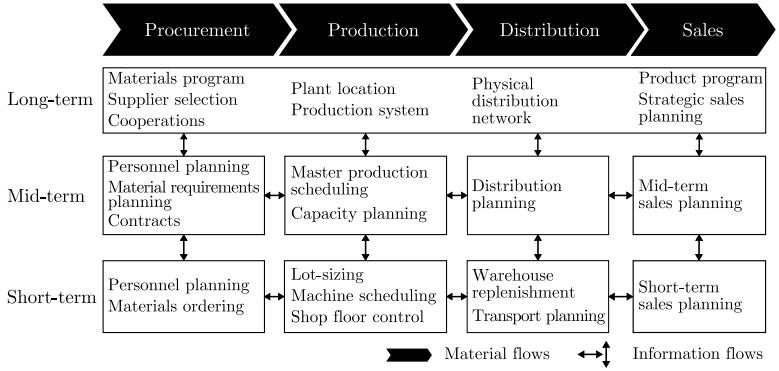


Figure 2.2. – Supply chain planning matrix (Fleischmann et al., 2008, Fig. 4.3)

The vertical division into a long-term, a mid-term and a short-term planning level not only caters to the hierarchical structure of the individual supply chain planning tasks, but also mirrors the different planning intervals. The vertical links, on the one hand, represent key temporal dependencies. On the other hand, they also correspond to the governance structure and the levels of responsibility of many companies. Strategic, aggregate decisions are usually made by higher-ranked planners whereas operational decisions are delegated to specialists.

Different levels of integration between the individual planning tasks are visualized by more aggregate modules. For example, most long-term decisions in a supply chain cannot be limited to a particular functional domain, but strategic procurement, production, distribution and sales decisions are closely related, as discussed above. Thus, they are represented as one module in the SCPM (top row). In many software solutions for supply chain planning, higher levels of integration can also be found at a mid-term planning level. For example, aggregate and master planning often comprises not only production, but also procurement and distribution decisions (see also Section 2.3).

The SCPM is a framework which can be employed for many different supply chains. However, this general applicability implies that not all planning tasks have to be present in a particular supply chain. For a general overview of the characteristics of different supply chain types and planning requirements, see Meyr and Stadtler (2008).

2.1.4. The Position of the Customer Order Decoupling Point

One particular aspect which is typically not captured by the SCPM representation is the position of the so-called *customer order decoupling point* (Fleischmann et al., 2008). Its importance stems from the fact that the four major processes in a supply chain—procurement, production, distribution and sales activities—can be either executed on an

anticipative or on a reactive basis. While reactive processes are triggered by an explicit order of a subsequent supply chain member or by a subsequent supply chain process, anticipative processes are pushed by forecasts regarding prospective, i.e. not yet placed orders (Fleischmann and Meyr, 2003, p. 462). The boundary between both types of processes is referred to as the CODP. The planners who are responsible for the forecast-driven processes usually cannot anticipate all orders from the subsequent supply chain members (or processes). Hence, buffer stocks are necessary and are held at the position of the CODP point to hedge against forecast errors.

There are a number of different options regarding the actual **position of the decoupling point**. Its choice is a highly strategic decision which has an impact on a number of supply chain characteristics. Properly chosen decoupling points allow for significant cost and efficiency improvements in a supply chain. Moreover, a well-defined CODP has the potential to be a key differentiating factor against competitors (see Sharman, 1984).

At an operational level, the CODP determines the lead time after which an order can be fulfilled (Hoekstra and Romme, 1992, p. 8). At higher planning levels, the choice of a particular CODP is closely linked to the pursued manufacturing strategy (e.g. job shop or flow production).[11]

The existing balance of power between customers and manufacturers in a particular industry constitutes a highly influential factor in determining where to place the CODP. In a very competitive industry with demanding customers, an order request which cannot be fulfilled on short notice by a company is typically lost. Hence, delivery times have to be shortened. Since procurement, production and assembly times are often longer than the acceptable customer lead time, as many steps as possible should be executed on an anticipative basis. In practice, this often implies executing all production steps based on forecasts and building up inventories of final products to hedge against uncertainties in demand. To keep the amount of finished goods inventory and the associated costs small, the ability to prepare accurate demand forecasts with small forecast errors is a crucial prerequisite (Silver et al., 1998, p. 244). In such an **MTS** environment, the entire production is decoupled from customer orders. The customer lead time is thus largely equivalent to the distribution time as transportation from the central inventory stock point is only initiated upon order confirmation. The position of an MTS decoupling point is outlined in Figure 2.3.

A further shortening of the customer order lead time can be achieved by also executing some or all distribution processes already on a speculative basis. This is often the case in supply chains for consumer goods where replenishments of regional warehouses are already planned centrally, based on forecasts. This is referred to as a **make-and-ship-to-stock** (MSTS) or deliver-to-stock environment (Hoekstra and Romme, 1992; Fleischmann and Meyr, 2004). This approach implies that the manufacturer designates some stocks early to serve a certain share of the spatially distributed demand. A more extreme version of MSTS is a vendor-managed inventory where all supply chain processes including the

[11] A more detailed discussion of these aspects can be found in Hoekstra and Romme (1992, p. 70).

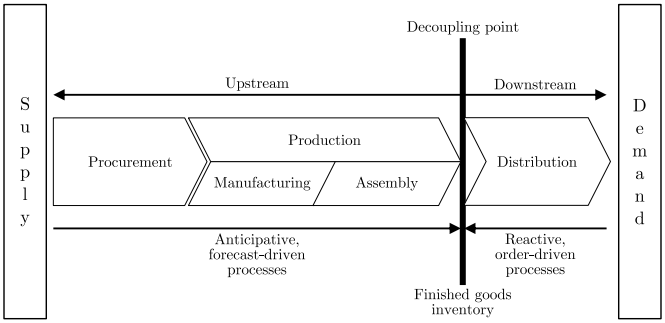


Figure 2.3. – Make-to-stock decoupling point
(adapted from Fleischmann and Meyr, 2004, Fig. 2)

entire distribution to the customer are executed based on forecasts. The manufacturer directly controls the receiving storage of the customer.

On the other extreme, it is possible to execute all processes, including development and procurement only after an order has been received. This leads to an **engineer-to-order** (ETO) decoupling point. Even the design of the product is dependent on an actual customer order. Practical examples of ETO decoupling points include major project-based, capital-intensive constructions such as ships or large buildings.

More typically, the decoupling point is often positioned after the procurement processes. This is referred to as a **make-to-order** (MTO) environment as all production-related steps are only executed once a concrete customer order is available. Depending on the actual manufacturing strategy employed, customer lead times may become rather lengthy. An intermediate strategy between MTO and MTS is referred to as **assemble-to-order** (ATO). This strategy exploits the fact that many final items are not produced in a single step but are assembled from a number of semi-finished items. If the assembly time of the final products is rather short compared to the production time of the constituent semi-finished items, (inventory) costs may be reduced and multiple variants of closely related products may be offered at competitive customer lead time durations.

The breadth of possible decoupling point locations and the associated production environments are summarized in Figure 2.4. When moving the CODP further downstream, the extent of forecast-driven, anticipative processes increases whereas the length of order-based, reactive processes shortens. In Figure 2.5, the order-driven planning tasks associated with different positions of the CODP have been indicated symbolically in the SCPM.[12] The lower right triangular area marked in dark gray, covering primarily the short-term sales planning task, corresponds to the order-driven tasks in an MTS production environment, but all procurement, production and most distribution decisions are initiated based on forecasts. In ATO environments, by contrast, not only are all short-term sales planning tasks executed based on individual orders, but also most dis-

[12] This representation extends Figure 2.2.

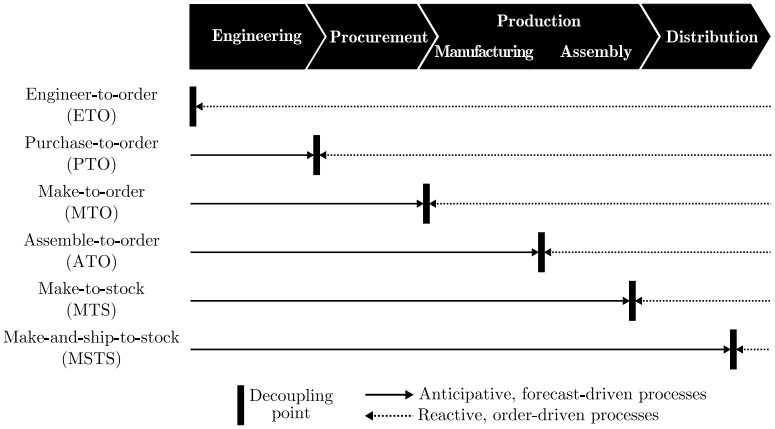


Figure 2.4. – Overview of different positions of the decoupling point and the associated production environments

tribution and many short-term production tasks since product assembly is dependent on customer specifications. Lastly, the combination of all highlighted areas corresponds to an MTO environment where most production and even some procurement decisions require a concrete customer order.

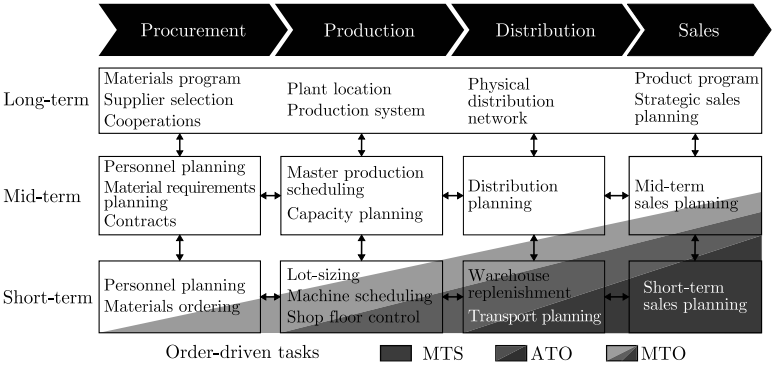


Figure 2.5. – Order-driven tasks in the supply chain planning matrix for different decoupling points (adapted from Fleischmann et al., 2008, Fig. 4.3)

It is important to note that this clear distinction between prototypical decoupling points and their associated supply chain environments primarily serves analytical purposes. In practice, mixed forms prevail:

- Companies often operate different supply chain configurations simultaneously to address different sales channels or different regions (Hoekstra and Romme, 1992, p. 68). For example, while orders from some regions may be served from final item inventories (MTS case), orders from other geographies may first trigger assembly (ATO case, e.g. to ensure a product is conform to national requirements in terms of language/labeling or power system).

- The differentiation between anticipative and reactive processes is often difficult to draw. In an ATO environment, the production of some subassemblies with short production times may only be started upon receipt of an actual order (MTO), while other subassemblies may be taken from inventory.

- Finally, there are hardly any examples of pure 'to-stock' strategies if customer orders consist of multiple order lines. Consider retailing as a typical MTS environment. Although individual orders are served from stocks of final goods, there is usually still a final assembly step needed for packaging. Hence, such an environment has a close similarity with an ATO setting. Several individual final items are combined or assembled 'on the fly' to full shipments which are then collectively dispatched with a common due date (see Okongwu et al. (2012) and Xu et al. (2009) for examples and models).

The choice of a certain position of the decoupling point is closely linked to the concept of *postponement*. Postponement means delaying activities until exact demands materialize in the form of individual orders. This strategy reduces the need to maintain costly inventories, at the cost of longer lead times. The concept of postponement was originally introduced into the marketing literature by Alderson (1950). Bucklin (1965) developed the postponement concept further as a means to manage and shift supply chain risks. Zinn and Bowersox (1988) introduced a differentiation of postponement with respect to time, place and form:

- **Time**: Delay all supply chain activities until orders have been received, e.g. as in an ETO, PTO or—more commonly—in an MTO environment. This strategy helps to reduce inventory levels, but comes at the cost of longer customer lead times.

- **Place**: Transportation processes are delayed until orders have arrived. This corresponds to the centralization of stocks typically found in MTS environments.[13] By contrast, MSTS is the corresponding decentralized approach, i.e. with an early commitment of stocks to demand regions.

- **Form**: All manufacturing and assembly activities which determine the form of the final items are delayed until demand is known exactly. To keep the customer lead time within an acceptable range, products are designed such that they are composed of a few common basic assemblies. A series of quick customization operations will

[13] A similar aspect arises in the DMC problem. The concept of so-called virtual safety stocks in multi-stage customer hierarchies will be discussed in Section 5.3.

transform the basic assemblies into a specific product variant. Such a form-related postponement is the most widely understood type of postponement and is often implemented in an ATO environment. An example is the widely cited computer assembly case study of Hewlett-Packard, see Lee and Billington (1995).[14]

Further theoretical extensions to the postponement idea were provided by Garg and Tang (1997) who studied supply chains with two or more points where product differentiation occurs. A more in-depth discussion of the benefits of postponement in stochastic planning environments was given in Aviv and Federgruen (2001).

Postponement ideas have also been embraced to solve a number of other planning problems in supply chains, notably allocation problems (e.g., see Bish et al., 2008). Moreover, the idea of *decision postponement* can even be considered as a general design principle for planning systems, especially hierarchical planning systems. As will be shown in the course of this thesis, this tenet is equally well applicable to the DMC problem.

2.1.5. Advanced Planning Systems

To close this broad overview of supply chain planning issues, a brief look will be taken at basic standard software for supply chain planning and supply chain management. Such APS have been positioned by software vendors as comprehensive solutions to support (intra-organizational) SCM. The 'advanced' nature of APS does not stem from particularly advanced optimization logic in its software modules. Rather, it is the implementation of a hierarchical planning concept based on the idea of integral planning and the use of true optimization in standardized, extensible software modules which has led to the wide adoption of APS in many SCM environments (Fleischmann and Meyr, 2003, p. 458).

As a tool primarily designed for supply chain planning, APS do not replace commonplace *enterprise resource planning* (ERP) systems. Rather, APS and ERP systems are complements. APS rely on ERP systems for various types of input data and send back some of their outputs in the form of instructions to be executed by the ERP system (Fleischmann and Meyr, 2003, p. 480). Hence, ERP systems ensure the execution and control of the plans which have been generated by the APS. They also provide important reporting functionalities. To better understand the relationship between APS and ERP systems, it is helpful to briefly review the historical development of enterprise software systems.[15]

An early computer-based system is *material requirements planning* (MRP), popularized by Orlicky (1975). MRP systems essentially consist of a database of material requirements and their dependencies. They provide three major functions:

1. To calculate inventory levels and requirements in an automated manner (particularly regarding work-in-process inventory),

[14] Hewlett-Packard was among the first electronics manufacturers which postponed adding country-specific power adapters and manuals to their printers only as late as in the regional distribution warehouses. Demand forecasts are much more reliable at regional warehouses which are closer to the final customers.

[15] The following points have been summarized from (Miller, 2002, p. 88ff).

2. to prioritize production jobs and

3. to determine production requirements at a detailed level, e.g. regarding subassemblies.

MRP systems can be seen as complements or as basic extensions to HPP systems (Miller, 2002, p. 90): While HPP systems are used to plan independent demands, MRP determines the dependent demand which derives from the independent demand. The major drawback of the MRP planning logic is that it does not explicitly consider capacity constraints. In practice, a series of MRP runs has to be executed and the associated plans have to be checked until the feasibility of the overall HPP plan is ensured (successive planning).

The focus of MRPII (Manufacturing Resource Planning) systems goes beyond materials. These systems seek to provide a common information basis for all major processes involved in a manufacturing environment, for example by integrating demand forecasts as well as financial or personnel-related data. At the core of MRPII is the *master production schedule* (MPS), a plan for raw material requirements, production quantities, staffing and inventory levels. Furthermore, MRPII systems usually consider limited capacities in the planning process and thus offer some functionality for due date and capacity planning for the major production resources. However, most MRPII systems still build upon a successive planning concept. Hence, plans obtained under MRPII logic usually fall short of the solution quality which may be obtained by a hierarchical planning method.

Evolutionary improvements to the MRPII concept led to ERP systems. These can be seen as more integrated software solutions due to a centralized database and a more thorough connection of individual modules. However, ERP systems continue to lack a systematic consideration of resource availabilities and are still predominantly based on a successive planning method. While offering adequate support for execution and controlling in a production environment or supply chain, ERP systems have only limited planning functionality and can only provide restricted decision support.

APS aim at filling this gap by providing a more holistic planning perspective and a truly capacity-oriented planning approach. While many APS are positioned as all-purpose solutions for all SCM applications, most commercially available solutions address only a subset of the planning needs which arise in SCM (cf. Bartsch and Bickenbach (2002, p. 28) and Knolmayer et al. (2009, p. 21)).

Adopting the concept of hierarchical planning, APS split a large planning problem into a set of smaller, loosely coupled planning modules. In many APS, this structure is similar to the elements of the SCPM (see Meyr et al., 2008b) which have been introduced in Section 2.1.3. By employing a simultaneous planning approach for tasks positioned at the same level and by using successive planning for problems positioned at different planning levels, APS aim at striking a balance between true optimization and integral planning on the one hand and between finding a feasible, flexible solution on the other hand (Fleischmann and Meyr, 2003, p. 457).

Usually, higher- and lower-level APS modules are linked by a non-reactive anticipation function since the plans of the higher level are forced upon the lower level in the form

of instructions and constraints. Reactive feedback is often only considered with a certain delay due to the rolling horizon planning methodology. However, some efforts have recently been made to enable some event-driven planning functionality to shorten response times (Meyr et al. (2008b, p. 114); see Lautenschläger (2008) for a case study).

Commercially available APS solutions differ significantly in terms of the functionality provided. Usually, not all software modules have to be installed simultaneously and many modules have been pre-customized to address the requirements of a specific industry. Nevertheless, the functionality of most APS can be clustered roughly into the modules depicted in Figure 2.6. Following Meyr et al. (2008b), the tasks associated with the key software modules depicted in Figure 2.6 can be summarized as follows:[16]

- The design of the supply chain network structure (plant locations and distribution system) and the associated key material flows between the nodes in the network are determined in the **strategic network design planning** module. As discussed above, these design decisions have to include purchasing options and long-term sales opportunities. These strategic decisions (see also page 26) are made with a long-term planning horizon.

- **Demand planning** primarily generates short- and mid-term demand forecasts at various levels of aggregation. These sales forecasts provide a key input into the master planning, distribution and demand fulfillment modules. Furthermore, demand planning also offers functionality to monitor and control the sales forecasts.

  In addition to that, demand planning often contains procedures to determine safety stock requirements. These requirements then serve as constraint in subsequent master planning (see below). This is justified as the amount of required safety stocks depends on the size of the forecast errors (see Fleischmann and Meyr, 2003, pp. 487–488).[17] Many demand planning modules also offer functionality for simulations and what-if analyses, e.g. for the planning of sales and marketing campaigns and for new product introductions or for product retirements. A broader perspective on typical demand planning functionalities will be provided in Section 2.2.

- Most APS synchronize all major mid-term decisions regarding procurement, production and distribution planning via a common **master planning** module. Based on mid-term sales forecasts from the demand planning module, master planning determines aggregate purchasing, production and transportation quantities, ensuring that the available supply chain network is utilized as efficiently as possible. As a result, master planning provides a holistic optimization of all decisions at a mid-term planning level. It is thus a significant improvement over prior capacity planning approaches in MRPII and ERP systems which only apply successive planning methods. Basic master planning is also covered in more detail in Section 2.3.

[16] As indicated in the following, some modules will be discussed in more detail in the subsequent sections.
[17] However, safety stock levels may also be determined as part of the master planning module (e.g. see Betge, 2006, p. 55).

- **Demand fulfillment** is a module for the short-term which takes care of expected and arriving customer orders. Basic demand fulfillment approaches merely match the available supply quantities with the arriving orders. More sophisticated modules have been designed with a profit improvement objective, e.g. by reserving scarce quantities for more important customer groups. The tasks of this module are addressed more comprehensively in Section 2.4.

- The actual functionality provided by the **production planning and scheduling** module depends largely on the actual production environment and industry. Often, this module is split into a separate production planning sub-module which determines lot sizes based on the constraints set by the master planning module. Subsequently, machine scheduling and shop-floor control issues are handled by a scheduling sub-module.

- **Transport and distribution planning** addresses all short-term transportation planning needs. Operating within the constraints set by strategic and mid-term planning in terms of the structure of the supply network and the aggregate production and distribution quantities, it is used to optimize vehicle loadings and schedules. This module fulfills an important function termed *deployment*, i.e. to physically link short-term product availabilities (from production planning and scheduling) with short-term demands (from demand fulfillment).

- **Purchasing & material requirements planning** usually complements existing MRP functionality which is embedded in most ERP systems. While the ERP system performs the operational MRP tasks such as providing basic bill-of-material explosions and generating purchasing orders, this APS module addresses more advanced planning needs, e.g. accounting for constraints such as limited resource availabilities, supporting the supplier selection process or determining order quantities in the case of quantity discounts.

