3. Multi-Stage Customer Hierarchies

The purpose of this chapter is to provide an introduction and background perspective on a number of properties and behavioral aspects associated with hierarchical structures in general and with customer hierarchies in particular. The characterization of these basic aspects allows for a more concise discussion of the DMC problem in the following chapters.

The five sections of this chapter can be subdivided into two parts: A step-by-step introduction to customer hierarchies (Sections 3.1–3.3) and a characterization of two fundamental management problems found in customer hierarchies (Sections 3.4 and 3.5).

- In the first part, various facets of hierarchies will be discussed, gradually moving from an abstract representation of hierarchies to the specifics of customer hierarchies. The analysis in Section 3.1 examines hierarchical structures from a purely formal point of view. Basic definitions will be provided, a notation will be introduced and aggregation and disaggregation operators in general hierarchies will be discussed. Section 3.2 contains a comprehensive framework from the literature illustrating four key aspects of distributed decision-making in hierarchies. Such decentralized planning is typical of the DMC problem. Finally, in Section 3.3, heterogeneous customer hierarchies will be introduced formally.

- The second part provides a more in-depth coverage of two management problems of customer hierarchies which are relevant for the following chapters. First, Section 3.4 will address agency problems and forecast misrepresentations in customer hierarchies, e.g. caused by dishonest sales agents reporting biased sales forecasts. The incentive properties of common compensation schemes will be analyzed and several options to remediate agency problems in customer hierarchies will be discussed.

  Then, different approaches to measure the actual level of heterogeneity in a given customer hierarchy will be discussed in Section 3.5. The most promising approach, Theil's index T, will be used later in Section 4.4 to establish a novel allocation scheme for customer hierarchies.

3.1. Formal Hierarchies

Section 3.1.1 starts with basic definitions and by introducing notations for multi-level trees and hierarchies. Then, a differentiation between hierarchy types will be given. A key problem in hierarchical planning and hierarchical forecasting is the need to distribute a given quantity to multiple lower-level receiving objects or to perform aggregation to a

S. Vogel, *Demand Fulfillment in Multi-Stage Customer Hierarchies,* Produktion und Logistik,
DOI 10.1007/978-3-658-02864-0_3, © Springer Fachmedien Wiesbaden 2014

higher hierarchical level. Hence, in Section 3.1.2, a brief introduction into aggregation operators and their inverse, disaggregation or allocation operators, will be given.

3.1.1. Trees and Hierarchies

To facilitate the subsequent discussions, it is essential to first define a hierarchy. An important building block for a hierarchy is a mathematical tree. Following Radner (1992, p. 1390), such a tree can be defined as follows:

Definition 3. *A tree is a set of nodes, denoted by the symbol $\mathcal{N}$, which are related to one another by superiority relations. Trees fulfill the following key properties:*

- *Transitivity*: If node a is superior to node b (denoted as $a \succ b$) and if $b \succ c$, then $a \succ c$.
- *Anti-Symmetry*: If $a \succ b$ and if $b \not\succ a$ (b is not superior to a), then b is subordinate to a.
- *Existence of a Unique Root*: There is a single, unique root node which is superior to all other nodes. This root will be referred to as node 0.
- *Unambiguous Parent Node*: Except for the root, each node b has exactly one immediate superior node a. Node a is the immediate superior node ('parent') of b if there is no node 'in between'.

The reverse of the last property is not true, so a particular node may be the parent node to multiple successor nodes. Generally, the set of immediate successor nodes of node a will be denoted as $\mathcal{D}_a$. All nodes which do not possess any successor nodes will be jointly referred to as the set of *leaf nodes* $\mathcal{L} \subset \mathcal{N}$ and obviously, $\mathcal{D}_l = \emptyset$ for all $l \in \mathcal{L}$. When particular nodes need to be indexed, the index i will denote all types of nodes, the index k will be used to refer to intermediate nodes from the set $\mathcal{N} \setminus \mathcal{L}$ whereas the index l will be reserved to indicate leaf nodes from the set $\mathcal{L}$.

These parent-successor relationships imply that a tree is a *partially ordered set*. Trees are not completely ordered sets since no parent-successor relationships can be established between nodes which are positioned in different branches. Nevertheless, to still establish a relationship between nodes in different branches, the concept of *levels* will be introduced. The level (or rank) of a particular node a corresponds to the number of individual edges on the direct path between node a and the root node 0. Hence, the root node 0, which has no parent node, corresponds to level 0 and all its immediate successor nodes in $\mathcal{D}_0$ correspond to level 1. This establishes a relation between nodes without direct parent-successor relationships in different branches of the tree. Nodes at the same level are thus positioned at the same distance from the root node.

Overall, there are M levels and the lowest level is $M-1$. By definition, all nodes at level $M-1$ are leaf nodes, but not all leaf nodes have to be positioned at level $M-1$, as will be discussed below. Note that other authors often assign the lowest level 0 to the

leaf nodes which are the farthest away from the root node. Yet, the definition used here[1] facilitates some notation which will be introduced below. Nevertheless, attention needs to be given to ensure correct verbal references. In line with common practice, a *pyramidal* graphical representation of trees and hierarchies will be assumed in the following. Here, the leaf nodes are found in the bottom part of the pyramid. Hence, 'lower' hierarchy levels such as those containing the leaf nodes are associated with large level indices m whereas 'higher' levels are associated with small indices.

The introduction of levels now permits defining a hierarchy (see Radner, 1992, p. 1391):

Definition 4. *A hierarchy is a ranked tree where each node has been assigned to a particular level specifying its distance from the root node.*

This definition of a hierarchy still allows for a number of conceptually different hierarchy types. Unfortunately, no dominant classification system has yet emerged as researchers and software vendors employ a number of different (and sometimes contradicting) names for the same types of hierarchies (see e.g. Malinowski and Zimányi (2006) for several examples). For convenience, the following terminology will be introduced with the help of Figure 3.1:

Balanced hierarchy In a balanced hierarchy, the path from each of the leaf nodes to the root node traverses along the same number of edges. Essentially, all leaf nodes are positioned at the same level of the hierarchy. The hierarchies in Figures 3.1a–3.1c are balanced.

Symmetric hierarchy A symmetric hierarchy is a special type of a balanced hierarchy where all nodes *at a particular level* m have the same number of successor nodes c at the next lower level, i.e. $|\mathcal{D}_k| = c$ for all k at level m. Both Figures 3.1b and 3.1c represent symmetric hierarchies.

Uniform hierarchy A uniform hierarchy is a symmetric hierarchy in which the number of successor nodes per intermediate node $k \in \mathcal{N} \setminus \mathcal{L}$ is identical at all levels of the hierarchy (except for the lowest where there are no successor nodes by definition). An example is depicted in Figure 3.1c with $|\mathcal{D}_k| = 2$ for all $k \in \mathcal{N} \setminus \mathcal{L}$.

Unbalanced hierarchy In an unbalanced hierarchy, the number of edges which need to be traversed from a leaf node to the root node are not identical for all possible paths. For example, in Figures 3.1d and 3.1e, the edge-count of the left leaf node to the root node equals one whereas both other leaf nodes in these two hierarchies have an edge-count of two.

Ragged hierarchy In a ragged (and simultaneously unbalanced) hierarchy, at least one path from a leaf node to the root skips one level. The number of edges on this path is less than the level of the associated leaf node, as depicted in Figure 3.1e. Ragged hierarchies are problematic in the context of aggregation and allocations since one

[1] Athanasopoulos et al. (2009) used a similar definition.

or several intermediate nodes are not available. An example of a ragged customer hierarchy will be discussed later in Section 3.3.2.

Independent of the actual hierarchy type, the main differentiating aspect between a hierarchy and a tree is that the constituent hierarchy nodes correspond to certain levels. Nodes at different levels therefore have a characteristic mutual relationship. The next section focuses on such hierarchical relationships, in particular on the aggregation and the disaggregation problem.

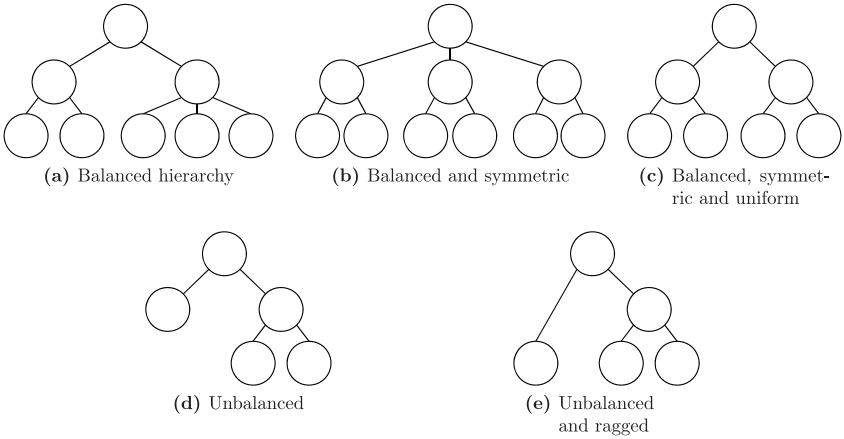


Figure 3.1. – Different hierarchy types

3.1.2. Aggregation and Disaggregation

In the course of this thesis, hierarchies will primarily be used to represent data at different levels of abstraction. In such a setting, the key problem consists of characterizing the relations between data elements at different hierarchical levels. From the perspective of a lower hierarchical level, there is an *aggregation problem* which consists of finding a higher-level representative (aggregate or parent). Essentially, in aggregating, one desires to trade complex, detailed information for less complex, compressed information. As will be shown shortly, two types of data need to be distinguished: In the case of *summable values*, the aggregation function and thus the parent value are unambiguous. In the following, also the case of *non-summable values* will be considered. Here, the aggregation problem is no longer trivial and multiple operators are available to choose from.