The APS modules employ mathematical models to represent the objectives and constraints of each planning task. The models are solved with the help of algorithmic approaches based on the principles of mathematical programming. For a comprehensive overview of the different solution methods used in APS modules, see Dudek et al. (2002, p. 50). A key problem of these solutions is that they are often not well understood by practitioners (Steven, 1994, p. 21). As a consequence, practitioners may at times become leery of these solutions and rather prefer to run their operations via proven rules-of-thumb. This is particularly important when an APS is being implemented for the first time. Lin et al. (2007) argued that the human factor is too often overlooked in most APS implementation projects. Challenges with the planning process to implement an APS, especially from a modeling perspective, have been analyzed in Zoryk-Schalla et al. (2004). Jonsson et al. (2007) and Rudberg and Thulin (2009) presented case studies which illustrate practical applications of APS in different industries, and Kjellsdotter Ivert and Jonsson (2010) gave a comprehensive discussion of the overall benefits which may be gained for a

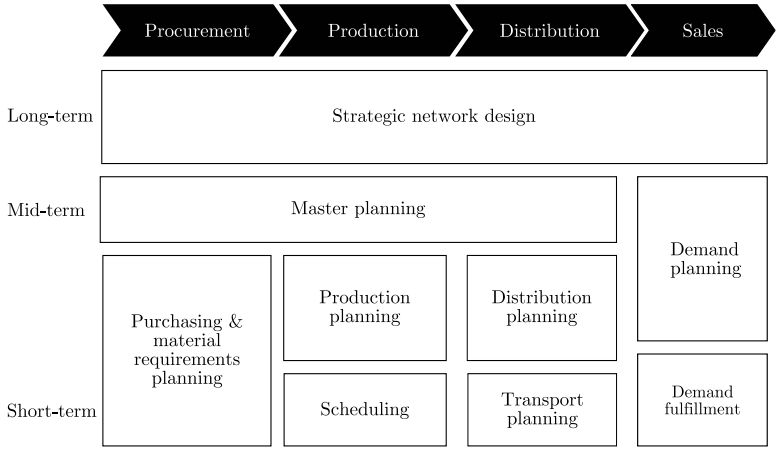


Figure 2.6. – Software modules in the SCPM (Meyr et al., 2008b, Fig. 5.1)

company. A recent textbook on APS fundamentals based on a simulation case study is Stadtler et al. (2012).

This concludes the basic overview of SCP in general and APS in particular. In the remaining sections of this chapter, the three key SCP tasks and their APS software modules which are most important for the DMC problem will be covered in more detail. As illustrated in Figure 2.7, demand planning will be addressed in the immediately following Section 2.2. Based on the demand planning inputs, master planning (Section 2.3) can then determine aggregate procurement, production and distribution plans. Lastly, Section 2.4 will cover the actual demand fulfillment process which matches supplies as determined by master planning with demand as forecast by demand planning. As indicated before, Section 2.5 will conclude the chapter.

2.2. Demand Planning

As outlined in the previous section, the ultimate goal of all SCP tasks is to optimally prepare a supply chain for the arrival of customer orders so that the available supplies match demand as closely and as profitably as possible. While the previous discussion centered on the general planning tasks to be performed, the purpose of this (demand planning) and the following section (master planning) is to focus on the major forecast-driven tasks, before discussing the handling of actual customer order arrivals in Section 2.4 (demand fulfillment).

The subsequent presentation is primarily geared towards settings where the CODP separates production and distribution processes, corresponding to an MTS environment. Starting in Section 2.2.1, the major objectives and planning tasks of demand planning

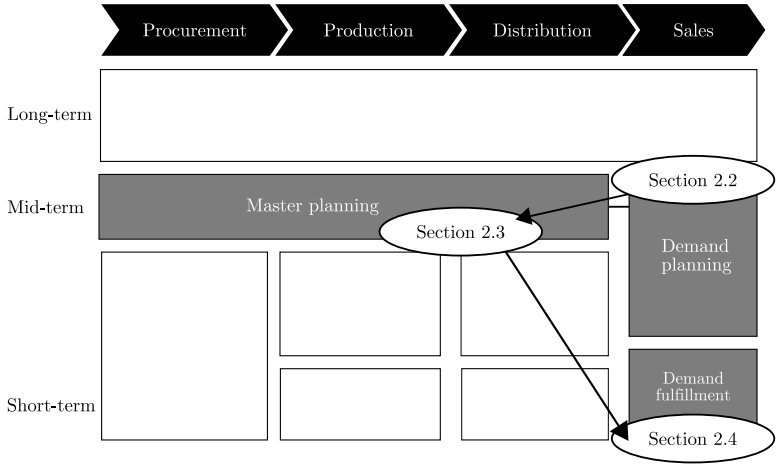


Figure 2.7. – Structure of the following sections

will be summarized. Afterwards, the major components of a demand planning system will be introduced:

- Demand planning **structures** (Section 2.2.2) refer to the planning and forecasting hierarchies which need to be managed within a company,
- demand planning **processes** (Section 2.2.3) comprise the key forecasting activities and
- demand planning **controlling** (Section 2.2.4) addresses procedures and measures to manage the quality and accuracy of the demand planning process.

These sections provide a general overview of demand planning. In many practical settings, demand planning structures have a hierarchical nature in the form of multi-stage customer hierarchies. Forecasting and demand planning in such hierarchical structures is rarely covered in textbooks and the standard forecasting literature. Hence, Section 2.2.5 will provide an introduction to hierarchical forecasting.

2.2.1. Objectives and Planning Tasks

Demand planning is often defined formally as the process of forecasting future customer demand (Kilger and Wagner, 2008, p. 133). It constitutes the first planning step at each supply chain planning level as its output affects the quality of all subsequent planning activities (Chen et al., 2007, p. 2269). Any errors and uncertainties which are present in demand planning are propagated and often magnified by all subsequent supply chain processes. Hence, the further up in the supply chain, the worse usually the planning

quality (Lee et al., 1997a). Bad planning has a particularly adverse influence on scheduling, on resource acquisition decisions and on the determination of resource requirements (Makridakis et al., 1998, p. 5).

Companies aim for exact forecasts to achieve long-, mid- and short-term objectives (see Eickmann, 2004): In the mid- and long-term, accurate forecasts help leveling out fluctuations in production and procurement planning. This reduces setup costs as well as direct and indirect inventory holding costs (e.g. due to obsolescence). In the short-term, accurate forecasts lead to improved service levels, shorter service times, a more flexible production and less exception management.

Good forecasting also requires catering to a variety of specific forecasting needs which have to be satisfied in the different functional areas of a company. Table 2.1, taken from Mentzer and Bienstock (1998), summarizes typical forecasting needs of different managerial functions. The overview highlights the requirements per corporate function in terms of level of granularity, horizon and interval after which forecast updates are required. This breadth of requirements underlines that not only good forecasting techniques but also adequate planning structures along multiple dimensions are necessary for successful demand planning.

Additionally, adequate controlling and forecast monitoring processes are another key component of demand planning to maintain the consistency and quality of the forecasts. To summarize, demand planning consists of three key components (Kilger and Wagner, 2008, pp. 133–135):

- **Demand planning structures**: Proper planning structures are necessary to handle the inputs and outputs, particularly along several key dimensions, including products, customers, location and time. As will be discussed in more detail in Section 2.2.5, aggregation and disaggregation functions are required to provide forecasts along these dimensions at different planning levels.

- **Demand planning processes**: This component refers to the actual preparation of the forecasts. After the input data has been collected, analyzed and possibly condensed, both statistical and judgmental forecasting methods are usually used to predict the future. Often a reconciliation of multiple forecasts from different sources into 'one number' is necessary.

- **Demand planning controlling**: In the spirit of a continuous improvement process, the quality of the resulting forecasts needs to be evaluated ex-post to trigger necessary adjustments both in the forecasting processes and in the planning structures. This requires the definition of basic forecast evaluation metrics. Furthermore, systems to calculate, to record and to present such performance indicators are necessary. Such forecast controlling systems also have to be complemented by suitable organizational processes to ensure incentives and responsibilities are set adequately (e.g. when using variable compensation schemes).

In the following, a more detailed discussion of each of those three components will be provided.

Managerial function	*Marketing*	*Sales*	*Finance / Accounting*	*Production: Long term*	*Production / Purchasing: Short term*	*Logistics: Long term*	*Logistics: Short term*
Needs	Annual plans for existing and new products, product changes, promotions, channel placement, pricing	Setting goals for salesforce, motivating sales agents	Projecting cost, profit levels and capital needs	Planning plant and equipment development	Planning specific production runs	Planning development of storage facilities and transportation equipment	Operational dispatching and transportation at product and location level
Level	Product (line)	Product by territory	Corporate, business unit, product line	Product (SKU)	Product (SKU)	Product by location (SKUL)	Product by location (SKUL)
Horizon	Annual	1–2 years	1–5 years	1–3 years	1–6 months	1 month–several years	1 day–1 month
Interval	Monthly–quarterly	Monthly–quarterly	Monthly–quarterly	Quarterly	Weekly–monthly	Monthly	Daily–monthly

Table 2.1. – Forecasting requirements of various managerial functions (Mentzer and Bienstock, 1998, p. 9)

2.2.2. Demand Planning Structures and Forecasting Hierarchies

A first issue is to clearly define appropriate data structures which hold the input and output data of the required forecasts. As exhibited in Table 2.1, most organizations require a series of interrelated forecasts which can be structured according to multiple *dimensions*. Miller (2002, p. 199) differentiates between eight typical forecasting dimensions. These are summarized in Table 2.2.

Dimension	Typical forecasting levels	Hierarchy
Product	End item, product family, product line, company, industry	x
Geography	County, province, state / country, continent	x
Sales organization	Sales territory, district, region	x
Planning horizon / time	Hour, day, week, month, quarter, year	x
Customer	Individual person, key account, customer segment	x
Sales channel	Retail, wholesale, online	
Company organization	Cost center, business unit, corporate	
Network location	Plant, distribution center	

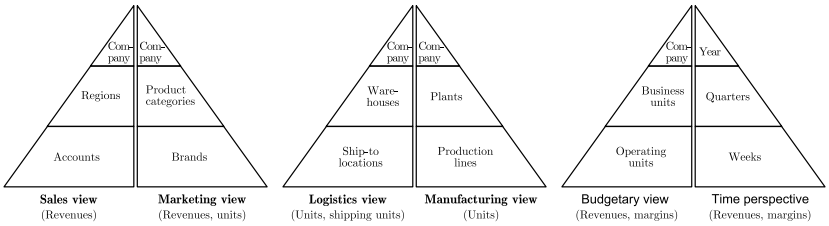
Table 2.2. – Typical forecasting dimensions (adapted from Miller (2002, p. 199))

The different forecasting levels within most dimensions form a hierarchy, i.e. these levels are related via generalization-specialization relationships (see also Section 2.1.2). For example, along the product dimension, inventory management and product scheduling require disaggregate forecasts for individual end items. Marketing, advertising and logistics planning, however, rather require forecasts at a product family level. Further forecasting needs exist at even more aggregate levels such as the product line level or even the entire company (for example for strategic network planning). In many industries, also the development of the entire competitive landscape (e.g. car units sold worldwide) needs to be predicted. At these higher levels of aggregation, the product dimension gradually blends into the measurement of general macroeconomic output.[18] In some dimensions, however, no such hierarchical relationships exist and only horizontal splits are possible, e.g. regarding forecasts per sales channel, company organization or network location.

Furthermore, users require forecasts in different units of measurement at many different levels of aggregation. Consider Figure 2.8, taken from Lapide (2006). It illustrates different views on the forecasting requirements of different functional areas of a company. While some of these stakeholders such as sales and marketing usually require forecasts in

[18] These latter forecasting needs are rarely addressed internally and the use of external forecasters such as (national) statistics bureaus or industry associations is more common, see Davidson and Prusak (1987).

terms of sales revenues (Figure 2.8a), others such as manufacturing and logistics require forecasts in terms of units (Figure 2.8b). Another important category of measurement units are budgetary units such as revenues, costs and margins (Figure 2.8c).



(a) Demand-side views **(b)** Supply-side views **(c)** Financial views

Figure 2.8. – Forecasting needs of stakeholders in various functional areas and hierarchy levels (Lapide, 2006, Fig. 2)

It is usually not feasible to prepare and manage individual forecasts at all levels in all forecasting dimensions. Rather, forecasts are managed in the form of *planning hierarchies* in which direct forecasts are only made at certain levels (see also Miller, 2002, Ch. 6.4). Forecasts for other planning levels can be obtained via aggregation or disaggregation operations. The definition of adequate planning structures to fulfill the various forecasting needs of the organization is thus a key strategic planning task of demand planning. A discussion of hierarchical forecasting will be postponed to Section 2.2.5.

A key structural requirement for hierarchical forecasting is a multi-dimensional database system to store, query and present forecasting inputs, forecasts and actual demand data. This functionality is typically provided by *online analytical processing* (OLAP) tools (for technical details, see Gray et al., 1997). OLAP tools allow navigating within the forecast and demand data. They provide the means to analyze data across multiple dimensions and hierarchy levels (e.g. roll-up and drill-down). In particular, they allow users to apply forecasting procedures at a certain aggregation level of the demand data and to propagate the forecast results to other levels via forecast aggregation and disaggregation. However, OLAP tools primarily serve operational purposes and do not provide planning functionality. For example, they do not provide recommendations how aggregation should be performed within the hierarchy (see Chen and Chen, 2004).

2.2.3. Demand Planning Processes and Forecasting Procedures

The process of forecasting, i.e. predicting the future, is at the heart of demand planning. There is an almost endless amount of literature on individual forecasting procedures. Hanke and Wichern (2009, pp. 2–3) introduced four main dimensions which may be used to classify individual forecasting procedures:

- **Type:** Quantitative vs. qualitative, or better: statistical vs. judgmental forecasts

- **Temporal span:** Long term vs. short term forecasts
- **Level of aggregation:** Position on the micro-macro continuum, i.e. item-level or corporate forecasts
- **Nature of output:** Point forecast (single best guess), interval forecast, density forecast (probability distribution for the future value)

The differentiation between quantitative and qualitative forecasts is often made in the following sense: Qualitative forecasts apply to situations where little or no quantitative information is available, but where sufficient qualitative knowledge exists to make an educated prediction (see Makridakis et al., 1998, p. 8). As both quantitative and qualitative forecasting methods result in a quantitative output, it is preferable to distinguish between **statistical** and **judgmental** forecasts.[19] A discussion of the key aspects of statistical and judgmental forecasts will follow shortly.

Note that some of the above dimensions are closely related. For example, forecasts for a long and very long horizon are often derived via judgmental forecasts as the necessary input data for statistical techniques are only rarely available (e.g. regarding long-term technology trends). The most typical forecasting needs relate to the short- and medium term and a single figure is the most widely used form of output.

In the following, first an overview of the key phases of the demand planning process will be given. Afterwards, the two most important of these steps, statistical and judgmental forecasting, will be characterized in more detail. In particular, the salesforce composite method will be introduced, a form of judgmental forecasts typically used for demand planning in multi-stage customer hierarchies.

Phases of the Demand Planning Process

In many companies, the actual forecasting activities are embedded into a formalized demand planning process. It consists of a regularly repeated sequence of up to five phases. This demand planning process is followed by demand planning controlling. Figure 2.9 contains a schematic overview (see for the following Kilger and Wagner, 2008, pp. 141–144).[20]

[19] Nahmias (2009, p. 56) makes a similar point and uses the terms 'objective' and 'subjective' forecasts.

[20] Kilger and Wagner (2008) in fact considered six phases of the demand planning process. They included at the fifth position a separate step for the planning of the so-called *dependent demand*. In an MTS environment, such a dependent demand is the amount of raw materials and components required to produce the predicted quantity of final items. This dependent demand is typically determined with the help of the MRP capabilities of ERP solutions (see Section 2.1.5). Often, however, the availability of raw materials and components will be constrained, rendering the production of the original final item demand forecast infeasible. Hence, taking care of the dependent demand is in fact a planning problem which needs to be handled separately from the demand planning and forecasting process, typically in the master planning step (see Section 2.3). Therefore, the step 'planning of the dependent demand' is omitted here. This perspective has also been taken in, e.g., Meyr (2012). He highlighted that the outcome of demand planning is simply an (unconstrained) 'forecast'. Once actual constraints are respected, it is more appropriate to refer to this output as a 'plan'; but planning is not in scope of forecasting.

Step 1: Preparation: In a first preparatory phase, historic demand data is gathered and—where necessary—corrected (e.g. to remove outliers or discontinued products).

Step 2-4: Core Forecasting Activities: The core forecasting activities consist of up to three phases. If sufficient quantitative information is available, a **statistical forecast** is computed and forms the basis for all subsequent steps. Statistical forecasts are often prepared with the implicit assumption that no systematic changes or departures from previously observed patterns are expected. Hence, statistical forecasting emphasizes predicting the future rather than explaining the past (Makridakis et al., 1998). If no or only limited quantitative information is available or if significant deviations from past observations are expected, a **judgmental forecast** may replace or adjust the statistical forecast. Both statistical and judgmental forecasting will be characterized in more detail below.

Often, closely related forecasts may be prepared by several managerial functions at the same time (e.g. by the corporate headquarter to complement the forecast of the regional sales organization). Furthermore, different statistical methods may be used in parallel to forecast the same data. Usually, the resulting forecast figures will differ and any discrepancies between different sources need to be removed in a **consensus forecasting** step. This step is necessary to synchronize the tasks and decisions within an organization or entire supply chain.

The combination of several forecasts—also known as *forecast pooling*—usually leads to an improvement of the overall forecast quality (Newbold and Granger, 1974). The underlying idea is portfolio diversification. Forecasting methods are usually affected to different degrees by structural breaks in a time series (i.e. they adapt to changes at different speeds). In many judgmental forecasts, private information is included which is unobservable to other forecasters and barely captured by statistical forecasts. A basic approach to forecast pooling is to simply calculate the average of the available forecast values to determine the one-figure forecast. More advanced procedures aim at finding optimal weights for pooled forecasts (e.g. see Mahmoud (1984), Clemen (1989), Granger (1989), Leitner and Leopold-Wildburger (2011) for reviews of the literature).

Forecast pooling is particularly popular in practice in the form of *collaborative forecasting*. This collaboration follows a formalized process which is often referred to as *sales & operations planning* (S&OP). S&OP combines the human knowledge from different

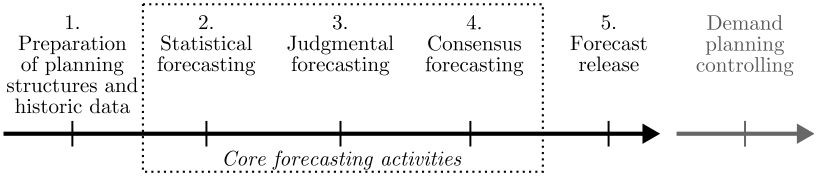


Figure 2.9. – Phases of the demand planning process (adapted from Kilger and Wagner, 2008, Fig. 7.7, p. 141)

sources such as sales, marketing, production and corporate headquarter with statistical methods.[21]

Step 5: Forecast Release: Finally, the resulting forecast can be released officially. It has to be shared within the supply chain, possibly also beyond the boundaries of legal entities in case of collaborative forecasting. Once the actual realizations of the forecast have been recorded, the quality of the forecast can be determined as part of the subsequent demand planning controlling activities. These analyses often yield important indications which may improve subsequent iterations of the demand planning process.

Statistical Forecasts

As mentioned above, statistical forecasts are applicable if sufficient quantitative information from the past is available which allows for an extrapolation into the future. Statistical forecasts can be roughly classified into *causal* and *time-series models*.[22] Time-series models only rely on past values for the prediction whereas causal models also include data from other sources to explain the future development of a value (Nahmias, 2009, p. 57).

In **causal models**, it is assumed that reliable dependencies exists between a *dependent variable* which is to be predicted and one or several *predictor variables* or *leading indicators* other than just time (Meyr, 2012, p. 73). An example for such a predictor variable is the 'outdoor temperature', which influences the dependent variable 'sales of ice cream'. If sufficiently accurate predictions of the outdoor temperature of the following day are available, this may allow making an accurate forecast of tomorrow's ice cream sales. Many causal models use the various techniques of *regression analysis* to identify stable relationships between the dependent and the predictor variables.

Time series models aim at identifying and exploiting patterns observed in past behavior to predict the future. Historic data is decomposed into different components. These usually include a baseline level, trend, seasonality factors and remaining random fluctuations. The simplest assumption is a stationary time-series which is often predicted with the help of the *simple exponential smoothing* method. This method produces a *one-step ahead forecast*, i.e. the immediately following period $t+1$ is being forecast. Given the stationary nature of the time-series, this forecast also holds for all following periods $t+2, \ldots, t+n$. The simple exponential smoothing forecast is given by

$$\hat{d}_{t+1} = \alpha d_t + (1-\alpha)\hat{d}_t, \quad 0 < \alpha \leq 1. \tag{2.1}$$

[21] For a general overview of S&OP, see Miller (2002, Ch. 6.5); for a more detailed discussion of the judgmental issues in S&OP, see Oliva and Watson (2009).
Combinations of judgmental and time-series-based forecasts were analyzed in a comprehensive study by Fildes et al. (2009). They found that judgmental adjustments of forecasts generated by statistical methods generally improved the overall forecast quality. This effect was greater in cases where the judgmental adjustment was large and furthermore, where the adjustment led to a *reduction* of the forecast figure.

[22] For a reference, see Kilger and Wagner (2008, p. 144).