A reciprocal situation exists if such an aggregate representative is available which needs to be distributed to a corresponding group of lower-level data objects (*disaggregation problem*). In that sense, the disaggregation is an inverse to the aggregation operation. Again, two situations may occur, depending on the type of data considered: While the

aggregation function in case of summable values implies an unambiguous parent object, it is the disaggregation (or allocation) procedure which is no longer unique. It may lead to many different possible solutions for the lower-level data objects, differing in the detailed information which is added in the process. Contrariwise, in the case of non-summable values, the disaggregation problem is again trivial while the aggregation problem often imposes challenges.

The most important areas of application in practice are the aggregation and disaggregation of numerical values such as demand forecasts, prices or (forecast) error measures. Examples of the aggregation and disaggregation problems have already been encountered in Section 2.1.2 when discussing hierarchical planning. Similar problems arise in hierarchical forecasting which have been discussed in Section 2.2.5. In the following, a general perspective on aggregation functions for such applications will be given, followed by corresponding comments on disaggregation—or allocation—functions.

Aggregation Functions

It is helpful to start by introducing a formal terminology. To characterize an aggregation function from a rather general perspective (i.e. in a wide sense), the following basic definition will be used (e.g. see Beliakov et al., 2007):

Definition 5. *An aggregation function (or operator) in the wide sense assigns a representative real number y to any n-tuple $(x_1, x_2, \ldots, x_n)$ of real numbers, i.e.*

$$y = \text{agg}\,(x_1, x_2, \ldots, x_n). \tag{3.1}$$

In many practical aggregation problems, the lower-level objects represent *summable quantities* x_i, e.g. product units or capacities. The aggregated value y has to stand for the entire group of the individual values x_i. A straightforward aggregation operator is the simple sum,

$$y = \sum_{i=1}^{n} x_i. \tag{3.2}$$

The problem is more challenging if the lower-level objects correspond to *non-summable quantities* such as unit prices or unit margins. Here, it is required to determine a *representative* quantity which may stand for any of the disaggregate values at the lower level. The aggregation of such non-summable values corresponds to what is typically referred to when speaking of a *mathematical aggregation function* (e.g. see Detyniecki (2001) or Beliakov et al. (2007)).

Grouping several previous definitions, Calvo et al. (2002) proposed a set of fundamental properties which have to be fulfilled by such a mathematical aggregation operator. Assuming that the disaggregate numbers x_i, z_i ($i = 1, \ldots, n$ and $0 \leq x_i, z_i \leq a$) are defined on the interval $[0; a]$, the operator

$$\text{agg} : \bigcup_{n \in \mathbb{N}} [0; a]^n \to [0; a] \tag{3.3}$$

satisfies:

$$\text{agg}\,(x_i) = x_i \tag{3.4}$$

$$\text{agg}\,(0,\ldots,0) = 0 \;\text{ and }\; \text{agg}\,(a,\ldots,a) = a \tag{3.5}$$

$$\text{agg}\,(x_1,\ldots,x_n) \leq \text{agg}\,(z_1,\ldots,z_n) \;\text{ if } x_1 \leq z_1,\ldots,x_n \leq z_n \tag{3.6}$$

The boundary property (3.5) and the monotonicity condition (3.6) together imply that the aggregated value should always lie between the extreme points of the disaggregate values, i.e.

$$\min_{i=1,\ldots,n}(x_i) \leq \text{agg}\,(x_1,\ldots,x_n) \leq \max_{i=1,\ldots,n}(x_i). \tag{3.7}$$

Obviously, property (3.5) and consequently also this last property (3.7) are not fulfilled by the simple sum (3.2). For convenience, the following additional distinction will be used:

Definition 6. *An aggregation function in the narrow sense differs from Definition 5 by fulfilling at least also the properties* (3.4)–(3.6).