In (2.1), $\hat{d}_{t+1}$, the forecast for period $t+1$, is a weighted average of the actual observation d_t in period t and the previous forecast $\hat{d}_t$. α is the smoothing constant which determines the weight placed on the previous realization d_t. A simple rearrangement of terms gives

$$\hat{d}_{t+1} = \hat{d}_t - \alpha(\hat{d}_t - d_t) = \hat{d}_t - \alpha \cdot \epsilon_t, \tag{2.2}$$

where ϵ_t is the observed one-step ahead forecast error for period t. In case the time-series follows a linear trend, *double exponential smoothing* or *Holt's method* may be used to forecast both the level and the slope of the time-series. Time-series with a seasonal pattern[23] may be predicted with the help of *Winter's method*. Introductions into these extensions of simple exponential smoothing have been provided in Makridakis et al. (1998, Ch. 4) or in Meyr (2012, Ch. 4.3). Note that for these time-series models which exhibit a trend and/or a seasonal pattern, it is also possible to generate multi-step ahead forecasts. However, the forecast error generally increases with the length of the step-ahead horizon, i.e. short-term forecasts tend to be more accurate.

Compared to these basic techniques based on simple exponential smoothing, the level of sophistication of modern time series forecasting approaches has increased considerably. It is not intended to provide a review here. Rather, Hyndman et al. (2008) as well as Gardner (1985, 2006) may be referenced for reviews of the current state of the art of time-series forecasting using exponential smoothing. Yet, in practice rather simple methods such as simple exponential smoothing (Equation (2.1) above) still prevail (e.g. see the survey by White, 1984, p. 5). Several authors have shown that such simple methods may perform sufficiently well in many practical situations, often matching the performance of many of the more complex alternatives (see Mahmoud, 1984; Fildes et al., 1998; Makridakis and Hibon, 2000).

Judgmental Forecasts and the Salesforce Composite Method

In many situations, historic information is insufficient to prepare a forecast and human judgment has to be used in forecasting. This may be necessary if there is information available in the present which suggests that a time-series will deviate from historic patterns. An example is the case of planned promotions or advertisement which will affect future sales (see Meyr, 2012, p. 73). Similarly, the historic time-series data may have been biased by singular events unlikely to be repeated in the future. In both cases, the application of statistical forecasting approaches will lead to erroneous predictions while human judgment may prove superior. Typical examples for structured forms of judgmental forecasts include *expert panels*, *surveys* or *salesforce composite* forecasts (Nahmias, 2009, p. 56). These judgmental or qualitative forecasts are not characterized by any explicit model, but rather knowledge, experiences and gut feeling (Caniato et al. (2005, p. 32), see also Armstrong (2001) for a broader perspective).

[23] These seasonal patterns can be of an additive or multiplicative nature.

Judgmental forecasts have to cope with a number of typical problems. Most result from the natural limitations of all human decision makers to process complex information. For example, Mentzer and Bienstock (1998, pp. 111–113) mention four essential aspects:

- Bias,[24]
- overconfidence of forecasters,
- anchoring effects (influences due to starting values, e.g. from quantitative forecasts) and
- the susceptibility of planners to political pressure (e.g. to make sales forecasts agree with company business plans).

Given these challenges, the individual judgmental forecasts in a company need to be coordinated, e.g. via an S&OP process. For example, over-budgeting is a typical issue with overconfident marketing and sales teams. Their customer representatives usually do not wish to disappoint customers (see, e.g. Celikbas et al., 1999; Larkin and Leider, 2011). Under-budgeting is often associated with the finance and controlling function, wishing to avoid or to reduce working capital requirements. Furthermore, production management generally has an incentive to understate forecasts in an attempt to increase the utilization rates of assets under their control, a key performance measure for operations staff (see Porteus and Whang, 1991). On top of this, executive management often requires a certain amount of revenue or has a margin target to meet investor relations (see Gilliland, 2002, p. 18). As these over- and under-budgeting effects may go in both directions, the exact impact depends on individual circumstances. Furthermore, judgmental forecasts are generally associated with a higher cost per forecast value than statistical methods.

Nevertheless, the use of judgmental forecasts is often unavoidable, particularly for demand planning in certain *business markets*. Generally, business markets have characteristics which are distinctly different from most *consumer markets*.[25] While consumer markets encompass individual consumers or households which buy or acquire goods and services for their personal consumption, products distributed in industrial markets are used in the production of other products or services which are sold, rented or supplied to third-party customers. Compared to consumer markets, order quantities in industrial markets are larger and quantity discounts are commonplace. Demand is often inelastic in the short term as it constitutes derived demand, i.e. the product is used in or for the production of other items. Furthermore, prices are often less stable than in consumer markets, for example regarding raw materials.

Many specialized industrial products require technically competent sales agents having good personal rapport with the customers. An adequate approach to forecasting in such markets is the *salesforce composite method*. In short, salesforce composite forecasting consists of relying on the views of the sales agents and their sales managers to determine a bottom-up outlook on individual products or on total sales (Cox, 1989, p. 307). This

[24] For a further discussion of bias in the context of forecast errors, see page 61.
[25] The following aspects have been summarized from Kotler et al. (2009, Ch. 8).

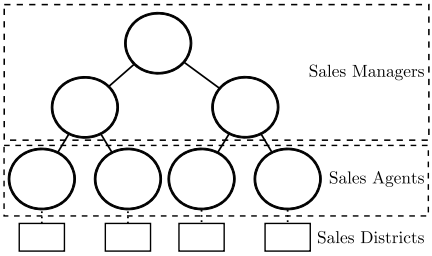


Figure 2.10. – Basic hierarchical sales organization

approach is also prevalent with apparel and fashion retailers, see Fisher et al. (1994) and Hausman and Thorbeck (2010).

The salesforce composite method implies that a hierarchical sales organization exists. *Sales agents*, positioned at the lowest level, can usually provide a fairly accurate perspective on the upcoming demands within their sales district or for their key accounts. Higher-level *sales managers* coordinate and supervise the sales agents and collect and aggregate the market data for the individual districts and segments. Usually, multiple layers of management are required as one manager can only supervise a limited number of subordinates efficiently.[26] A major advantage of the resulting hierarchy is its ability to closely scrutinize the market information as it is passed on through the forecasting system, being reviewed by several individuals with different personal vantage points (Weinstein, 1987, p. 453). A basic hierarchical sales organization is depicted in Figure 2.10.

The salesforce composite method ranges among the most widely used methods since it often yields very satisfactory results (e.g. see Dalrymple, 1987; McCarthy et al., 2006). However, it is not free from weaknesses. In addition to the general problems which apply to all judgmental forecasting techniques, Weinstein (1987) discussed three major issues which are often associated with salesforce forecasts:

- **Expertise and contagion errors:** When relying on personal judgment, two different types of error may occur: *Expertise errors* refer to individual errors of judgment which may occur at each hierarchy level in the sales organization. These individual errors are usually unrelated. Their impact on the aggregate error can be mitigated by having each agent only forecast a small share of overall demand (Staelin and Turner, 1973). *Contagion errors*, by contrast, can be introduced if some or all sales agents and managers rely on identical pieces of information, e.g. regarding the macroeconomic outlook, production capacities or promotional plans. This type of error usually has a much stronger impact on the aggregate error.

- **Loss and distortion of information:** Due to the aggregation process, some data may inadvertently be lost, e.g. regarding the level of uncertainty associated with each detailed forecast. A partial solution to this problem is to rely on several scenario

[26] This aspect is addressed in more detail later in Section 3.2.

forecasts. Furthermore, a close monitoring of key accounts may help to obtain early warnings of unexpected demand changes.

A greater problem is posed by sales persons actively distorting or biasing their forecasts or hiding additional (strategic) information (e.g. regarding material changes within one of their client organizations). These problems often arise either due to negligence or intentionally, e.g. to obtain personal advantages for individuals. While negligence may be addressed with training programs, intentional behavior requires a proper alignment of incentives.

- **Confusion between forecasts and objectives:** Lastly, (sales) forecasts often serve a dual purpose: On the one hand, they constitute key inputs into many subsequent SCP processes; on the other hand, they are used to set personal objectives for individual agents and managers. Objectives help to reduce fluctuations of output, to set expected norms of performance and can also provide motivation. Nevertheless, objectives and goals not necessarily reflect an appropriate prognosis of the future. However, the latter is usually required for good plans for the subsequent supply chain processes. The closer the relationship between forecasts and salesforce objectives, the more likely is also the introduction of bias (see previous point).

In addition to the above aspects, Kerkkänen et al. (2009) highlighted that salesforce composite forecasts may involve motivational problems. Sales people perceive selling to be more attractive and better rewarded than administrative tasks such as forecasting. Furthermore, White (1984, p. 39) mentioned that salespeople tend to be poor estimators when tasked with identifying long-term trends.

However, many of these challenges can be handled. On the one hand, salesforce composite forecasts are rarely used alone. Rather, aggregate salesforce composite forecasts are usually complemented by statistical (aggregate) forecasts (e.g. computed by the corporate headquarter). Hence, reconciliation between both data sources has to be performed to determine an overall single-figure forecast (Weinstein, 1987, p. 451). On the other hand, problems caused by game-playing behavior can be mitigated, at least to some extent, by properly designed incentive and compensation schemes. This last aspect will be discussed in more detail in Section 3.4.

2.2.4. Demand Planning Controlling and Forecast Error Measures

As frequently emphasized in the forecasting literature, forecasts are always wrong (e.g. see Nahmias (2009, p. 52), Meyr (2012, p. 67)). The extent to which a forecast $\hat{d}_{t,r}$ differs from the actual value d_t is an obvious measure of the forecast quality. Here, t refers to the time period of the event to be forecast and r is the period in which the forecast was prepared. $\hat{d}_{t,r}$ is the $t-r$-step ahead forecast and $\hat{d}_{t,r} - d_t = \epsilon_{t,r}$ is the associated forecast error (Kilger and Wagner, 2008, p. 150).

In the following, first an overview of popular forecast error and accuracy measures will be given. Then, the problem of biased forecasts will be addressed briefly. Lastly, a

few empirical results regarding the typical level of forecast accuracy in practice will be reported.

Forecast Error and Accuracy Measures

The primary objective of demand planning controlling is to reduce and manage the overall level of forecast errors (see Kilger and Wagner, 2008, p. 149). Lower forecast errors lead to reduced overall costs, e.g. due to lower safety stock requirements. This objective is pursued by providing adequate measures to track and analyze the evolution of forecasts, actual data and resulting forecast errors. An important component is the definition of adequate measures for forecast errors—or conversely, for the level of forecast accuracy. Some authors differentiate strictly between both types of measures. The term 'error' focuses more on the deviation from the actual whereas 'accuracy' rather considers the reverse, the level of agreement between forecast and actual. Nevertheless, both concepts capture the same issue: How 'good' is a particular technique in predicting a time series or an event. Therefore, the subsequent presentation will be simplified by not differentiating between error and accuracy measures, unless this is required by the context.[27]

Many different measures are available to quantify forecasting errors.[28] Most measures are suited for particular areas of application and may fail when applied in other, unsuitable areas. Reporting a measure for a single forecast error is rarely useful in practice. Rather, a summary statistic which provides an aggregate evaluation of the forecast accuracy over several data points is usually required. Typical dimensions include the aggregation over time (aggregation of the most recent n individual forecast errors), over different products and geographies (Kilger and Wagner, 2008, pp. 151–152). In the remainder of this section, the focus of the presentation will remain on the aggregation over time whereas a discussion of further aggregation dimensions will be postponed to Section 2.2.5.

A simple categorization of aggregate forecast error and accuracy measures for the dimension *time* has been suggested by Mentzer and Bienstock (1998, pp. 20–21) and Mentzer and Moon (2005). Employing their categories, a number of accuracy measures which are frequently used either in academia or practice have been summarized in Table 2.3.

The measures in the first group are all related to the basic forecast error $\epsilon_{t,r}$ and hence capture the differences between the forecast and the actual value in absolute terms. The ME comes with the often undesired side effect that positive and negative forecast errors cancel out. This deficiency is cured by the MAE or MAD by only considering the absolute value of the deviation, irrespectively of the direction. However, both the MAD and the ME put equal weight on all errors. Frequently, it is preferable to penalize large deviations more heavily. This can be achieved by the MSE, which corresponds to the variance of the forecast error over the time horizon considered (Kilger and Wagner, 2008, p. 151).

Neither of the measures in the first group does allow for an easy comparison between different time series as the measures are not standardized. This problem is countered by the measures in the second and third group. For example, the MAPE corresponds to a

[27] This perspective is also taken in Mentzer and Bienstock (1998).
[28] For a comprehensive overview, see e.g. Armstrong (1978).

Actual measures of forecasting accuracy		
Mean error	ME	$\frac{1}{n}\sum_{t=1}^{n}\epsilon_{t,r}$
Mean absolute error, mean absolute deviation	MAE, MAD	$\frac{1}{n}\sum_{t=1}^{n}\lvert\epsilon_{t,r}\rvert$
Mean squared error	MSE	$\frac{1}{n}\sum_{t=1}^{n}\epsilon_{t,r}^2$
Accuracy measures relative to a perfect forecast		
Mean absolute percentage error	MAPE	$\frac{1}{n}\sum_{t=1}^{n}\frac{\lvert\epsilon_{t,r}\rvert}{d_t}$
Mean absolute percentage accuracy	MAPA	$\frac{1}{n}\sum_{t=1}^{n}x_t$, with $x_t = \max\left(1-\frac{\lvert\epsilon_{t,r}\rvert}{d_t};0\right)$
Coefficient of variation of the root mean squared error	CV-RMSE, CV	$\frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\epsilon_{t,r}^2}}{\sum_{t=1}^{n}d_t} = \frac{\sigma}{d_t}$
Accuracy measure relative to a naïve forecasting technique		
Theil's U-statistic	U	$\sqrt{\frac{\sum_{t=1}^{n-1}\left(\frac{\epsilon_{t+1}}{d_t}\right)^2}{\sum_{t=1}^{n-1}\left(\frac{d_{t+1}-d_t}{d_t}\right)^2}}$

Table 2.3. – Frequently used forecast error and accuracy metrics, adapted from Mentzer and Bienstock (1998, pp. 20–21), Mentzer and Moon (2005) and Kilger and Wagner (2008)

relative assessment of the forecast error (or the accuracy) compared to the actual value. The problem with the MAPE metric is that it is not defined in case the actual demand realization equals zero. Therefore, Kilger and Wagner (2008) proposed using the MAPA metric which has the advantage of being defined even if $d_t = 0$ in one period. The third measure of the second group can be obtained in a similar manner as the MAPE. Rather than normalizing the MAD, in the CV-RMSE the square root of the MSE is used in the numerator. As stated above, the MSE corresponds to the variance of the forecast error; correspondingly, the square root of the MSE equals the standard deviation of the forecast error. Following Reddy (2011, p. 145), this normalized standard deviation of the forecast errors will simply be referred to as the *coefficient of variation* (CV), unless this would result in a disambiguation.[29] Note that it is possible to define a corresponding accuracy measure in the form of MAPA also for the CV.

Unfortunately, MAPE, MAPA and CV only allow for relative comparisons between different forecasting techniques. This means that it is only possible to determine whether one forecasting technique is better than the other, but apart from a comparison with a perfect forecast (i.e. the actual value), it is impossible to given an absolute quality assessment.

[29] Nevertheless, a proper usage of the concept of the coefficient of variation would require dividing the standard deviation by the mean. Given that the mean demand forecast error equals zero, the normalization here occurs with respect to the actual demand.

This lack of a standard is cured in the third category of accuracy measures. Here, the forecast error resulting from a certain forecasting method is compared to the forecast error from a naïve forecasting technique. As an example for measures of the third category, Theil's U statistic is given.[30] Like the MAPE and the CV metrics, it is not defined in case $d_t = 0$. Theil's U statistic compares how well a certain forecasting model performs against a naïve 1-step ahead prognosis. Such a naïve forecast simply results from using the most recent observation as the forecast, i.e. $\hat{d}_{t+1} = d_t$. If $U > 1$, the forecast model is worse than the naïve forecast. Clearly, values smaller than 1 are preferred, and a value of $U = 0$ corresponds to a perfect forecasting technique.

Biased Forecasts

Fildes and Kingsman (2011) have stressed that the above definition of the forecast error is actually the combination of two components: The first component is the inherent randomness in the process which generates the time-series. The second component corresponds to the error arising from not using the optimal forecasting model (e.g. by assuming a time series is stationary while in fact it follows a trend). It is only this latter component which leaves room to improve the quality of the forecast. Unfortunately, the generation process for most time-series is usually unknown in practical applications. However, separating the inherent randomness in the data generation process from the problem of choosing a suboptimal forecasting model permits to distinguish—at least conceptually—between several different aspects of forecast errors (Fildes and Kingsman, 2011, p. 487).

A typical such error type is *bias*, the systematic over- or underestimation of a variable. In practice, bias is due to safety thinking of the forecaster or triggered by incentive misalignments (see above). In the presence of bias, the individual forecast errors $\epsilon_{t,r}$ are either consistently positive or negative. This results in an ME which is strictly different from zero (Kilger and Wagner, 2008, p. 153). If both the magnitude and direction of bias do not change over time, the forecast can be adjusted manually to remove this distortion. Corresponding de-biasing schemes have been discussed, e.g. in Weinstein (1987) and in Utley (2011).

As judgmental forecasts are particularly susceptible to bias, statistical methods often yield better results in practice (Makridakis et al., 1998). Nevertheless, human judgment is often essential in many business contexts, e.g. for the salesforce composite method. The introduction of appropriate compensation schemes may provide important incentives to mitigate biased and misrepresented forecast reports in such situations. An overview of suitable schemes will be presented in Section 3.4.

Forecast Accuracy in Practice

In the absence of bias, there are two dominant factors which affect $\epsilon_{t,r}$ (Kilger and Wagner, 2008, p. 150),

[30] The measure U given here is sometimes also referred to as $U2$. Note that Theil proposed two error measures, but the historically older $U1$ has serious flaws and should be discarded (see Bliemel, 1973).

- the *lead time* $t - r$ between forecast and actual value and
- the *forecast granularity*, i.e. the level within the demand planning hierarchy at which $\hat{d}_{t,r}$ has been determined.

While longer lead times in general entail a higher forecast error,[31] less granular forecasts usually lead to lower forecast errors (see also Section 2.2.5). Furthermore, when forecasting demand, certain characteristics of the market environment have a strong impact on the magnitude of the resulting forecast errors, e.g. the current market dynamics (e.g. market growth), the competitive situation and the phase in the product life-cycle (e.g. new product introduction vs. stable, established product). Forecast errors are higher in volatile than in stable markets.

Empirical observations regarding the level of forecasting accuracy achieved in practice depend significantly on the individual industries, markets and company particularities and are therefore difficult to generalize. Unfortunately, it is almost impossible to obtain independent inter-company and cross-industry assessments of forecast error performances. The only practicable way is to resort to (unverifiable) surveys. Three major cross-industry surveys have been reported in the literature. To conduct each of these surveys, managers have been asked to assess their familiarity with different forecasting methods, to state their level of satisfaction and to disclose the observed forecasting errors.

The three surveys by Mentzer and Cox (1984), Mentzer and Kahn (1995) and McCarthy et al. (2006) use a similar survey design and have been conducted roughly ten years after their immediate predecessor, allowing for comparisons over time. For later reference, their main results regarding forecasting accuracy as well as some similar survey results among British managers by Fildes and Beard (1992) are stated in Table 2.4. More specifically, the table lists MAPE values at different levels of aggregation and for different lead times. The first data series, abbreviated by MC84, represents results of the first survey by Mentzer and Cox (1984) whereas the third and fourth data series are the subsequent survey updates by Mentzer and Kahn (1995) (MK95) and by McCarthy et al. (2006) (MC06). The second data series, abbreviated by FB92, is from a British study by Fildes and Beard (1992). These values have been referenced frequently in the literature (see e.g. Wacker and Lummus (2002), Fildes et al. (2009)).

As noted by McCarthy et al. (2006), the level of overall forecasting accuracy appears to have dropped since the mid-1980s at almost all hierarchical levels. Independent of the temporal span of the forecasts, accuracy is generally highest at the corporate level and decreases at more disaggregate levels such as SKU or SKU per location. Note further that forecasts covering a period of three months up to two years are generally less reliable than short-term forecasts over less than three months.[32]

[31] Empirically, this holds when comparing short- and mid-term forecasts. As indicated below in Table 2.4, forecast accuracy in practice seems to improve again in the longer run ($>$ 2 years). It is beyond the scope of this thesis to discuss possible reasons for this observation.

[32] However, the two more recent studies (MK95 and MC06) indicated that for more distant periods of time (more than 2 years) overall forecast accuracy at the SKU level is higher than at the mid-term level (3 months up to 2 years). Neither of the author teams provided an explanation for this phenomenon.

Overall, the values in Table 2.4 give a rough indication of the accuracy which can be achieved in practice. In particular, the forecast error at the SKU per location-level can be expected to be roughly in the range 13–40%. In the numerical experiments reported in Chapter 5 for the DMC problem with forecast errors, values from this range will be employed to generate realistic forecast error values.

Horizon	Short ($\leq$ 3 months)				Medium ($\leq$ 2 years)				Long ($>$ 2 years)			
	MC84	FB92	MK95	MC06	MC84	FB92	MK95	MC06	MC84	FB92	MK95	MC06
Corporate	7	–	28	29	11	–	14	16	18	–	12	11
Product line	11	12	10	12	16	12	14	21	20	20	12	21
SKU	16	16	18	21	21	20	21	36	26	26	14	21
SKU by location	–	–	24	34	–	–	25	40	–	–	13	

Table 2.4. – Forecast accuracy survey results (MAPE, in %) by level and time horizon (based on McCarthy et al., 2006); MC84 = Mentzer and Cox (1984), FB92 = Fliedner and Mabert (1992), MK95 = Mentzer and Kahn (1995), MC06 = McCarthy et al. (2006)

2.2.5. Hierarchical Forecasting

Given the ongoing trends towards globalization (more demand and supply markets), customer orientation (more customer segments) and increased product differentiation (more product varieties),[33] many organizations now require enormous amounts of different demand- and supply-side as well as financial forecasts. Already 20 years ago, Fildes and Beard (1992) indicated that 10,000 appeared to be a common figure for the number of individual product forecasts which many companies had to handle in their production and inventory forecasting systems. From today's perspective, this is likely to be a rather conservative estimate.

Given these large numbers of forecasts, the costs and effort associated with maintaining individual forecasts for all relevant time series usually cannot be justified (Gross and Sohl, 1990). Rather, hierarchical forecasting schemes have been suggested as an alternative. Not only do they result in a reduced forecasting effort, but also often lead to a better forecasting performance.

The following discussion of hierarchical forecasting consists of five paragraphs which are organized as follows: After introducing the essential concepts of direct and derived forecasts, the benefits of aggregate forecasts will be analyzed. Then, appropriate strategies to generate disaggregate forecasts at lower levels of planning will be investigated. Afterwards, the prior discussion of demand planning controlling will be extended to hierarchical forecasting situations. This section will close by briefly discussing a suggestion from the literature regarding the optimal design of demand planning and forecasting hierarchies.

[33] See also the introduction.

Direct and Derived Forecasts

In discussing hierarchical forecasts, a fundamental differentiation needs to be made between *direct* and *derived* forecasts (Theil, 1954). Consider the problem of deriving a statistical demand forecast for a particular product in a certain market:

Direct forecast Only the historical demand data of the given product in the given market is used to prepare the forecast.

Derived forecast The forecast is calculated from other existing forecasts, e.g. forecasts for other products or forecasts from other markets.

Derived forecasts are typical of hierarchical forecasting situations. On the one hand, the m direct forecasts from a lower hierarchical level may be combined via *bottom-up* (BU) summation, yielding an aggregate derived forecast. On the other hand, a directly made higher-level forecast may be *prorated* in a *top-down* (TD) manner to result in a number of lower-level, derived forecasts (Shlifer and Wolff, 1979).

For example, as depicted earlier in Figure 2.8a, the marketing function usually needs sales volume forecasts at the level of individual brands or at the level of individual products (SKU) for a particular market. In addition, also aggregate product category level forecasts are required. Rather than preparing a separate forecast at the aggregate level, one may simply sum the m different forecasts at the SKU per location-level. This bottom-up forecasting approach is depicted in Figure 2.11a.

As introduced in Section 2.2.2, many other aggregation dimensions are possible besides a product-based hierarchy, e.g. time or geography.

Conversely, lower-level forecasts per individual product may be obtained by disaggregating an existing aggregate forecast at the product category level in a top-down manner, as shown in Figure 2.11b.

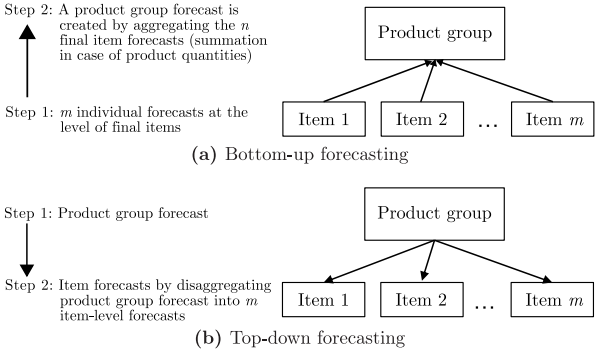


Figure 2.11. – Basic hierarchical forecasting methods (Miller, 2002, Fig. 6.8)

Aggregate Forecasts: Direct or Derived Bottom-up Approach

An extensive discussion has emerged over the past 60 years, examining the various arguments for and against individual hierarchical forecasting approaches. Two directions of this discussion can be differentiated (see Fliedner, 1999, p. 1136):

- In the older literature, predominantly in the economics domain, the focus was placed on strategies to forecast an aggregate value. The key question is whether an aggregate relation can better be forecast by a direct forecast at the aggregate level or by combining the results of several more granular forecast models in a bottom-up manner. Typically, advanced forecasting methods are used in this context. An outline of this discussion will be given below.

- More recently, hierarchical forecasting has also been adopted in the business management literature. Here, the focus has been placed on forecasting strategies both for the aggregate and for the disaggregate level. A typical area of application is supply chain planning where both aggregate and disaggregate forecasts need to be prepared. In such settings, rather simple forecasting methods such as exponential smoothing prevail. The research contributions in this stream of the literature not only consider the question of predicting aggregate values by either a direct or a derived bottom-up approach, but also address the reverse question whether disaggregate forecasts should better be prepared by a direct or a derived top-down approach.

The first stream of literature starts with Theil (1954), who argued in favor of direct forecasts at the lower level using specific micro-models. His modeling environment contains the strong assumption that the parameters of the detailed forecasting models are known with certainty. He then showed that a direct forecasting model at the aggregate level introduces bias and forecasting errors. However, in (econometric) practice, as contended by Grunfeld and Griliches (1960), forecasting models at the micro-level are usually less accurately specified than at the macro-level, allowing for a direct macro forecasting model to better capture relationships at the aggregate level. Lütkepohl (1984) sided with this view and showed that the direct approach at the aggregate level is superior if the parameters of the disaggregate forecasting models are unknown and have to be estimated. Both Orcutt et al. (1968) and Edwards and Orcutt (1969), arguing in favor of a bottom-up approach, pointed out that using only aggregate (econometric) data to construct an aggregate forecasting model may entail a substantial information loss.

Dunn et al. (1971) and Dunn et al. (1976) were among the first to report empirical evidence. They gave practical illustrations from the telecommunications industry and showed that the summation of forecasts obtained from detailed forecasting models per geographical demand region can predict overall demand at an aggregate level consistently better than a direct forecasting model at the higher level. More empirical support for the derived approach has been provided by Kinney (1971) and Collins (1976). They found that aggregate corporate performance can better be predicted by first forecasting detailed-level earnings data and aggregating. A similar result in favor of the derived approach has been

reported by Foekens et al. (1994). They analyzed retail promotion effects at different levels of aggregation such as store, chain or the market level. In particular, they established that detailed forecasting models can better accommodate heterogeneity among the lower-level series which otherwise gets lost when forecasting at higher levels of aggregation. Zellner and Tobias (2000) and Weatherford et al. (2001) arrived at similar conclusions when forecasting gross domestic product (GDP) data and data for hotel revenue management, respectively.

While the empirical results tend to point towards a superiority of the derived bottom-up approach, there is no general consensus whether a direct or derived approach performs better. Rather, as pointed out by Wei and Abraham (1981, p. 1340), there is no "one-way unconditional inequality" which could specify whether a direct aggregate or a derived aggregate linear forecast[34] performs better.

The key challenge thus consists of identifying specific conditions under which either the direct or the derived approach to forecast an aggregate time-series works better. This call has been taken up by some of the more recent contributions. These younger publications also put a stronger emphasis on the closely related problem of specifying conditions under which a disaggregate time-series can better be predicted by either a direct approach or by top-down proration.

In this context, it is important to note that the phrasing in many publications gives the false impression that the actual problem is whether a top-down or a bottom-up approach is better.[35] Technically, these two approaches are not alternatives. Rather, the key problem is—as indicated above—whether a direct or a derived forecast is better suited to make predictions *at a given hierarchy level*.[36]

In the following, the benefits of direct and of derived forecasts will be investigated with the help of a simple formal model. Consider the problem of forecasting the demand of m product items which form a particular product family. This setting constitutes a simple two-level hierarchy. For period t, denote the aggregate demand at the family level by D_t. D_t is the sum of m disaggregate time-series and $d_{i,t}$ will be used to denote the (actual) demand of item i in period t. The entire family demand is then given by

$$D_t = \sum_{i=1}^{m} d_{i,t} \tag{2.3}$$

For later reference, note that $D_t > 0$ if at least one of the $d_{i,t} > 0$, $i = 1, \ldots, m$.

Assume that a forecast for period $t + 1$ is required. Using the above concepts, there are two ways to generate an aggregate forecast $\hat{D}_{t+1}$. First, using a direct forecast-

[34] In a linear forecast, the individual historic observations of a particular time-series are related only by linear functional relationships. For example, this is true for the exponential smoothing or moving average forecasting models.

[35] A few examples where already the title is misleading: Schwarzkopf et al. (1988), Kahn (1998), Wanke and Saliby (2007), Viswanathan et al. (2007), Widiarta et al. (2008, 2009).

[36] This applies if only a lower and an upper hierarchy level are considered. However, if there are three or more hierarchy levels, the middle level may be forecast by a direct forecast, by top-down proration or by a bottom-up summation. The emerging problem of *consistency* will be addressed shortly.

ing approach, one may resort to the last n historical observations of aggregate demand $D_t, D_{t-1}, \ldots, D_{t-n+1}$ to produce a forecast $\hat{D}^{dir}_{t+1}$, e.g. via exponential smoothing or any other time-series-based procedure. Here, the superscript ^{dir} indicates a direct forecast at the aggregate level. Alternatively, one may produce individual forecasts $\hat{d}_{i,t+1}$ for each of the m disaggregate time-series and use summation to determine a derived bottom-up (BU) forecast for the aggregate family demand, i.e.

$$\hat{D}^{BU}_{t+1} = \sum_{i=1}^{m} \hat{d}_{i,t+1}. \tag{2.4}$$

Similarly, there are two ways to determine disaggregate forecasts at the item-level. In addition to the direct forecast $\hat{d}_{i,t+1}$, there is also the derived, top-down approach by prorating a direct aggregate forecast. For example, by multiplying the aggregate direct forecast $\hat{D}^{dir}_{t+1}$ by the ratio of the disaggregate demand of item i to the aggregate demand in period t, $\left(\frac{d_{i,t}}{D_t}\right)$, the top-down (TD) derived forecast will be obtained as

$$\hat{d}^{TD}_{i,t+1} = \hat{D}^{dir}_{t+1} \cdot \frac{d_{i,t}}{D_t}, \quad i = 1, \ldots, m, \quad \text{if } D_t > 0. \tag{2.5}$$

In many situations, the direct and the derived forecast do not coincide. At the aggregate level, $\hat{D}^{dir}_{t+1} \neq \hat{D}^{BU}_{t+1} = \sum_{i=1}^{m} \hat{d}_{i,t+1}$ and similarly, at the lower level $\hat{d}^{TD}_{i,t+1} \neq \hat{d}_{i,t+1}$.

The principle of *vertical consistency* implies that the sum of the forecasts made at a lower level needs to equal the forecast value at the upper level. This is seldom the case. In such situations, processes for a proper reconciliation of forecasts across levels are required. An early contribution in this respect was the *pyramid principle* by Muir (1979). He proposed a forecasting system in which in a first step, the most recent demand data at the item level is collected. Then, statistical forecasts at the item-level will be produced for all SKUs. These will be complemented by judgmental (management) forecasts for selected SKUs. Aggregate values at the product group level will be obtained by summing, on the one hand, the most recent demand information and, on the other hand, the forecast values per SKU group.[37] The group-level figures of the most recent demand information then form the basis of the statistical forecast at this hierarchical level, some of which may again be complemented by judgmental forecasts.

The summed group-level forecasts will then be used for a subsequent disaggregation or 'downward-forcing' operation to obtain the final item-level forecasts. These must sum either to the statistical or to the judgmental forecast of their respective group. Muir (1979) did not give guidelines regarding the choice of the downward-forcing method and only stated several methods to obtain item-level forecasts. For a description of an early practical implementation of a forecasting system based on the pyramid principle see Kuehne and Leach (1982). Forecast reconciliation is now a standard functionality of many planning

[37] This is done separately for the statistical and the management forecast; if no judgmental forecast exists for a particular item, the statistical forecast will be used instead.

software tools, for example see Bartsch and Bickenbach (2002, pp. 128–130) or Kilger and Wagner (2008, p. 140).

In the following, it will be highlighted why a direct forecast at the aggregate level is usually superior to a derived bottom-up forecast for the case of stationary demand time-series. For this, it will be shown that aggregation leads to a smaller relative forecast error. To obtain simpler analytical expressions, only two products ($i = 1, 2$) at the lower-level will be considered, hence $m = 2$.

First, assume that the demand for each product i ($i = 1, 2$) is stationary; the mean values $\bar{d}_i$ can thus be used to make predictions. The forecast error associated with this method can be measured with the MSE. The CV-RMSE (see Table 2.3), or CV for short, is the corresponding normalized forecast accuracy measure:

$$CV_i = \frac{\sigma_i}{\bar{d}_i}. \tag{2.6}$$

A demand-weighted average of the two individual values CV_i allows characterizing the dispersion of the joint relative forecast error of both disaggregate time-series for the bottom-up approach. The corresponding expression amounts to

$$CV_D^{BU} = \frac{\bar{d}_1}{\bar{d}_1 + \bar{d}_2}\frac{\sigma_1}{\bar{d}_1} + \frac{\bar{d}_2}{\bar{d}_1 + \bar{d}_2}\frac{\sigma_2}{\bar{d}_2} = \frac{\sigma_1 + \sigma_2}{\bar{d}_1 + \bar{d}_2} = \frac{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2}}{\bar{d}_1 + \bar{d}_2}. \tag{2.7}$$

Now consider the time-series of the aggregate demand $D = d_1 + d_2$ with a mean value of $\bar{d}_1 + \bar{d}_2$. The variance of the forecast error corresponds to

$$MSE_D = MSE_{(d_1+d_2)} = \sigma_1^2 + \sigma_2^2 + \mathrm{cov}\,(d_1, d_2) = \sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2 r_{d_1,d_2}. \tag{2.8}$$

Here, $\mathrm{cov}\,(d_1, d_2)$ is the covariance between both time-series and r_{d_1,d_2} is the corresponding correlation coefficient. Now the relative dispersion of the aggregate demand process can be specified, yielding

$$CV_D = \frac{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2 r_{d_1,d_2}}}{\bar{d}_1 + \bar{d}_2}. \tag{2.9}$$

Clearly, unless both lower-level time-series are perfectly positively correlated with $r_{d_1,d_2} = 1$, the relative dispersion of the aggregate forecast error is strictly smaller than the demand-weighted dispersion of the sum of the two disaggregate forecast errors,

$$CV_D < CV_D^{BU} \quad \text{if } r_{d_1,d_2} < 1. \tag{2.10}$$

Therefore, a direct forecast at the aggregate level is usually superior to a derived forecast as it is associated with a lower relative forecast error.[38]

This result has been largely confirmed by Fliedner (1999) in a number of simulation experiments for the above setting with two demand streams. He generated forecasts at the disaggregate and at the aggregate level via exponential smoothing and moving av-

[38] A similar argumentation can be found in many SCM textbooks, e.g. in Vollmann et al. (2005, p. 40).

erages and tested different degrees of correlation between the two lower-level demand streams. Irrespective of the forecasting technique and the level of correlation between the two demand streams, the direct forecasting approach performed better than the derived approach. However, the performance gap between both approaches was smaller for highly correlated demands (positive and negative correlation) than for not or only mildly correlated lower-level demand streams. As the first conclusion is in contradiction to some earlier studies, Fliedner (1999) argued that other results may be possible if the number of disaggregate demand streams becomes larger (and thus the correlations become more complex), if the disaggregate streams exhibit trends and seasonal variations or if more sophisticated forecasting techniques are used. Hence, the performance of hierarchical forecasting is not only dependent on the hierarchical level, but also on the type of the aggregation which is used.

Disaggregate Forecasts: Direct or Top-down Approach

The analog problem to the issue discussed above is the question whether forecasts at the lower-level should better be prepared using a top-down derived or a direct approach. Strijbosch et al. (2008) stated four factors which seem to have a decisive influence whether a top-down forecasting approach is superior, but gave no empirical support for the claims. In the following, several key findings from the existing literature will be related to the statements of Strijbosch et al. (2008). These four factors consist of

1. the criteria used to form aggregates based on similar lower-level series,
2. the accuracy of the aggregate forecast,
3. the magnitude of the individual disaggregate time-series in relation to the aggregate time-series and
4. the accuracy of the factors used in prorating the aggregate forecast.

Criteria to Form Aggregates: Traditionally, it has been suggested in the literature to form product families by grouping items with similar demand patterns (see also the discussion on hierarchical planning in Section 2.1.2). This homogeneity criterion aims at preserving the common demand pattern at the aggregate level to exploit it in the forecasting process (see, e.g., Lapide, 1998). This approach often leads to the claim that item-level demand time-series should be correlated positively at the aggregate level (see, e.g., Fliedner, 1999, p. 1146). A simulation study of this problem has been reported by Chen and Blue (2010) who studied direct and derived forecasting approaches and who investigated both the autocorrelation and the cross-correlation of two demand time-series. They observed that the top-down forecasting approach is beneficial if the aggregated demand is very predictable due to a strong positive autocorrelation component. The authors found that the predictability at the aggregate level in their model setting benefits from a positive correlation of the two demand time-series.

However, the argumentation in the previous paragraph has suggested that it is generally negative correlation which will lead to lower forecast errors at the aggregate level. The apparent contradiction between both perspectives has been discussed by Chen and Boylan (2009). They favored negative correlation as the key criterion to form aggregates, and their argumentation focused on the types of forecasting models used. If the *same forecasting model* is used for the disaggregate and the aggregate time-series (e.g. simple exponential smoothing with same weights), they argued that it is in fact negative correlation between the demand time-series which reduces variability at the aggregate level, and this favors the derived top-down approach (Chen and Boylan, 2009, p. 176).

However, this no longer holds if different forecasting models are used. In fact, this was the case in the Chen and Blue (2010) study. They exploited the autoregressive nature of the demand streams and thus used *different forecasting models* for the two lower-level demands and the aggregate demand (due to different values for the autoregressive coefficients). Following Chen and Boylan (2009, p. 176), "the more consistent the [forecasting] model forms are, the more this favors the grouping approach; and consistency of [forecasting] model forms is associated with positive correlations between series, not negative correlations."[39]

Moreover, positive correlation is a bad indicator whether series follow the same model, as positive correlation may also be incurred by a trend component (Chen and Boylan, 2009, p. 176). This may explain the results of the experimental research results by Fliedner (1999) which have been reported in the previous paragraph—he actually found both types of correlation (positive and negative) to result in a superior forecasting performance at the aggregate level.

Overall, the choice of aggregation criteria remains a challenging task as unequivocal criteria have not yet been reported in the literature. For example, when testing different criteria to group item-level forecasts into an aggregate family-level forecast, Fliedner and Lawrence (1995) found some empirical confirmation that sophisticated clustering approaches do not perform better than random clusterings.

Accuracy of the Aggregate Forecast: The second factor appears to be more important. Generally, proponents of a top-down approach argue that lower-level data is often more error-prone and more volatile (e.g. see Zotteri et al., 2005). An aggregate time series is more stable, making forecasts at the aggregate level more reliable which can then easily be broken down to the lower level. In the particular in case of intermittent demand at the lower levels, a top-down approach has been found useful by Moon et al. (2012) who reported a case study involving military spare parts. Dekker et al. (2004) presented evidence from two case studies for large Dutch wholesalers involving products with highly seasonal demand. They showed that seasonal factors can be determined more precisely at the aggregate product group level and that a top-down strategy thus performs better.

[39] The two inserts were added by the author of this thesis.

Magnitude of Disaggregate Values: Schwarzkopf et al. (1988) stressed that although the top-down approach reduces the effect of random errors on the disaggregate (derived) forecasts, the proration may introduce a complex interaction between bias and outlier effects. This effect was found to be particularly strong if the lower-level time-series are of similar magnitude and thus represent similar proportions of the family demand. By contrast, Gordon et al. (1997) established that the top-down approach is superior if detailed-level demands are roughly of equal size or if detailed-level demands are negatively correlated. Similarly, Fliedner and Mabert (1992) found the top-down approach to perform better for product families which were formed by items with similar volumes (i.e. where the individual components had comparable proportions of the aggregate) than for other clustering criteria such as seasonality or forecast performance.

Accuracy of Prorating Factors: In the few analytical contributions on the performance of top-down forecasting schemes, for example by Widiarta et al. (2007, 2008, 2009), the simplifying assumption was made that the prorating factors, i.e. the expected value of the share $p_{i,t} = d_{i,t}/D_t$ of time-series i with respect to the family time-series is known with certainty. However, this is rarely the case in practice, and these prorating factors need to be estimated.

Gross and Sohl (1990) tested a large variety of such factors experimentally. They found that simple sample averages over the last n periods lead to satisfactory results. In particular, the two most promising candidates were the *proportional-mean* and the *mean-proportional* factor. Using the previous notation, the proportional-mean is given by

$$p_i = \frac{1}{n}\sum_{t=1}^{n} x_{i,t}, \qquad \text{with } x_{i,t} = \begin{cases} \frac{d_{i,t}}{D_t}, & \text{if } D_t > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.11}$$

The mean-proportional factor is

$$\bar{p}_i = \frac{\frac{1}{n}\sum_{t=1}^{n} d_{i,t}}{\frac{1}{n}\sum_{t=1}^{n} D_t}, \qquad \text{if } \sum_{t=1}^{n} D_t > 0. \tag{2.12}$$

Chen et al. (2008) showed that the proportional-mean factor p_i minimizes the sum of squared differences between $d_{i,t}/D_t$ and p_i. They also suggested another factor,

$$p_i^* = \frac{\sum_{t=1}^{n} d_{i,t} \cdot D_t}{\sum_{t=1}^{n} D_t^2}, \qquad \text{if } \sum_{t=1}^{n} D_t^2 > 0, \tag{2.13}$$

and proved that it minimizes the sum of squared differences between $D_t \cdot p_i^*$ and $d_{i,t}$. They found that p_i^* is more accurate than p_i and $\bar{p}_i$.

Note that the expressions (2.12) and (2.13) are not defined if the respective denominators equal zero. However, these cases will only occur if all products in the group are characterized by intermittent demand, with long periods of no demand at all. But if the

denominators equal zero, also the numerators must equal zero, due to Equation (2.3). Should such a situation arise, it is convenient to require that $\bar{p}_i = 0$ and $p_i^* = 0$ must hold.

The problem with these rather static factors p_i, $\bar{p}_i$ and p_i^* is that they do not capture characteristics of the individual time-series such as trend or seasonality which lead to different proportions over time. Athanasopoulos et al. (2009) suggested forecasting the proportional factors by producing independent forecasts for all lower-level time series and then calculating the proportion of each individual forecast to the aggregate, i.e. to determine $\hat{p}_{i,t} = \hat{d}_{i,t} / \sum_{j=1}^{n} \hat{d}_{j,t}$. Then, they used this factor to prorate an aggregate forecast, i.e. they calculated $\hat{p}_{i,t} \cdot \hat{D}_t$. Obviously, this approach based on forecast proportions introduces significantly more effort, particularly since a simple direct forecast is already being produced in this process.

Another method to update the proportional factors is to use exponential smoothing to dynamically adjust the weights of each individual time-series over time. This way, the rise and decline of the demand can be captured better for individual products as they pass through different stages of their product life cycle. The problem here, as pointed out by Chen et al. (2008), is that the optimal smoothing coefficient needs to be updated regularly. This involves a significant computing effort per time-series to find the smoothing coefficient which minimizes the sum of squared differences. As an alternative, the authors suggested a dynamic updating scheme for the weighting factors based on the sample one-lag autocorrelation statistic. This scheme simplifies the search for the optimal smoothing coefficient considerably. The overall dynamic updating approach was found to perform well using demand data from the semiconductor industry.

To conclude this discussion, it can be inferred that the superiority of a direct or a derived forecasting approach depends on a large number of situation-specific factors. These include the variability of the lower-level demand streams, their correlation, the relative magnitude of the individual shares and the accuracy of available forecasting techniques both at the lower and at the aggregate level. More specifically, the following guidelines have emerged:

- Forecasts at aggregate level: Preparing a direct forecast at the aggregate level is often beneficial in hierarchical supply chain planning, particularly if the lower-level time-series are negatively or mildly positively correlated.

- Forecasts at lower level: Top-down approaches seem to work well for positively correlated time series with only limited variability or if the forecasting accuracy is significantly higher at the aggregate level, e.g. in case of highly seasonal time-series.

While these previous discussions have mainly focused on situations involving only two hierarchical levels, most practical applications of hierarchical forecasting involve multiple hierarchy levels. For these situations, many authors have also suggested that a combination of direct and derived approaches—both bottom-up and top-down—can be beneficial (e.g. see Schwarzkopf et al., 1988; Kahn, 1998). This is often referred to as a *middle-out* approach where a direct forecast is made at a particular intermediate level and disaggregation is used to determine lower-level forecasts and aggregation yields forecasts for more

aggregate levels. Lo et al. (2008) presented a practical example applying hierarchical forecasting to the LCD monitor market. Using a three-level planning hierarchy (level 0: all products, level 1: clustering by product size, level 2: clustering by size and by region), they found that a middle-out approach performs best.

Zotteri et al. (2005) stressed that one of the most crucial decisions in hierarchical forecasting is the choice of the hierarchy level at which the direct forecast will be made. They argued that the overall ability to produce a good forecast is determined by two major factors. To use their wording, planners have to make a trade-off between *managing* variability and *capturing* variability. This trade-off is sketched in Figure 2.12: On the one hand, at a high level of aggregation, variability is lower and thus easier to manage, for example when determining parameters of (aggregate) forecasting models which are usually more accurate (dotted line). On the other hand, one loses the ability to exploit heterogeneity at an aggregate level. Pro-rated forecasts usually do not capture the heterogeneity which exists at lower level of aggregation (dashed line). In addition to the hierarchy level, it is also the choice of the criteria used to form the aggregates which has a significant impact on the ability to capture and to manage variability. Kalchschmidt et al. (2006) have given several case studies which illustrate the benefits on forecasting performance by choosing adequate criteria for the aggregate clusters.

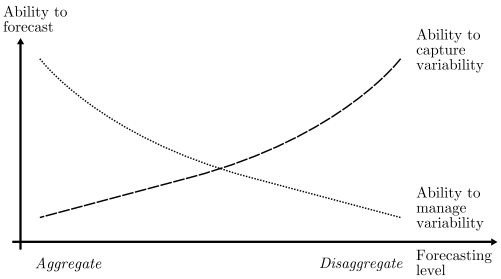


Figure 2.12. – The ability to forecast at different aggregation levels (Zotteri et al., 2005, Fig. 3)

Controlling of Hierarchical Forecasts

Demand planning hierarchies not only require an aggregation and disaggregation of forecasts, but also a calculation of forecast accuracy or error measures at different levels of aggregation. While the following brief discussion will be limited to simple accuracy measures at aggregate levels, Flores and Wichern (2005) gave a broader perspective, particularly addressing problems with aggregate bias.

As will be discussed in more detail in Section 3.1.2, forecast accuracy measures represent non-summable quantities. Thus, special care needs to be taken when performing aggregation in hierarchical forecasting systems. Given the many different dimensions along which

(dis)aggregation can occur (see Section 2.2.2), two basic cases must be distinguished (see Kilger and Wagner, 2008):

- Aggregation by time
- Aggregation by other dimensions such as product, geography, sales channel etc.

The first case poses no special requirements and a number of approaches have already been presented in Section 2.2.4. The second case requires making sure that only matching items are compared. Two essential requirements will be discussed in the following.

First, aggregation of forecast measures should only be performed for time-series with **similar units of measurement**. For example, in markets with volatile prices, a certain forecast accuracy measured for a sales revenue forecast has a different interpretation than a forecast accuracy measured for a forecast in sales units. In the latter case, the effect of changed prices is not accounted for. As discussed earlier, most planning environments require forecasts in terms of revenues *and* units. Hierarchical forecasting systems need to provide for functionality to convert one into the other while maintaining consistency of the overall database. This requires that prices and profit margins are available at different hierarchical levels to allow for conversions between units and revenues. In practice, prices and profit margins are usually aggregated via demand-weighted arithmetic averages (see Section 3.1.2). Disaggregation of prices may be more cumbersome, but in the absence of additional information, the higher-level figure can simply be copied to the lower level (in SAP's APO system, this disaggregation procedure is termed 'average of key figures', see SAP AG (2011)). For example, if a price is changed at an aggregate level, it is helpful to propagate this change immediately to the lower levels of the planning hierarchy. If revenue data is stored separately from the unit and price information, consistency can be ensured by immediately multiplying the updated price data with the number of units (Kilger and Wagner, 2008, p. 155).

The second requirement is that forecast **accuracy measures** are **weighted appropriately**. Middle and senior management often requires forecast accuracy metrics at higher hierarchical levels. If direct forecasts at the aggregate level (and the disaggregate actual values) are available, such metrics can be calculated easily. However, if a bottom-up forecasting approach has been used, also the accuracy metrics require an aggregation. Since simple sums may not be used for the aggregation of non-summable figure,[40] forecast accuracy figures such as MAPE are aggregated to higher hierarchical levels in practice by calculating demand-weighted arithmetic means, similar to the case with prices or profit margins.

Kilger and Wagner (2008, p. 152) suggested using weights of the form

$$w_i = \frac{d_i + \hat{d}_i}{\sum_{j=1}^{m} (d_j + \hat{d}_j)}, \tag{2.14}$$

[40] This aspect will be discussed in more detail in Section 3.1.2.

allowing to calculate the aggregate (bottom-up) forecast accuracy $MAPE_{BU}$ as

$$MAPE_{BU} = \sum_{i=1}^{m} (w_i \cdot MAPE_i). \tag{2.15}$$

Here, $MAPE_i$ is the forecast accuracy obtained for the lower-level time-series i and m is the number of time-series which are to be aggregated. This type of weight is particularly helpful for the aggregation involving time-series with intermittent demand where either the actual demand or the forecast value may be zero in some periods or intervals. A standard demand-weighted approach will not consider the corresponding contribution to overall forecast accuracy as $d_i / \sum_j d_j = 0$.

Optimal Design of Forecasting Hierarchies

All previous comments on hierarchical forecasting have largely focused on aggregation and disaggregation operations within given demand planning hierarchies. Nevertheless, this leaves the more important strategic question unanswered 'in which direction' aggregation should ideally be performed. More precisely, when building demand planning and forecasting hierarchies, one often may choose between different dimensions to represent data at higher hierarchical levels. For example, assume that at the lowest level of planning, a company requires forecasts per SKU per month per customer. Ignoring the dimension time, should aggregate data at the next higher level of planning be represented at the level of regions, or rather at the level of product families?

Most firms have a legacy demand planning and forecasting hierarchy, but it is often unclear whether this is the best setup. In the following, a number of ideas will be discussed which have originally been presented by Chen and Chen (2004).

They defined a *demand planning hierarchy* as a "sequence of steps which starts from the highest aggregation level, and ends at the most detailed disaggregated level" (Chen and Chen, 2004). In the following, the inverse, yet equivalent direction will be used. For example, consider the demand perspective (Figure 2.8a) with the two dimensions customers (sales view) and products (marketing view). Both dimensions have a hierarchical structure as customer accounts are grouped on a geographical basis to regions. The highest level of planning comprises the entire customer base of the company. Similarly, products (brands) are grouped to product families, to product types (product categories) up to the corporate level. Starting at the most disaggregate level with forecasts at the SKU per account-level—should the next higher level of demand planning data be represented at the country level (geography dimension) or better at the level of the product family? Similar choices exist at each of the next higher levels.[41] Ultimately, a number of different

[41] As discussed by Chen and Chen (2004), this problem also arises if the dimensions do not have a hierarchical relationship, but if a number of functional attributes are considered. For example, they discussed various options to define a demand planning path in the semiconductor industry. Here, one has a choice among several functional attributes to define product families at different levels of: functions (memory, logic, ASIC), manufacturing technology (i.e. the width of the structures, usually measured in μm or nm) or the number of metal layers.

paths may be taken to traverse the hierarchy from the bottom to the top (or vice versa). All these paths consist of the same number of steps. Several possible paths are indicated in Figure 2.13.

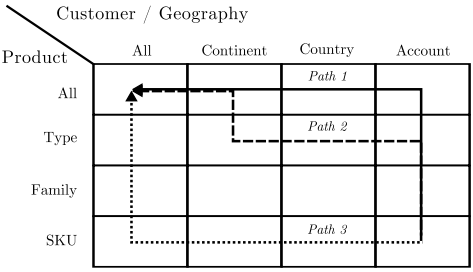


Figure 2.13. – Different paths in a demand planning hierarchy (adapted from Chen and Chen (2004))

Following Chen and Chen (2004), a good demand planning hierarchy path will minimize the demand-weighted coefficient of variation of the demand data. Alternatively, the demand planning path can also be determined with respect to the forecast accuracy. In this latter case, the demand-weighted relative forecast error will be minimized which is measured by the root of the MSE divided by the mean of the aggregate demand $\bar{D}$, i.e. the CV-RMSE measure.[42] Chen and Chen preferred to term this value the *coefficient of forecast error*.

If correlations between demand series can be ignored, the demand-weighted coefficient of variations of m time-series with mean $\bar{d}_i$ and coefficient of variation of $CV_i = \sigma_i/\bar{d}_i$ is given by

$$CV = \sum_{i=1}^{m} \left(\frac{\bar{d}_i}{\sum_{j=1}^{m} \bar{d}_j} \cdot \frac{\sigma_i}{\bar{d}_i} \right) = \frac{\sum_{i=1}^{m} \sigma_i}{\sum_{i=1}^{m} \bar{d}_i}. \tag{2.16}$$

Similarly, the demand-weighted and normalized root of the squared forecast error $\sqrt{MSE}/\bar{D}$ is calculated in a similar bottom-up manner via

$$CV - RMSE = \sum_{i=1}^{m} \left(\frac{\bar{d}_i}{\sum_{j=1}^{m} \bar{d}_j} \cdot \frac{\sqrt{MSE_i}}{\bar{d}_i} \right) = \frac{\sum_{i=1}^{m} \sqrt{MSE_i}}{\sum_{i=1}^{m} \bar{d}_i} = \frac{\sum_{i=1}^{m} \sqrt{MSE_i}}{\bar{D}}. \tag{2.17}$$

The optimization problem consists of finding a demand planning hierarchy which has smaller variations or errors at the top (Chen and Chen, 2004). The authors presented both a greedy and a dynamic programming algorithm for this problem. In the simple greedy approach, the next aggregation step is chosen by considering all possible aggregation dimensions at the next higher level. For each possible adjacent aggregation step, the weighted-average CV or CV-RMSE according to (2.16) and (2.17) will be calculated. The

[42] Note that here the additional qualification CV-RMSE rather than CV must be used for the forecast accuracy measure to avoid ambiguity.

aggregation dimension leading to the lowest CV or CV-RMSE will be chosen. For example, consider applying this greedy approach to the two dimensions depicted in Figure 2.13. In a first step, demand data or forecasts are available at the level 'SKU, per account'. Now determine CV or CV-RMSE at the next higher levels 'SKU, per country' and 'product family, per account'. Whichever dimension leads to the lowest CV or CV-RMSE is chosen as the first aggregation step. This check is repeated at each hierarchy level until the top level has been reached. Alternatively, a dynamic programming algorithm may be used to find an entire path along which the sum of CV or CV-RMSE is minimal. This problem is equivalent to finding the shortest path in a network.

Yet, this approach is often too simplistic as it neglects important interdependencies with other planning tasks. In practice, the choice of a demand planning hierarchy must be aligned with the overall planning system, rather than simply focusing on minimizing the variability of the demand or of the forecast data. For example, to reduce the complexity of the models used in master planning, a particular minimum level of aggregation must be employed. Hence, direct or derived demand forecasts at this level of aggregation must be available in an acceptable quality to allow for reliable plans. This already fixes some parts of the demand planning path. As a result, the choice of the demand planning hierarchy is strongly dependent on the overall planning system used in a particular planning environment.

This closes the discussion of demand planning in general and of hierarchical forecasting in particular. As indicated before, the output of this important planning task influences a number of other SCP problems, especially master planning. This will be covered in the next section.

2.3. Master Planning

Based on the demand forecasts provided by demand planning, master planning synchronizes the flow of materials in the entire supply chain over a mid-term time horizon (see Fleischmann and Meyr, 2003). In line with the hierarchical planning concept, master planning decisions are made within the limits imposed by the higher strategic planning level. Thus, master planning aims at using the established infrastructure as effectively as possible. Its results in turn constitute constraints and targets for short-term operational planning. In the following, first the objectives and planning tasks of master planning will be discussed in Section 2.3.1. Afterwards, a basic master planning model will be introduced in Section 2.3.2.

2.3.1. Objectives and Planning Tasks

Mid-term master planning is important because most production environments—particularly in the case of MTS—involve significant lead times before a final product becomes available for sale to a customer. Typically, three types of lead times exist:

- **Procurement lead times:** Raw materials, subassemblies and other procured items usually need to be ordered early to ensure availability at the time of production. For example, in the previously mentioned planning situation in the refining industry described by Roitsch and Meyr (2008), crude oil procurement decisions have to be made 2–8 weeks ahead of actual production based on aggregate forecasts.

- **Manufacturing lead times:** Machine capacity has to be considered as fixed, at least in the short- and medium term. If aggregate demand exhibits seasonal fluctuations, a possible solution in master planning consists of anticipatively building seasonal stocks. This may fix a large part of the master plan early in the planning process. Furthermore, an adjustment of employment levels (e.g. overtime work to maximize machine runtime) has to be announced with sufficient lead time in many jurisdictions and is typically subject to negotiations with employee representatives.

- **Deployment lead times:** If planning takes place in an MSTS environment, not only production, but also some distribution processes will be executed based on forecasts (e.g. in the consumer goods industry). Hence, long legs in the distribution system, e.g. via long-distance cargo vessels, may necessitate an early finalization of the production and distribution plan.

Master planning uses a central perspective and considers all relevant costs, constraints, temporal dependencies and bottlenecks in the SC to determine feasible aggregate plans for the mid-term horizon. Typical measures include (see Rohde and Wagner, 2008, p. 162):

- to build seasonal stocks,

- to temporarily increase capacity by working overtime,

- to produce at different sites while incurring higher or lower production costs, possibly offset by additional transportation costs,

- to outsource production to an external third party and

- to employ alternative transportation modes and delivery routes.

The objective of most master planning models is either cost minimization or profit maximization. Such models are usually of the LP or MIP type. Like all mid-term planning tasks, also master planning is based on deterministic planning data. Apart from internal production-related data, e.g. regarding costs, capacities and inventory levels, the most important set of planning data is provided by demand planning. To cope with unavoidable forecast errors, safety stocks are a typical way to handle the stochasticity of demand. Safety stocks are often included in master planning in the form of minimum inventory constraints. Their aggregate size is usually determined in demand planning (see Section 2.2).[43]

[43] In the case of demand planning in multi-stage hierarchies, determining safety stock levels at the disaggregate level is a challenging task. However, this problem is beyond the scope of this thesis.

To simplify the planning problem, to increase the likelihood that a feasible solution will be found and to cope with the unavoidable inaccuracy of planning data, master planning requires an aggregation of input data, of resources and of constraints. Aggregation in master planning usually occurs regarding time, decision variables and input data (Rohde and Wagner, 2008, pp. 172–174).

Master planning is a typical planning task in which rolling schedules are employed to allow for regular plan adjustments. While plans are made for a number of periods into the future, only the plan for the first few periods is binding and will be implemented. The plans for the subsequent periods are only of a tentative nature. They may still be adjusted if either updated or more detailed information becomes available, e.g. in the form of improved demand forecasts.

As discussed in the context of hierarchical planning, the demarcation between mid-term master planning tasks and those addressed either in operational or strategic planning may often be fuzzy. The allocation of planning tasks to the different planning levels usually depends on the industry and the actual production environment. For example, in the base chemicals industry, lot-sizing is often already determined on a mid-term planning horizon and thus part of master planning. In other industries, e.g. food and beverages, lot-sizing is part of the short-term planning and scheduling tasks (see Wagner and Meyr, 2008). This choice is closely related to the lengths of the time buckets employed in master planning. Since master planning is a typical mid-term planning task, its planning horizon spans at least one seasonal cycle, i.e. 12–18 months. This period is usually divided into monthly or weekly time buckets. Hence, if the manufacturing time per production lot is rather in the range of days or hours, lot-sizing decisions are made at a more disaggregate planning level.

2.3.2. Basic Master Planning Model

While strategic network planning is used to determine the markets to serve, no actual supply quantities can be determined at this early stage in the planning process. But better data is usually available at a mid-term planning level. Hence, in master planning, not only the quantities to be produced, but also their allocation to different warehouses and sales regions will be determined. The supply network as defined in the strategic network planning task has to be utilized as efficiently as possible by making adequate production, distribution and sales decisions.

In contrast to strategic network planning where only rough profitability estimates per market may be used, more reliable information regarding revenues and individual cost components is usually available in master planning at the geography level. Since sales regions have different revenue potentials, numerous trade-offs need to be solved involving production, inventory and transportation decisions. To prevent that strategically important, though not yet highly profitable regions are under-served, lower bound constraints may be set in the master plan to ensure a minimum service level.

To illustrate the fundamental mid-term decision problems in an MTS supply chain, a basic master planning model will be stated for a two-stage supply chain which produces a

single product. The following LP model was presented by Fleischmann and Meyr (2003, pp. 493–494). Assume that there are multiple manufacturing plants, warehouses and sales regions, indexed by $p \in \mathcal{P}$, $w \in \mathcal{W}$ and $s \in \mathcal{S}$. The model will determine optimal values for three types of decision variables within the planning horizon $t = 1, \ldots, T$:

- The production quantities $x_{pwt} \geq 0$ which are to be manufactured in plant p and sent to warehouse w in period t.
- The sales quantities $y_{wst} \geq 0$ to be sold via warehouse w in sales region s in period t.
- The inventory level $I_{wt} \geq 0$ of warehouse w in period t.

All other parameters of this basic model are summarized in Table 2.5. Note that transportation times are assumed to be sufficiently small in comparison to the length of each planning period t, allowing omitting them from this model.

Data and model parameters	
r_{st}	Unit revenues in sales region s in period t
c^p_{pw}	Per unit costs for production in plant p and transportation to warehouse w
c^h_w	Per unit holding costs in warehouse w
c^d_{ws}	Per unit transportation costs from warehouse w to sales region s
I_{w0}	Initial inventory in warehouse w
a_p	Time required to produce one unit of output in plant p
K_{pt}	Available production capacity in plant p in period t
$\hat{d}^{min}_{st}$,$\hat{d}^{max}_{st}$	Minimum sales requirements and maximum sales forecast in region s in period t

Table 2.5. – Notations of the basic master planning model

The resulting LP is a profit maximization problem:

$$\text{Max} \sum_{w\in\mathcal{W},s\in\mathcal{S},t=1}^{T} r_{st} \cdot y_{wst} - \sum_{p\in\mathcal{P},w\in\mathcal{W},t=1}^{T} c^p_{pw} \cdot x_{pwt} - \sum_{w\in\mathcal{W},t=1}^{T} c^h_w \cdot I_{wt} - \sum_{w\in\mathcal{W},s\in\mathcal{S},t=1}^{T} c^d_{ws} \cdot y_{wst} \tag{2.18}$$

subject to

$$I_{wt} = I_{w,t-1} + \sum_{p\in\mathcal{P}} x_{pwt} - \sum_{s\in\mathcal{S}} y_{wst} \qquad \forall w \in \mathcal{W}, t = 1, \ldots, T \tag{2.19}$$

$$a_p \sum_{w\in\mathcal{W}} x_{pwt} \leq K_{pt} \qquad \forall p \in \mathcal{P}, t = 1, \ldots, T \tag{2.20}$$

$$\hat{d}^{min}_{st} \leq \sum_{w\in\mathcal{W}} y_{wst} \leq \hat{d}^{max}_{st} \qquad \forall s \in \mathcal{S}, t = 1, \ldots, T \tag{2.21}$$

In the objective function (2.18), total profits, i.e. the difference between revenues and the various cost components, are maximized. The set of equations (2.19) represents the inventory balance constraints. The constraints (2.20) limit production in each plant to the available capacity. Finally, constraints (2.21) ensure that the sales quantities remain within lower and upper bounds. While the lower sales bounds often correspond to (contractual) minimum service commitments, the upper bounds typically reflect the maximum demand forecasts.

The key results of this master planning model are the product quantities x_{pwt} to be transported to each regional warehouse w per time period t. In line with the hierarchical planning concept, these regional allocations constitute constraints on the number and size of orders which may be fulfilled in each sales region. A further refinement of these quantities will take place in the subsequent short-term planning tasks. In particular, differences in terms of customer heterogeneity have not yet been observed explicitly. This is the purpose of the demand fulfillment task which will be characterized in the following section.

2.4. Demand Fulfillment

The process of handling a customer order after it has entered a company's planning system is generally referred to as *demand fulfillment* (Fleischmann and Meyr, 2004). This supply chain planning task extends the long- and mid-term sales-related tasks strategic network planning and demand planning to the short term. More specifically, demand fulfillment refers to all order-driven activities downstream of the CODP. Furthermore, the planning tasks associated with demand fulfillment are also closely linked to other short- and mid-term tasks upstream of the CODP such as master planning, distribution planning and deployment.

After discussing the objectives and key planning tasks of the demand fulfillment problem in Section 2.4.1, an overview of basic demand fulfillment system types will be given in Section 2.4.2. Afterwards, in Section 2.4.3, basic models will be presented for the demand fulfillment problem in MTS environments with a flat partitioning of the customer segments. Finally, an overview will be given of the current state-of-the-art of demand fulfillment in MTS environments (Section 2.4.4). The conclusion from this overview confirms the need for the research presented in this thesis. It will be shown that the existing approaches primarily apply to a flat partitioning of the customer segments and do not provide sufficient support yet for the case of multi-stage customer hierarchies.

2.4.1. Objectives and Planning Tasks

With many supply chains being pressed hard to ensure lean operations while at the same time satisfying high customer service levels, demand fulfillment is now a core business activity for many firms. Its decisions have a strong and immediate impact on profits in a supply-constrained supply chain. Starting as a mere availability record, demand fulfillment has developed into a powerful planning and decision support system.

Its most important functionality is to distinguish between customers or between orders of different priorities. Given limited resources (e.g. as specified by master planning), enhanced service levels should primarily be offered to the most important customers who guarantee the long-term success of a company. While the importance of demand fulfillment is undisputed, no consensus seems to exist among researchers and practitioners alike regarding the exact decision problems which are to be supported (see Framinan and Leisten, 2010). In this thesis, the argumentation by Kilger and Meyr (2008, p. 181) will be followed. They postulated three key objectives of demand fulfillment:

1. To generate reliable quotes and thus improve the on time delivery,
2. to increase the number of business opportunities by searching effectively for feasible quotes and
3. to increase the average sales price and in turn improve revenue and profitability.

For analytical purposes, it is helpful to distinguish between two key phases of demand fulfillment: *order promising* and *order fulfillment*[44] (see Fleischmann and Meyr, 2004). These two phases, as well as the associated decision problems and their outputs are summarized in Figure 2.14.

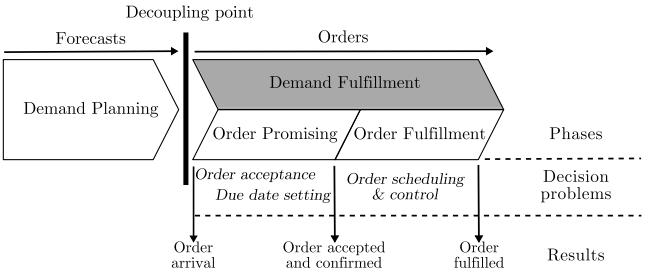


Figure 2.14. – Phases, decision problems and results of the demand fulfillment task

A key driver which determines the exact planning tasks of demand fulfillment is the position of the CODP, separating the forecast-driven and the order-driven tasks in a supply chain. The order promising phase can be loosely summarized as processing and replying to a particular customer order request. This requires first an *order acceptance* decision for all incoming order requests. In most cases, the order requests are characterized by some flexibility in terms of the order due date. Hence, the problem of *due date setting* must also be solved. The subsequent order fulfillment phase takes care of the accepted customer orders. The underlying decision problem will be referred to as *order scheduling & control*. It ensures that the order will be fulfilled as originally confirmed to the customer. Order scheduling & control is a particularly complex problem in environments with many order-driven processes like MTO and ATO.

[44] The order fulfillment phase is sometimes also termed *order execution*, e.g. in Okongwu et al. (2012).

Obviously, several other classifications of the demand fulfillment tasks and decision problems are also possible. The three main decision problems introduced above correspond to the split discussed in Framinan and Leisten (2010). A similar logic is also used by Fleischmann and Meyr (2004) who use the term *demand-supply matching* rather than order scheduling & control. They mention *shortage planning* as a further decision problem which comprises adequate actions if capacities or quantities are in short supply. On the one hand, such situations occur if a new order is to be accepted. On the other hand, shortage planning is also an inherent task in order scheduling & control if a single or a batch of (already accepted) orders needs to be re-scheduled (or in the worst case re-promised). This latter situation is more likely to occur in MTO and ATO environments due to the duration of all order-driven tasks. For the scope of this thesis, shortage planning will be split. Its two main sub-tasks will be covered as part of the order promising and as part of the order scheduling & control subproblems.

In the following, the three decision problems order acceptance, due date setting and order scheduling and control will be explained in more detail. A summary of the individual tasks per decision problem is provided in Table 2.6.

Phase	**Decision problem**	**Main tasks**
Order promising	Order acceptance	Order receipt ATP availability check Search rules for supply alternatives Shortage planning
	Due date setting	Due date assignment Order confirmation
Order fulfillment	Order scheduling and control	Order execution control & demand supply matching Order rescheduling Shortage planning Deployment

Table 2.6. – Decision problems and main tasks of demand fulfillment (based on Fleischmann and Meyr, 2004; Framinan and Leisten, 2010)

Order Acceptance: All demand fulfillment steps are invoked once a customer submits an order request. Most firms provide many ways for the order submission, e.g. online, via phone or with a sales agent. The **order receipt** task ensures that the order request is recorded properly, entered into the company's order management system and that any obvious errors or missing data entries are corrected (for more such operational aspects, see Croxton, 2003).

Subsequently, a check is necessary whether the company has sufficient resources available to fulfill the order request. For now, focus on the case of an MTS environment in

which final items are produced according to forecasts. Here, it typically suffices to conduct an **ATP availability check** which consists of merely verifying the available inventories to determine if the order can be fulfilled. If stocks are insufficient, a feasible due date can be determined by simply quoting the default lead time required for manufacturing and distribution. More reliable commitments may be obtained and more promises may be kept by explicitly observing the constraints within the supply chain. Essentially, orders should rather be quoted against the amount of yet uncommitted stock in the inventory and against the planned production quantities which become available according to the MPS (McClelland, 1988). These quantities are commonly referred to as available-to-promise or ATP quantities (Schwendinger, 1979).[45]

For a planning horizon of T periods, the ATP quantities for a given single product may be calculated as follows (see Fleischmann and Meyr, 2003, p. 507): Use I_t to denote the inventory in period $t = 0, \ldots, T$, where I_0 is the amount of initial inventory, and let S_t and C_t designate prospective supply arrivals and the amount of already accepted customer orders in period t, respectively. First, the inventory position in each period t can be calculated via a simple forward-pass calculation,

$$I_t = I_{t-1} + S_t - X_t \quad \forall t = 1, \ldots, T \tag{2.22}$$

Provided no negative value of I_t exists in $t = 1, \ldots, T$, the ATP quantities can be derived by working backwards from period T. Introduce $I^\star_T = I_T$ and $I^\star_t = \min\{I_t; I^\star_{t+1}\}$ for $t = 0, \ldots, T-1$. This gives

$$ATP_t = I^\star_t - I^\star_{t-1} \tag{2.23}$$

for the amount of ATP quantities which become available for order promising in period t or later. As will be seen later, in some situations, it is more important to consider the amount of ATP quantities which are available in a particular period, including those which remain from previous periods. This cumulative ATP quantity can be defined as

$$CATP_t = \sum_{s=1}^{t} ATP_s = I^\star_t. \tag{2.24}$$

The second part of the above equality holds if $I_0 = 0$.

Note that order acceptance based on the ATP quantities as defined in (2.23) or (2.24) is a mere bookkeeping function without any planning functionality (Chen et al., 2001, p. 477). Corresponding simple order promising systems check the availability of the existing or planned resources to determine if an order request can be accepted. More sophisticated systems (sometimes also referred to as *advanced ATP* systems) provide real planning functionality. They may be configured with a set of multi-dimensional **search**

[45] In parts of the literature, the term ATP is sometimes used very broadly, e.g. to also designate the entire demand fulfillment planning task (e.g. in Chen et al. (2001)) or to refer to the supporting software modules (e.g. in Framinan and Leisten (2010)). In this thesis, the term ATP is used more restrictively to refer to the available quantities of final items.

rules to identify supply alternatives if shortages occur. This is particularly relevant in MTS and MSTS environments. Three search dimensions are typically used in this **shortage planning** step:

- Time (delivery earlier or later than requested by the customer),
- product (search for a substitute product, e.g. from same product family) or
- geography (delivery from a more distant supply location).

Obviously, combinations of these dimensions are possible. To ensure a positive contribution to overall company profits, an important planning functionality in this search process is to differentiate between orders of different importance and to explicitly observe heterogeneity in the customer base, e.g. due to different levels of customer profitability. This functionality is also referred to as *profitable-to-promise*, e.g. by SAP AG (2003). Basic models for the order acceptance decision with heterogeneous customers will be introduced in Section 2.4.3.

The order acceptance decision problem is different if the CODP lies upstream of or between production tasks. To be able to accept an order in an MTO environment, a company must have sufficient free capacity at all bottleneck resources which are required in the production process. This *capable-to-promise* (CTP) availability check replaces the ATP availability check in MTO environments where the availability of raw materials only rarely constitutes a limiting factor. The situation is more complex in case of an ATO environment. Here, the order acceptance decision needs to be based on a joint ATP/CTP availability check as both component inventory and assembly capacity may constitute bottlenecks.[46] A comprehensive discussion of the different nature of the order acceptance planning task in the MTO, ATO and MTS environment has been provided in Fleischmann and Meyr (2004, Ch. 3).

Due Date Setting: The tasks within the *due date setting* decision problem again depend on the degrees of freedom which may be exploited by the supply chain: In many MTS and MSTS environments, the customer either has specified a mandatory due date or requires immediate delivery, typically within a range of 24–72 hours (Fleischmann and Meyr, 2003, p. 505). If feasible, the order is confirmed to the customer as requested or it will be rejected. If the customer order allows for some due date flexibility, due date setting corresponds to simply returning the result of the ATP availability check look-up function to the customer as the **order confirmation**.

Due date assignment is therefore particularly relevant in the context of MTO and ATO and closely related to the order scheduling problem. Often, order completion times

[46] As in the MTS case, companies operating in an MTO or ATO environment should focus on identifying particularly profitable order requests. For ATO cases, Ervolina et al. (2007) proposed an 'availability to sell' functionality for customizable products such as computers. It allows for component substitutions and suggests up-selling opportunities (i.e. replacing components by more profitable alternatives) to improve overall profits from the order.

are stochastic, rendering the due date assignment problem particularly difficult. Keskinocak and Tayur (2004), Framinan and Leisten (2010) and Slotnick (2011) provided overviews of the current state of the art. Note that the due date which is ultimately committed to the customer is often set to a later period than the estimated completion time of the production orders associated with the order request. Leaving some slack in the initial due date quotation increases the degrees of freedom for the subsequent *order scheduling and control* tasks particularly in MTO and ATO environments.

Order Scheduling & Control: Starting and completion times of the production orders associated with the accepted customer orders need to be controlled to efficiently utilize the available resources and raw materials (**order execution control & demand supply matching**). This is important if the already committed orders have to be rescheduled due to unanticipated shortages, e.g. concerning raw materials or subassemblies (**shortage planning**). Furthermore, later-arriving, more important orders may be granted preferred access to production capacities and materials, necessitating an adjustment of the production schedule (**order rescheduling**). These tasks are particularly challenging in MTO and ATO environments.

In MTS/MSTS environments, delivery orders are usually released immediately upon order acceptance. For single-item orders, order scheduling & control simply consists of reserving and dispatching the supply quantities which have been determined as part of the ATP availability check. In case of multi-line orders, this process is equivalent to **deployment**. Different types of stock, often from different inventory locations need to be allocated to a particular order (Fleischmann and Meyr, 2004).

The above overview of planning tasks and decision problems illustrates that demand fulfillment systems need to be carefully tailored to the requirements of specific industries, business contexts and production environments. In the following, a basic classification of demand fulfillment systems will be introduced.

2.4.2. Types of Demand Fulfillment Systems

In the following, a classification of demand fulfillment systems due to Pibernik (2005) will be summarized to illustrate the breadth of the existing demand fulfillment approaches. This classification allows characterizing demand fulfillment systems along the three dimensions *interaction with manufacturing resource planning*, *availability level* and the *operating mode*. As each dimension describes two major contrasting characteristics, in total eight different basic demand fulfillment system types can be distinguished. These types are indicated in Figure 2.15.

While the resulting basic types are not truly mutually exclusive, this classification permits illustrating the key trade-offs involved when designing a demand fulfillment system. In the following paragraphs, the three dimensions of this classification will be explained in more detail to provide an overall picture of possible demand fulfillment systems. Later, in Section 2.4.3, the focus will be placed on a particular subset of demand fulfillment

systems, passive systems for finished goods. This selection has been highlighted in gray in Figure 2.15.

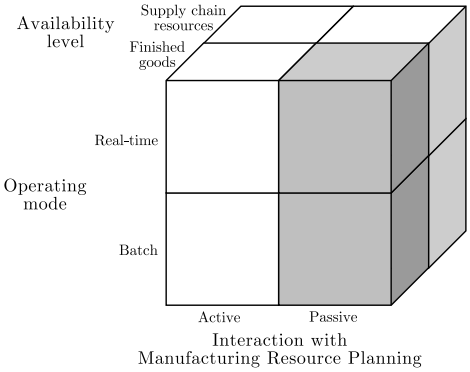


Figure 2.15. – Generic demand fulfillment system types (Pibernik, 2005, p. 242)

Availability Level: A demand fulfillment system may promise orders either based on *finished goods* or on the basis of *supply chain resources* such as raw materials, subassemblies and production capacities. This differentiation follows the position of the CODP in the production environment. While in MTS environments production is independent of actual orders and inventories are thus held at the level of final products, this is different in MTO and ATO environments. Here, demand fulfillment always needs to consider supply chain resources since production and assembly, respectively, are not triggered before the customer order has been received and confirmed.

Interaction with Manufacturing Resource Planning: On the one hand, a demand fulfillment system may be *passive* in the sense that it does not change the production schedule. The system is merely capable of checking the availability at the level of final goods or at the level of supply chain resources. If suitable products or resource capacities have been found, they will be reserved by the demand fulfillment system once an order has been accepted. On the other hand, an *active* demand fulfillment system has the capabilities to directly adjust the production schedule when scheduling individual orders and determining due dates.

Passive systems rely on powerful search rules for the order acceptance module. These rules help identifying fulfillment alternatives to ensure consistently high service levels to the most important customers. It is of utmost importance to prevent promising the last available stock units or capacities to a less important customer if chances are high that a more important customer will request this last item on stock immediately after-

wards.[47] This is particularly relevant in a passive demand fulfillment system which is used to promise finished goods. Here, the order promising decision is usually irrevocable as accepted orders are immediately deployed and shipped.

An active demand fulfillment system is inherently more complex with respect to the due date setting and the order scheduling & control decision problems. Active demand fulfillment is more appropriate for ATO and MTO environments where assembly and production operations, respectively, are only triggered upon order acceptance. However, active demand fulfillment functionality may also exist in MTS environments. These systems trigger the inclusion of an additional, customer-specific production order via the CTP functionality. In comparison to passive systems, active demand fulfillment systems usually have more flexibility to respect different customer priorities. For example, the order scheduling & control decision problem may determine appropriate counter measures if an unexpected high-priority order arrives, e.g. by rescheduling or even re-promising the existing orders.

Operating Mode: From the customer's perspective, the two typical operating modes *batch* and *real-time* differ with respect to the timing of the responses: In real-time mode, customers receive an immediate reply in direct response to their individual order. In batch mode, on the contrary, replies are only given to the customers at regular intervals. A number of orders are collected during a batching interval.[48] The set of incoming orders is then processed jointly. Framinan and Leisten (2010) further distinguished between two types of real-time mode and an entirely off-line process:

- *Real-time process*: This is the typical situation usually implied by real-time order promising. It frequently occurs in electronic markets where the customer requires an instant response to his order request and furthermore, immediately afterwards must confirm or deny such a quote. This quick customer acceptance is particularly important in some ATO/MTO environments. Each order quotation requires an upfront scheduling step to check order feasibility, to determine the availability of the supply chain resources and/or to determine the costs of the order request. As other orders may arrive in the meantime which may compete for the same supply chain resources, it is costly to freeze a tentative schedule with unconfirmed orders for an extended period of time.

- *Real-time quotations* differ from real-time processes in that the customer does not immediately have to confirm an issued order promise. This is more acceptable in MTS situations with order promising at the level of final goods since no scheduling activities are affected.

[47] Unfortunately, many commercially available demand fulfillment systems do not provide for the possibility to observe customer heterogeneity. For an overview of academic models with this capability, see Section 2.4.4.

[48] The length of the batching interval may typically range from an hour up to several days (Ball et al., 2004). In the experimental studies of Chen et al. (2001), 1–7 days were tested whereas Lin et al. (2010) considered the range 12–72 hours.

- *Off-line quotations* cover the typical batch-promising situation: Neither party expects an immediate reply or confirmation. This gives the manufacturer more flexibility to optimally schedule the processing of the order. Furthermore, real-time access to the required production planning and scheduling systems is often not possible in the same moment when an order arrives (see Chen et al., 2002, p. 426). Nevertheless, to still reduce waiting time for a customer, sometimes a two-step, *hybrid* approach is taken: First, an initial 'soft' and often coarse promise is given to the customer in real-time (e.g. the delivery week). This is followed by a more detailed confirmation (e.g. delivery day) after some detailed scheduling has been run, usually as part of a batch process (see Ball et al., 2004, Sec. 2.3).

In ATO environments, the order acceptance decision always requires a feasibility check. If nevertheless a real-time reply is required, a passive demand fulfillment system is usually more adequate as the computational requirements have to be very modest to allow issuing a real-time order confirmation. Hence, simple availability checks which do not cause any schedule changes are more appropriate (Akkan, 1997, p. 172).

As stressed before, the focus of this thesis lies on MTS environments with order promising at the level of final goods and without scheduling (passive systems). Therefore, the remainder of this demand fulfillment discussion will address the two suitable system types for this setting which have been indicated in Figure 2.15. The next section will present basic mathematical problems for passive batch and the real-time order promising for MTS environments.

2.4.3. Basic Models for Demand Fulfillment

The major advantage of a batch order promising system is the ability to make a selection from a group of order requests and to only fulfill the most important or most profitable orders if the available resources are constrained. This involves a key trade-off: A longer batching interval is more beneficial from the firm's perspective, but results in a severe degradation of customer service as replies will be given with a substantial delay. Furthermore, a long batching interval may lead to missed business opportunities if orders arrive which have due dates before the end of the current batching interval (see Chen et al., 2001).

In choosing between a batch and a real-time system, the balance of power between the manufacturer and the customers is crucial: In a buyer's market, customer expectations, e.g. in terms of response or delivery lead time, are usually not negotiable. This is typical of many standard, high-volume products produced and sold in MTS / MSTS environments. The situation is different if a manufacturer has a strong competitive position or manufactures products which require customization to customer specifications. In these latter cases, the manufacturer will usually have more lead time in responding to and more flexibility in promising a particular order.

In the following, two typical approaches for demand fulfillment with a flat partitioning of the customer segments will be characterized in more detail. First a basic batch order

promising model will be presented and discussed, highlighting its weaknesses. Afterwards, a real-time approach will be introduced which is based on an allocation planning step. Individual orders will be handled by a separate consumption planning model upon arrival. It is the objective of the subsequent discussion in Chapters 4 and 5 to extend the allocation planning-based approach with subsequent consumption planning to the case of multi-stage customer hierarchies.

Basic Batch Order Promising Model

In the following, a basic batch order promising model as suggested by Fleischmann and Meyr (2004, pp. 310–311) will be introduced. Table 2.7 summarizes the notation. The model consists of the following LP:

Indices	
$i \in \mathcal{I}$	Individual order from the set of all open customer orders $\mathcal{I}$
$t = 1, \ldots, T$	Time periods
Data	
q_i	Desired quantity of order i
d_i	Desired delivery date of order i $(1 \leq d_i \leq T)$
ATP_t	ATP quantity becoming available in period t $(t = 1, \ldots, T)$
p_{it}	Profits = revenues − costs, from serving one unit of order i with ATP_t; revenues and costs may differ per order i; costs include: • Costs for early allocation $(t = 1, \ldots, d_i - 1)$ • Costs for backlogging $(t = d_i + 1, \ldots, T)$
$p_{i,T+1} < 0$	Negative profits (=penalty) for not fulfilling order i within the planning horizon
Decision variables	
$o_{it} \geq 0$	Part of order i which is served with ATP_t from period t
$o_{i,T+1} \geq 0$	Part of order i which is not fulfilled within the planning horizon T

Table 2.7. – Notation of the batch order promising model

$$\text{Max} \sum_{i,t=1}^{T+1} p_{it} \cdot o_{it} \tag{2.25}$$

subject to

$$\sum_{i \in \mathcal{I}} o_{it} \leq ATP_t \quad \forall t = 1, \ldots, T \tag{2.26}$$

$$\sum_{t=1}^{T+1} o_{it} = q_i, \quad \forall i \in \mathcal{I} \tag{2.27}$$

The objective (2.25) of this network-flow type model is to find profit-maximizing order acceptance and fulfillment decisions for all orders in the set $\mathcal{I}$ which have been received during the previous batching interval. Each fulfillment alternative of each order differs

in terms of the associated profits. For simplicity, it will be assumed that unit revenues are given and fixed over the planning horizon. Thus, the focus of the model lies on the order-specific *tangible* and *intangible* costs in the profit terms p_{it}. The former refer to the direct costs incurred by serving one unit of order i with ATP_t while the latter penalize non-adherence to order specifications. They capture the actual (e.g. contractual penalties) as well as the virtual (e.g. customer annoyance, reduced future purchase probability and generally reduced customer retention) short- and long-run costs of suboptimal customer service. The resulting per unit profits p_{it} allow discriminating between orders of varying importance and between different fulfillment alternatives (see Pibernik (2006) and Jung (2010) for similar formulations). Nevertheless, the quantification of the associated cost components is a crucial task, particularly regarding the long-run effects (e.g. see Anderson et al. (2006) for an example concerning stockout costs).

Constraints (2.26) limit the consumption of ATP in each period to the quantities which are 'available to promise'. A major characteristic of demand fulfillment systems for MTS environments is that production and order promising decisions are decoupled, i.e. the timing and quantities of ATP replenishments are given, e.g. by prior master planning or scheduling (see Section 2.3). Such fixed replenishment schedules are common in industry due to efficiency gains and cost reduction potentials in production and distribution (Graves, 1996; Ernst and Kamrad, 1997). A different approach is adopted in the literature on inventory rationing where the replenishment decision is determined endogenously.[49]

The second set of constraints (2.27) ensures that all units of each order i are ultimately taken care of. Usually, orders are served within the planning horizon $1 \dots T$ from the available ATP quantities. In case overall ATP inventory is insufficient to meet total demand, parts or entire orders may ultimately be fulfilled from an infinite supply in dummy period $T + 1$, i.e. after the current planning horizon. This corresponds to a typical approach in practice as many companies employ a policy of never denying an order. Rather, in shortage situations, delivery due dates will be confirmed only for the very distant future after the planning horizon T (see e.g. Ball et al., 2004).

The model is run regularly, i.e. every b periods. Here, b corresponds to the length of the *batching interval* in terms of time periods. Each model run considers the orders $i \in \mathcal{I}$ which have been received during the last batching interval. All these orders are processed simultaneously and order acceptance decisions are generated for a planning horizon of T periods. Generally, it holds that $b \ll T$ to ensure that the due date d_i of all orders $i \in \mathcal{I}$ lies within the current planning horizon. After each run of the model, i.e. every b periods, the available ATP quantities have to be updated to reflect the previously accepted order quantities. These relationships between the batching interval and the planning horizon are also illustrated in Figure 2.16.

[49] The general relationship between inventory rationing and demand fulfillment was discussed in Quante et al. (2009b). Nguyen et al. (2012) extended this discussion to networks with multiple stock points. General overviews of the comprehensive literature on inventory rationing approaches have appeared in Kleijn and Dekker (1999) and in Teunter and Klein Haneveld (2008).

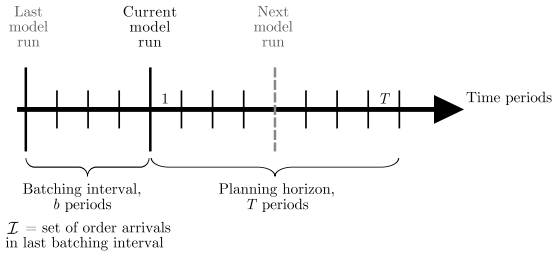


Figure 2.16. – Time structure in the batch order promising model

The main drawback of the above batch order promising model is the time delay $\leq b$ before a reply can be given to the customer. In practice, it is mandatory in many situations to give a real-time response. One may approximate real-time order promising behavior with the above model by shortening the batching interval b. This obviously reduces the number of elements in $\mathcal{I}$. In the limit, $|\mathcal{I}| = 1$, and an immediate, myopic reply can be given to the (single) customer order. The resulting order acceptance decision corresponds to a first-come-first-served scheme. This is often unsatisfactory in practice since sales agents are usually well aware of different degrees of importance within their customer base. One way to game such a myopic system is to book 'pseudo orders' based on forecasts of upcoming orders from the more important customers (Zhao et al., 2005, p. 70). This way, sales agents are able to reserve critical resources for these anticipated high priority demands without running the risk of disappointing them. A better approach will be explained in the next section.

Basic Allocation Planning Model for Real-Time Order Promising

In the following, a more thorough implementation of the above reservation idea will be presented. For important customers, product quantities will be set aside to enable a true real-time process for a given flat partitioning of the customer segments. Note that batch order promising and the related greedy real-time method may also be referred to as *pull-based* models.[50] In pull-based models, resource allocation decisions are made dynamically and in response to one or several actual order requests.

An alternative are *push-based* models. In anticipation of potentially arriving order requests, push-based demand fulfillment models pre-allocate (i.e. reserve) resources (final good inventory quantities in MTS; material, production and distribution capacities in ATO/MTO environments) for individual customers or customer segments. This reservation step is referred to as *allocation planning*. The actual order promising is made at a later point in time, upon arrival of each individual order. This *consumption planning* step is greatly simplified compared to order promising in the batch model. It merely has to check whether the remaining quotas[51] which may be consumed by a particular

[50] The differentiation of push- and pull-based models is due to Ball et al. (2004).
[51] The terms ATP 'reservations' and 'quotas' will be used interchangeably.

request are sufficient to fulfill the order. This spares the need to execute a computationally expensive optimization model and allows giving fast responses. More sophisticated consumption planning solutions consist of a set of interrelated search rules, e.g. to check quota availabilities in different time periods, at other locations or for substitute products.

This two-step process incorporates longer-term profitability considerations already into the allocation planning step. The customer segment-specific quota reservations may then be translated into reliable consumption planning decisions in a real-time process (see Ball et al., 2004, pp. 459–461).

Accurate forecasting is a necessary prerequisite for the allocation planning step. Compared to MTS environments, this is particularly difficult for ATO/MTO settings. Here, the number of necessary forecasts is significantly higher than in MTS. Forecasts are also required at the level of materials or subassemblies. Furthermore, the availability of production capacities needs to be considered. Since customers usually accept a longer lead time for products which are tailored to their specifications, real-time order promising is often not mandatory in ATO or MTO production environments. Hence, allocation and consumption planning predominantly apply to MTS environments.

Note that allocation planning does not consider individual order lines or orders, but rather operates at the more aggregate level customer segments, i.e. the grouping of similar customers. The basic relationship between order lines, orders, customers and customer segments is depicted in Figure 2.17. In allocation planning, quotas are reserved at the level of customer segments, and in consumption planning, the available fulfillment options are determined based on the customer segment associated with the customer who just placed a new order. As stated before, it is assumed in this thesis that orders consist of just one order line, i.e. a single product (as reflected by the gray right part of Figure 2.17).

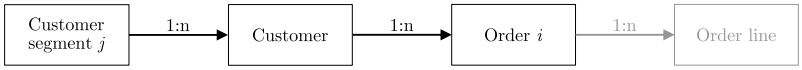


Figure 2.17. – Grouping orders to customers to customer segments

As an example for a basic real-time order promising model based on allocation and consumption planning, the model suggestion by Meyr (2009) will be described in the following. Any additional or adjusted notation compared to the previous batch model is summarized in Table 2.8.

The most important feature of the model is that the firm is assumed to be able to differentiate between different types of customers. On the one hand, there are important customers where order denial is comparably costly, e.g. regular customers who always had high and steady turnovers, or customers with whom service level agreements exist which may only be broken at high penalty costs. On the other hand, other customers are less important, e.g. casual spot-market customers. Further background on the reasons for this customer heterogeneity has already been presented in Section 1.2. Since reservations for individual customers are not viable, customers with similar importance are grouped into customer segments.

Sets and indices	
$j \in \mathcal{J} = \{1, \ldots, n_J\}$	Customer priority class
$t, \tau = 1, \ldots, T$	Time periods
Data	
ATP_t	Yet unassigned ATP quantity becoming available in period t
$\hat{d}_{j\tau}^{min}, \hat{d}_{j\tau}^{max}$	Forecast of maximum (minimum) demand occurring in customer segment j during period τ
$\bar{p}_{jt\tau}$	Profits from serving one unit of demand occurring in customer segment j during period τ with ATP_t becoming available in period t
Decision variables	
$z_{jt\tau} \geq 0$	Demand in customer j occurring during period τ, served using ATP_t which becomes available in period t
$f_t \geq 0$	Unallocated part of ATP supply in period t

Table 2.8. – Indices, data and variables of the allocation planning model (see Meyr, 2009)

The basic allocation planning problem has the form of a simple LP (Meyr, 2009, pp. 239–240):

$$\text{Max} \sum_{j,t=1}^{T+1} \sum_{\tau=1}^{T} \bar{p}_{jt\tau} \cdot z_{jt\tau} \tag{2.28}$$

subject to

$$\sum_{j,\tau=1}^{T} z_{jt\tau} + f_t = ATP_t \quad \forall t = 1, \ldots, T \tag{2.29}$$

$$\hat{d}_{jw\tau}^{min} \leq \sum_{t=1}^{T+1} z_{jt\tau} \leq \hat{d}_{j\tau}^{max} \quad \forall j, \tau = 1, \ldots, T \tag{2.30}$$

The objective (2.28) of the allocation planning problem is to allocate the available ATP supplies to the priority classes $j \in \mathcal{J}$ in order to maximize overall profits (decision variables $z_{jt\tau}$). Furthermore, a second set of decision variables f_t captures the amount of ATP quantities from period t which shall remain unallocated. Note that the profits from serving one unit of demand in period τ using ATP quantities of period t, $\bar{p}_{jt\tau}$, is an average value for all customers in segment j (and implicitly, also for all their individual orders).

The first set of constraints (2.29) specifies the utilization of the ATP quantities which become available in each period—either they are allocated to individual customer segments and time periods ($z_{jt\tau}$) or they remain unallocated (f_t). The unallocated quantities f_t may be used during actual order promising to fulfill order requests of any customer priority segment. The second set of constraints (2.30) ensures that the resulting allocations observe the upper and lower bounds on the demand per customer segment. Recall that there is unlimited supply in period $T + 1$ which ensures that a feasible solution always exists.

The optimal solution $z^{\star}_{jt\tau}$ to the problem in (2.28)–(2.30) corresponds to a fine-grained reservation of allocated ATP (aATP)[52] quantities, simultaneously specifying the target priority class and the consumption period for each unit of ATP supply. Following Meyr (2009), the allocated ATP quantity immediately after the allocation planning procedure, before the arrival of the first order, can be described as

$$aATP^{1}_{jt\tau} := z^{\star}_{jt\tau} \quad \forall j, t, \tau \tag{2.31}$$

Forecasts on such a detailed level of customer priority segment and consumption period are often not very reliable.[53] Summing over one or several dimensions of the decision problem will yield a coarser allocation result. For example, the aggregate amount of ATP quantities becoming available in a particular period t to be reserved for consumption by a particular customer segment j can be described by

$$aATP^{1}_{jt} := \sum_{\tau=1}^{T} z^{\star}_{jt\tau} \quad \forall j, t. \tag{2.32}$$

When applying the above allocation planning model to a setting in practice, a major difficulty is specifying the shortage costs per unit time and customer segment which enter $\bar{p}_{jt\tau}$. In particular, the indirect costs of sub-optimal customer service are often difficult to quantify. This challenge has caused some researchers to prefer a service-level based approach to determine quota reservations (e.g. see Pibernik and Yadav (2009)). However, this does not solve the underlying problem as it is still necessary to define an appropriate stock-out probability per customer segment. These problems also occur in inventory management. Here, Silver et al. (1998, Ch. 7) illustrated that a cost minimization and a service level objective approach are equivalent representations of the same problem. In particular, they showed that the different types of stockout probabilities (e.g. fill rate, cycle service level, ready rate) can be converted into equivalent shortage cost representations, e.g. fixed costs per stock-out occasion, fractional charges per unit short, or shortage costs per unit time. In a similar manner, Nahmias (2009, p. 275) used the term 'imputed' shortage cost to refer to the implicit shortage costs which correspond to a particular service level objective.

An alternative way to circumvent the problem of explicitly specifying shortage costs per unit time is to replace the LP-based allocation planning model by a set of simple rules. For example, it may often be easier to only specify a priority order in which the quotas for the individual customer segments should be set. Procedures for such rule-based allocation planning are provided by many commercially available APS; for a basic overview, see Meyr et al. (2008a). Rule-based allocation planning is particularly attractive for situations with

[52] aATP is also used for advanced ATP in parts of the literature. For this latter usage, see Lee et al. (2006) and Okongwu et al. (2012).

[53] Recently, approaches based on ideas from revenue management have been proposed to handle the stochasticity of the demands (Quante, 2009; Quante et al., 2009a). More details will be presented in Section 2.4.4.

decentral decisions as in multi-stage customer hierarchies with limited data transparency. Examples for such rules will be given in Section 4.3.

Basic Consumption Planning Model

Once allocations have been determined for all relevant customer segments, the order acceptance problem is greatly simplified. Individual orders may now be promised with respect to these quota reservations in a real-time process. This single-order promising after allocation planning (SOPA; see Meyr (2009)) corresponds to a simple search for the least expensive fulfillment alternative. In the following, the basic LP-based consumption planning model of Meyr (2009) for a flat partitioning of the customer segments will be presented.[54]

The model is invoked every time an order arrives. Assume that iteration s comes next. An order $\hat{\imath}(s)$ of size $q_{\hat{\imath}(s)}$ arrives with delivery due date $d_{\hat{\imath}(s)}$. The order originates from customer priority segment $j_{\hat{\imath}(s)}$. It can be fulfilled by using both allocated (o^s_{jt}) and unallocated (f^s_t) ATP quantities, up to the maximum available quantities $aATP^s_{jt\tau}$ and $uATP^s_t$.

In practice, not all possible fulfillment options may be desired. The simplest option is a dedicated consumption, i.e. orders may only be fulfilled by consuming the corresponding quota reservations. Enhanced options can be described with the help of sets which specify the permitted search alternatives. For example, the set $\mathcal{J}_i$ designates the customer priority classes to which order $\hat{\imath}$ of priority class j_i has access. By setting $\mathcal{J}_i := \{j' : j_i \geq j' \geq n_J\}$, it can be ensured that order $\hat{\imath}$ can only consume ATP which has been reserved for its own priority class j_i or ATP quantities for lower priority classes. Furthermore, note that not all available ATP quantities may have been allocated in the prior allocation planning step for consumption by a particular customer segment j in a particular period τ. Those unallocated quantities $uATP^s_t$ can also be used to fulfill a particular order $\hat{\imath}(s)$ in iteration s; this is controlled by the decision variable f^s_t. A summary of the additional or adjusted notation for the consumption planning model is given in Table 2.9.

The consumption planning model for SOPA in iteration s for order $\hat{\imath}(s)$ is given as follows (Meyr, 2009, p. 242):

$$\text{Max} \sum_{j\in\mathcal{J}_i, t=1}^{T+1} p_{\hat{\imath}(s),t} \cdot o^s_{jt} + \sum_{t=1}^{T} p_{\hat{\imath}(s),t} \cdot f^s_t \tag{2.33}$$

subject to

$$0 \leq o^s_{jt} \leq aATP^s_{jtd_{\hat{\imath}(s)}} \qquad \forall j \in \mathcal{J}_i, t = 1, \ldots, T \tag{2.34}$$

$$0 \leq f^s_t \leq uATP^s_t \qquad \forall t = 1, \ldots, T \tag{2.35}$$

$$\sum_{j\in\mathcal{J}_i, t=1}^{T+1} o^s_{jt} + \sum_{t=1}^{T} f^s_t = q_{\hat{\imath}(s)} \tag{2.36}$$

[54] Note that for an actual implementation, a simple one-pass search algorithm will be sufficient.

Sets and indices	
$\mathcal{J}_{\hat{\imath}}$	Permitted customer classes for order $\hat{\imath}$
Data	
$aATP^s_{jt\tau}$	ATP quantity which becomes available in period t which is reserved for consumption by customer segment j in period τ
$uATP^s_t$	Unallocated ATP quantity becoming available in period t
$p_{\hat{\imath}t}$	Profits = revenues − costs, from serving one unit of order $\hat{\imath}$ with ATP in period t
Decision variables	
$o^s_{jt} \geq 0$	Part of aATP reservation for customer segment j becoming available in period t which has been used in iteration s to fulfill order $\hat{\imath}(s)$ with requested due date $d_{\hat{\imath}(s)}$
$f^s_t \geq 0$	Part of uATP supply from period t which has been used in iteration s to fulfill order $\hat{\imath}(s)$ with requested due date $d_{\hat{\imath}(s)}$

Table 2.9. – Indices, data and variables of the consumption planning model (see Meyr, 2009)

The objective function (2.33) aims at minimizing the total costs of fulfilling order $\hat{\imath}(s)$. Constraints (2.34)–(2.35) restrict the fulfillment quantities to the reserved $aATP$ and to the available $uATP$ quotas, respectively. Constraints (2.36) ensure that the order quantity is fulfilled either from the accessible $aATP$ reservations, from any of the unallocated $uATP$ quantities or from the dummy supply in period $T+1$.

After each execution of this consumption planning model, the values for the $aATP$ and $uATP$ quantities need to be updated for the next iteration $s+1$. Denoting the optimal solutions of iteration s by $o^{s\star}_{jt}$ and $f^{s\star}_t$, the update corresponds to

$$\begin{aligned} aATP^{s+1}_{jtd_{\hat{\imath}(s)}} &:= aATP^s_{jtd_{\hat{\imath}(s)}} - o^{s\star}_{jt} && \forall j \in \mathcal{J}_{\hat{\imath}}, t = 1, \ldots, T \\ uATP^{s+1}_t &:= uATP^s_t - f^{s\star}_t && \forall t = 1, \ldots, T \end{aligned}$$

The above consumption planning model allows for a number of extensions to fine-tune the search for fulfillment quantities.

As a first step, one may easily introduce search priorities. The current consumption planning model effectively fulfills the first unit of the order request with the most cost-efficient fulfillment option in terms customer priority penalty costs. Once that source of supply has been exhausted, the next best alternative is chosen freely within the search space defined by $\mathcal{J}$. Meyr (2009) has shown an efficient way to establish a fixed search sequence within the search set. Additional types of search rules may easily be defined and implemented. For example, Jeong et al. (2002, p. 195) allowed for order promising in a multi-site environment. They presented simple heuristics to determine the sequence of warehouses to check for available reservations.

An alternative strategy is to define the $aATP$ quota reservations at a more aggregate level. For example, it is possible to move from a very fine-grained reservation as used above to a quota where the consumption period τ is not pre-defined, as in (2.32). However, Meyr (2009, p. 248) and Quante (2009, p. 89) reported rather disappointing results for this aggregate reservation and consumption policy.

While the concepts discussed above have been presented for a situation with a flat partitioning of customer segments and a central planner, they allow for an adaptation to decentral planning and the case of multi-stage customer hierarchies. Such an extension of allocation and consumption planning will be provided in Chapters 4 and 5 when presenting solution approaches to the DMC problem.

The batch and the allocation planning-based demand fulfillment approaches which have been introduced in this section only represent a fraction of the available demand fulfillment systems. To conclude this discussion of demand fulfillment, the next section will give an overview of alternative approaches which have been discussed in the literature.

2.4.4. State of the Art: Demand Fulfillment in MTS Environments

Many ideas for demand fulfillment have been driven by practitioners rather than by the scientific community. Vendors of ERP and of APS systems have implemented comprehensive algorithms to determine order due dates and to calculate quantities which can be promised to customers (Fleischmann and Geier, 2012, p. 162). For example, SAP's 'Global ATP' module extends order promising to an enterprise-wide basis and a global scale. It checks product requirements not only against availability, but also against allocations. This is possible due to the close interrelation of the demand fulfillment system with other key APS modules such as master and demand planning. A detailed presentation on the demand fulfillment capabilities of SAP's Advanced Planner and Optimizer (APO) was given in Bartsch and Bickenbach (2002, p. 280ff) and in Dickersbach (2009, Ch. 7).[55]

Besides the steady development in practice, demand fulfillment has also increasingly attracted the attention of academic researchers over the last years. Initially, many scientific papers on demand fulfillment only summarized needs and potential features and highlighted the importance of adopting demand fulfillment systems. As pointed out by Chen et al. (2001, 2002) ten years ago, only few papers actually addressed the underlying decision models.

This has changed in the meantime. In the following, an overview of the major contributions will be presented which describe demand fulfillment models for MTS environments with order promising at the level of final goods. The focus lies on models which account for customer or order heterogeneity. Most models solve the *order acceptance* problem, either for a batch of orders or for an individual order in case of a real-time process. Some

[55] Within the range of open source enterprise software, ATP functionality is in the best case reduced to a basic quantity calculation without any advanced planning capabilities, as in the free ERP suite *Compiere* (Christou and Ponis, 2008, p. 22).

models also provide functionality for due date setting or search for alternative fulfillment options, e.g. by checking the availability of substitute products.

Table 2.10 summarizes the identified demand fulfillment contributions which meet the criteria 'MTS environment' and 'customer heterogeneity'. For each surveyed paper, the model type and the operating mode are given in columns two and three. The fourth column states which form of customer heterogeneity is exploited in the model. Where available, also the industry is specified to which each model applies.

In the following paragraphs, a brief characterization of the papers listed in Table 2.10 will be given. The first two paragraphs present different models for the two operating modes batch and real-time. Then, the introduction of ideas from revenue management into demand fulfillment will be highlighted. The last paragraph addresses models designed for multi-location networks which consist of multiple warehouses.

Batch Demand Fulfillment Models

Fischer (2001) was one of the first authors to consider heterogeneous customers. He analyzed several demand fulfillment approaches using data from the lighting industry. In particular, he described implementations of a simple order promising approach, a batch LP model and a combined allocation and consumption planning model to enable real-time order promising. The batch order promising logic operates on a coefficient which characterizes the suitability of each available ATP quantity to fulfill a particular order, taking into account both order profitability and customer importance as well as lost sales costs. While Fischer found the batch approach beneficial, his LP model has met some critique regarding the use of the 'suitability coefficient'. Pibernik (2002) criticized the use of overlapping suitability criteria, non-trivial aggregation weights and the reliance on identical lost sales costs for all priority groups. He presented an alternative batch model which exploited differences in order-specific contribution margins and penalty costs.

Fleischmann and Meyr (2003, 2004) and Günther and Tempelmeier (2003) described basic batch models for the order acceptance decision in MTS environments where individual orders have different values or costs. Pibernik (2003) presented a multi-product order acceptance and due date setting model for MTS. His model is an adaption of a prior paper by Chen et al. (2001) originally designed for ATO/MTO environments. A more extensive version of the former model has also been presented in Pibernik (2005).

Lečić-Cvetković et al. (2010) provided an outline of a heuristic algorithm to make order acceptance decisions for a batch of orders. Their paper is one of the few contributions which contain a rudimentary clustering algorithm to form customer segments. However, the impact of their approach remains unclear as they neither state any objective functions nor give quantitative results for their method. An alternative approach to clustering customers has been presented by Meyr (2008). The application of this latter approach has been described in Meyr (2009) (see below).

Paper	Model	Operating mode	Customer heterogeneity	Industry	Comments
Fischer (2001)	MIP	Batch + Real-time	Suitability coefficient[a]	Lighting	
Pibernik (2002)	MIP	Batch	Margin, penalty		Max. 2 deliveries
Fleischmann and Meyr (2003)	LP	Batch	Cost		
Günther and Tempelmeier (2003)	MIP	Batch	Penalty		
Fleischmann and Meyr (2004)	LP	Batch	Cost		Also network model
Pibernik (2003)	MIP	Batch	Profit, penalty		
Ball et al. (2004)	LP	Real-time	Profit		
Pibernik (2005)	MIP	Batch	Profit, penalty		
Lee et al. (2006)	Allocation rules	Batch	Priority, volume		Neural network approach
Pibernik (2006)	MIP	Batch + Real-time	Priority	Pharma	
Kilger and Meyr (2008)	Allocation rules	Real-time	Priority, volume		
Meyr (2009)	LP	Real-time	Value[b]	Lighting	
Pibernik and Yadav (2009)	Non-linear program	Real-time	Priority	Electronics	Service-level perspective
Quante (2009), Quante et al. (2009a)	Stochastic dynamic program	Real-time	Revenue		Revenue mgmt. model
Dhakar et al. (2010)	Clustering + allocation rules	Batch	Promise / order date, volume	Apparel	
Huaili and Yanrong (2010)	Rule-based	Real-time	Profit		
Jung (2010)	LP	Batch	Priority, penalty	TFT-LCD	Network model
Lečić-Cvetković et al. (2010)	Clustering + allocation rules	Batch	Not formalized[c]		Service-level
Nguyen et al. (2012)	LP	Batch	Profit, penalty		Network model

Table 2.10. – Literature overview: Demand fulfillment models with heterogeneous customers for MTS environments

[a] Weighted average of profit and priority factor.
[b] Piece-wise linear function of the strategic value of the order (unit profit + assessment of the strategic importance of the customer) and a penalty to account for early or late delivery or order denial; similar to the suitability coefficient of Fischer (2001).
[c] Criteria mentioned include revenue, profit, development potential, service rate, partnership level.

Real-Time Demand Fulfillment Models with an Allocation Planning Step

While allocation planning-based approaches have previously been addressed by software vendors, the dissertation by Fischer (2001) is again the first reference in the academic literature. Another basic push-based allocation model for multiple demand classes was given in Ball et al. (2004). In a technical sense, their model is suited to MTS supply chains as it allocates final items to demand classes. But at the same time, the model also supports decisions regarding material inventories and production capacities and thus better fits an ATO environment, as noted by Meyr (2009). Unfortunately, Ball et al. did not test their model with real-world data. However, due to the significantly higher number of possible customer-product configurations in ATO than in MTS, it is reasonable to expect that push-based allocation in ATO will only yield satisfying results if the bill of material (BOM) has a very simple structure, if only few customer classes exist and generally, if all required forecasts have very low errors.

Pibernik (2006) described several order promising mechanisms and discussed modeling aspects for the short- and long-term costs associated with insufficient order fulfillment. He made the proposition to change from single-order processing to a batch approach in the case of shortages. Additionally, his paper contains a description of an inventory pre-allocation logic based on an ordinal ranking of (given) customer classes to enable real-time order promising. Unfortunately only a naïve reservation policy for this allocation planning step has been tested in the case study.

Kilger and Meyr (2008) provided an overview of demand fulfillment approaches commonly supported by APS, focusing on the allocation planning step. They illustrated various allocation rules to split the available ATP quantities among the competing customer classes. In contrast to the remainder of the literature, they accounted for the case where the customer segments are structured according to a tree hierarchy; for example, as a result of splitting customers on a geographical basis (continents, countries, sales districts). It is reasonable to assume that the individual segments at the leaf nodes of such a tree differ in terms of profitability as a result of different transportation costs, taxes or exchange rate differences. As already highlighted in Section 1.2, Kilger and Meyr illustrated the use of a rank-based, proportional and fixed split allocation rule, but offered no quantitative analyses of these policies.

This gap was partially addressed by Huaili and Yanrong (2010) who assessed the performance of rank-based, proportional and fixed-split allocation planning schemes in a setting with a flat partitioning of the customer base. They compared these rule-based allocation methods with two benchmarks, an ex-post best case strategy under full information and with a naïve FCFS strategy. The contribution of their paper lies in attempting to quantify the loss in total profit and overall service level which results from using any of the rule-based allocation planning approaches under different shortage rates and forecast errors, compared to an ex-post global optimization. However, Huaili and Yanrong fell short of a systematic assessment of the available levers and did not provide any managerial recommendations. A related study addressing rule-based allocation schemes is Dhakar

et al. (2010). They showed how a very basic pre-allocation rule can be used to smooth the workload over time in a distribution center.

Meyr (2009) presented deterministic allocation and consumption planning models based on LP formulations and reported comprehensive numerical results. His characterization of customer heterogeneity is similar to the approach taken by Fischer (2001). There is a time-independent component as customers are assumed to possess a specific unit profit contribution; and customers also differ in terms of their strategic importance. However, the latter term is not formalized in the paper. In addition to these customer-specific/time-independent components, there is a time-dependent component to account for early or late delivery as well as order denial. The allocation planning and consumption planning models of Meyr (2009) have already been summarized in Section 2.4.3. The author showed that substantial improvements can be achieved by an allocation planning-based approach compared to simple FCFS order acceptance. However, his deterministic setting is particularly susceptible to forecast errors.

A special aspect of forecast errors has been studied in the paper by Lee et al. (2006). They investigated allocation rules for situations in which forecasts for customer segments are biased. Instead of considering bias due to strategic gaming, the authors focused on what they called 'surplus demand'. In essence, this surplus demand results if local planners, when preparing forecasts of the expected demand in their customer segment, deliberately report a higher demand forecast (compared to the unbiased point forecast quantities) to account for prognosis errors.[56] As the sum of the regular and these surplus demands of all customer segments often exceed the available capacity, an allocation is required. The paper by Lee et al. (2006) describes how a basic neural network-based approach may be employed to identify these surplus demands. They went on to illustrate how adjusted allocations may be calculated if any of the three basic rule-based allocation policies rank-based, proportional and fixed split as described by Kilger and Meyr (2008) are used.

Revenue Management-Based Demand Fulfillment Approaches

A more rigorous approach to better account for the stochasticity of the customer demands and to consider forecast errors already in the allocation planning step was suggested by Ball et al. (2004). They observed that the push-based demand fulfillment approaches with an allocation planning step can be viewed as a type of yield or *revenue management* problem. Revenue management is a set of methods and procedures to maximize the yield or revenue and is typically encountered in service industries. Here, a fixed capacity needs to be utilized as efficiently as possible to maximize profits. The main components of revenue management systems include mechanisms for capacity (re-)allocation, dynamic pricing as a means to influence external demand and overbooking rules (see Weatherford and Bodily, 1992). A standard revenue management problem is to decide which service requests to accept if the requests belong to different service or fare classes. These fare classes have different revenue potentials. By contrast, cost differentials can usually be

[56] This resembles the determination of safety stocks in inventory management.

neglected in many service industries. The necessary capacity to render the service (e.g. airplane seats, hotel rooms) is costly to establish and to adjust. Furthermore, this capacity is usually perishable, thus unused capacity is lost (see Belobaba, 1987; Brumelle et al., 1990; Weatherford and Bodily, 1992). For comprehensive introductions into the key revenue management concepts, see McGill and van Ryzin (1999) and Talluri and van Ryzin (2004).

Harris and Pinder (1995) were the first to transfer these basic revenue management concepts from service operations to manufacturing settings. They presented basic stochastic models for the determination of critical reservation levels, for optimal pricing and capacity reallocation decisions for a static revenue management problem with multiple customer classes. In manufacturing settings, the focus is less on exploiting differentials in the willingness to pay, but rather on differentiating between different levels of strategic importance as measured by tangible and intangible costs. Quante et al. (2009b) highlighted that the planning tasks of demand fulfillment in manufacturing environments largely correspond to the problem in service industries of choosing among service requests of different priorities. Their paper also provided a framework for analysis and detailed discussions of the interrelations between demand fulfillment, revenue management and inventory rationing.

In many MTO manufacturing environments, the similarities to revenue management in service industries are particularly close. Assembly capacity is often the key bottleneck as it is usually difficult to adjust in the short-term and lost if unused (the availability of materials is typically less constrained). Examples for demand fulfillment and order promising approaches in MTO environments can be found, e.g., in Spengler et al. (2007) or Volling and Spengler (2011). The application of revenue management ideas is more difficult in MTS environments as the allocation decision has to be made regarding the allotment of final items. These usually do not perish in the short- and medium term and thus inventory holding decisions must be included, leading to a significantly larger state space and resulting in computational challenges.[57]

The task of introducing revenue management ideas into an MTS demand fulfillment setting has been undertaken in Quante (2009) and Quante et al. (2009a). They explicitly modeled the stochasticity of customer demands. After deriving the optimal fulfillment policy, the authors demonstrated the superiority of their stochastic approach compared to a deterministic allocation planning step as well as in comparison to FCFS. While the underlying idea of this revenue management-based approach has its merits, a few drawbacks still need to be addressed. For example, the approach is computationally expensive, relies on the assumption of a particular (given) stochastic process and cannot account for different backorder and lost sale costs per customer segment.

[57] In ATO environments, both scarce perishable capacities and storable inventories of materials and subassemblies need to be considered simultaneously. See Chen et al. (2001), Ervolina et al. (2007) or Robinson and Carlson (2007) for examples.

Demand Fulfillment Models for Networks of Multiple Warehouse Locations

Few papers on demand fulfillment explicitly provide for the opportunity to satisfy an order request from alternative locations. In a recent overview paper, Nguyen et al. (2012) identified only seven contributions for network demand fulfillment in an MTS/MSTS environment. Of those, just three consider customer heterogeneity. A few more network demand fulfillment models exist which apply to ATO/MTO environments (e.g. Jeong et al. (2002), Tsai and Wang (2009) or Venkatadri et al. (2008)). However, the focus in these contributions lies more on the due date setting functionality and the order scheduling & control problem in a network. Furthermore, network demand fulfillment models for ATO/MTO environments are often tailored to specific industries such as microelectronics or TFT/LCD assembly.

The three papers for network demand fulfillment in MTS environments with heterogeneous orders represent classical optimization models of the LP type. The models use the flexibility of the batch approach to exploit different aspects of network-related fulfillment alternatives: A first basic model has been described by Fleischmann and Meyr (2004) which only considered general cost differentials associated with different fulfillment alternatives without specifying any detailed cost components. In the extended version described in Nguyen et al. (2012), both transportation time and costs were included; additionally, the model provided for different transportation modes. Jung (2010) described a practical application in the TFT/LCD industry. His model focused on transportation time and contained capacity constraints. In contrast to single-location models, no real-time models with prior allocation planning have yet been presented for the network case.

Overall, demand fulfillment is an established area of research. More recent publications have focused particularly on issuing order confirmations in real-time and on incorporating the stochasticity of the underlying demand processes. Differences in the importance of individual orders or customers can be accounted for in a prior allocation planning step and can be respected by flexible consumption search rules. These approaches have the potential to improve overall company profits and to raise customer service levels for the most important customer segments.

However, given the many different (and often conflicting) design alternatives, there is room for more quantitative assessments to derive concrete recommendations to improve the use of demand fulfillment systems in practice. For example, which allocation schemes at which temporal granularity should be employed given a certain level of forecast error? Which consumption rules are appropriate?

A major limitation of most current demand fulfillment models is the assumption of a flat partitioning of the customer segments with central control. This allows sorting the individual classes unambiguously and to determine an optimal fulfillment decision. In practice, however, multi-stage customer hierarchies with distributed decision-making prevail. Fulfillment decisions often have to be made on a decentral basis and without full data transparency. Then, the ranking of the customer segments in terms of fulfillment priorities is rarely obvious. Pibernik (2006, p. 730) noted that an allocation planning

scheme for practical applications will have a hierarchical structure. He suggested employing a fixed-split rule to first allocate supply quantities to sales regions and then to refine these regional allocations to the level of customer segments. This idea will be studied in the following chapters in an attempt to extend allocation planning and consumption rules to such hierarchical settings.

2.5. Conclusions: Allocation Problems in Supply Chain Planning

In this chapter, an overview of the key tasks in SCP has been provided and a differentiation between the major planning tasks at the different planning levels has been presented. To achieve the overall objective of SCM, i.e. to profitably match supply with demands to satisfy customer needs, a number of interrelated decision problems need to be aligned and solved. The concept of hierarchical planning has been shown to be an adequate planning framework both for theoretical and practical reasons. A prime argument is that it allows for decision postponement. This is particularly relevant in demand fulfillment where the actual value of each customer order is only known upon its arrival. However, a number of important decisions need to be made well in advance, particularly in an MTS environment with forecast-driven manufacturing processes. As a consequence of capacitated production, it is important to exploit the heterogeneity among the arriving order requests to maximize overall profits in the supply chain. Order acceptance decisions should be made dependent on the relative value and importance of each individual request. To allow for a real-time process, most supply chains employ a series of prior allocation and refinement decisions which ultimately determine which customer orders will be served.

These refinements are an integral characteristic of the hierarchical planning logic. Recall that at a long-term planning level, companies choose among the markets to serve. Planners specify the geographical areas, sales channels and aggregate customer segments to address. As only a limited amount of (usually unreliable) demand data is available at this planning stage, customer heterogeneity is only considered coarsely, e.g. by assessing the rough sales and profit potential in the relevant segments.

At a mid-term level, planning is usually done with the objective of satisfying all demands which have been predicted by demand planning. Hence, master planning determines aggregate production quantities, sales quantities per customer region as well as inventory replenishment quantities per warehouse location. If shortages occur and if an allocation is required, many mid-term planning models exploit differences in direct profits between competing customer segments. Often, decisions are made based on a geography-based segmentation for which revenue and cost differentials can be established easily. For example, different regions may have different sales prices, distribution costs, transportation costs, salesforce compensation schemes, etc. Minimum service level constraints may be set to ensure that sales regions with strategic importance are served adequately. This case may arise if certain regions have promising long-term sales and profit potential although they are currently less profitable than other geographies.

At a shorter time horizon, distribution planning and deployment provide for a further refinement of master planning quantities, e.g. by determining replenishments to individual sites within aggregate regions. Here, the focus is predominantly on transportation issues. Heterogeneity among customer segments is again primarily respected to the extent that it relates to direct cost differentials from employing the existing supply network.[58] Neither individual orders nor single customers are considered.

Allocation planning as introduced in Section 2.4.3 complements these mid- and short-term allocation decisions by introducing a broader notion of customer heterogeneity. Allocation planning models also consider the short- and long-term costs of suboptimal customer service. As stated by Ball et al. (2004, p. 461), allocation planning fills a niche downstream in time from inventory control. It is complemented by consumption planning for the immediate short-term decision of how to fulfill a particular order request.

Overall, a series of allocation problems is solved in supply chain planning to continuously match demand with supply. Each associated model refines the decisions which have been made at the prior, i.e. higher planning level. This decomposition is necessary due to the staggered availability of reliable demand and profitability data. Allocation and refinement decisions are postponed until more accurate and more detailed demand data becomes available, starting at an aggregate geographical level and proceeding to smaller customer segments until individual order data is available.

This hierarchical allocation and refinement process in supply chain planning has been summarized in Table 2.11. For each planning problem, the table states the typical time-horizon, lists the key decision variables and indicates how customer heterogeneity is accounted for. As a result, the successive refinement from the global markets to address down to the question of which orders to accept becomes obvious.

The DMC problem as introduced in the introductory Chapter 1 can be interpreted as a generalized form of the allocation and consumption planning problems. Considering the more limited time-horizon in the mid- and short-term, actual product and sales volumes need to be allocated within many sales organizations to match a given amount of supply with volatile demands and orders of individual customers. The research presented in the following chapters differs from previous contributions by explicitly accounting for the hierarchical structure within the customer base.

[58] For a model formulation, see Grunow and Farahani (2012).

Planning problem	**Time horizon**	**Allocation entities / decision variables**	**Form of Customer Heterogeneity**
Strategic network planning	Long-term	Markets to address	Revenue and profit potential
Master planning	Mid-term	Aggregate product quantities per sales region	Revenues and direct costs
Deployment / inventory control	mid-term / short-term	Replenishment quantities per warehouse	Direct costs
Allocation planning	Mid-term / short-term	Sales quotas per customer segment	Average unit profits + indirect penalty costs
Consumption planning	Short-term	Orders to fulfill	Unit profits + indirect penalty costs

Table 2.11. – Allocations and quota refinements in hierarchical supply chain planning