The most obvious approach to defining an aggregation operator in the narrow sense is to take some 'middle value' of the arguments x_i. If the x_i are ordered such that the smallest argument corresponds to x_1 and the largest argument corresponds to x_n, a simple representative for all n arguments is their middle value, i.e. the median, defined as

$$\text{median}\,(x_1,\ldots,x_n) = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{\frac{n}{2}+1} + x_{\frac{n}{2}}\right) & \text{if } n \text{ is even.} \end{cases} \tag{3.8}$$

Another simple operator which fulfills the properties (3.4)–(3.6) is the average value or *arithmetic mean*, i.e.

$$\text{agg}\,(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{3.9}$$

The arithmetic mean can be extended by placing non-negative weights w_i on the arguments x_i. This leads to the *weighted average*

$$\text{agg}\,(x_1,\ldots,x_n) = \sum_{i=1}^{n}(w_i \cdot x_i) \tag{3.10}$$

where the weights sum to unity, i.e. $\sum_{i=1}^{n} w_i = 1$. The arithmetic mean (and also its weighted version) can be extended further into an entire family of *quasi-arithmetic* means,[2] defined as follows:

$$\text{agg}\,(x_1,\ldots,x_n) = \left[\frac{1}{n}\sum_{i=1}^{n} x_i^{\alpha}\right]^{\frac{1}{\alpha}} = \left[\sum_{i=1}^{n}\left(\frac{1}{n}x_i^{\alpha}\right)\right]^{\frac{1}{\alpha}}. \tag{3.11}$$

[2] This has been shown simultaneously by Kolmogorov (1930) and Nagumo (1930).

By choosing different values for the parameter α, different alternative aggregation operators result. The most frequently used operators are summarized in Table 3.1, taken from Detyniecki (2001, p. 13). With $\alpha = 1$, Equation (3.11) turns into (3.9) for the arithmetic mean. For practical applications, aggregation operators need to be as simple as possible,

Parameter α	Operator	Functional term
$\alpha = 1$	Arithmetic mean	$\frac{1}{n}\sum_{i=1}^{n} x_i$
$\alpha = -1$	Harmonic mean	$\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}$
$\alpha \to 0$	Geometric mean	$\sqrt[n]{\prod_{i=1}^{n} x_i}$
$\alpha \to +\infty$	Minimum	$\min_{i=1,\ldots,n}(x_i)$
$\alpha \to -\infty$	Maximum	$\max_{i=1,\ldots,n}(x_i)$

Table 3.1. – Aggregation operators of the class of quasi-arithmetic means

but also have to lend themselves to an intuitive interpretation. Overall, this favors the use of the arithmetic mean with according weights. Consider two products a and b which form a joint product family. Sales volumes equal $q_a = 90$ and $q_b = 10$ and unit prices are given by $p_a = 3$ and $p_b = 1$. Clearly, when looking for an aggregate price at a product family level, the simple arithmetic mean $p = 2$ is misleading since the sales volume of a is significantly larger than that of the low-price product b. A demand-weighted average via (3.10) with weights $w_a = \frac{90}{90+10}$ and $w_b = \frac{10}{100}$ leads to $p_{agg} = 2.8$. This aggregation approach is usually perceived to be more appropriate.

Using the weighted average to aggregate non-summable figures is also in line with the case studies which have been reported in the literature. Vollmann et al. (2005, p. 41) suggested determining aggregate prices in hierarchical forecasting via the demand-weighted arithmetic mean price, Roitsch and Meyr (2008, p. 410) adopted the same approach for unit profits in their case study. As mentioned in Section 2.2.5, Kilger and Wagner (2008, p. 155) defined an aggregate price by dividing aggregate revenues by aggregate quantities. Using the example given above, this leads to $p_{agg} = \frac{90 \cdot 3 + 10 \cdot 1}{90+10} = 2.8$. Essentially, their approach also corresponds to a demand-weighted average.

The use of other aggregation operators for non-summable values is extremely rare. One notable exception is Andrawis et al. (2011) who focused on forecast pooling. They tested different approaches to combine multiple forecasts for the same time series (in their case, inbound tourists to Egypt from monthly and from annual data). As discussed in Section 2.2.3, forecast pooling may reduce the overall forecast error and the conventional approach is to use a (weighted) arithmetic average. Among other schemes, Andrawis et al. (2011) also tested the geometric and the harmonic mean to combine two forecast values and found them to lead to surprisingly competitive results in several of their test instances.

Nevertheless, there is sufficient evidence in the literature and in practice that quantity-weighted arithmetic means are an adequate means to determine aggregate prices and profits.[3] Hence, this method will also be used in the following to aggregate non-summable quantities.[4]

Disaggregation and Allocation Functions

As stated above, disaggregation can be thought of as the inverse operation to aggregation. In the same manner as in case of the wide-sense aggregation function, also a wide-sense disaggregation function can be defined. The following definition will be used:

Definition 7. *For a given aggregate real number y, a disaggregation function in the wide sense determines an n-tuple $(x_1, x_2, \ldots, x_n)$ of values for the n successor objects of y at the next lower hierarchical level.*

In analogy to the aggregation case, it is useful to also differentiate between the disaggregation of summable and non-summable figures. For the disaggregation of summable quantities, a disaggregation in the narrow sense can be defined:

Definition 8. *A disaggregation function in the narrow sense divides or splits an aggregate quantity y into an n-tuple $(x_1, x_2, \ldots, x_n)$ of real numbers such that $\sum_{i=1}^{n} x_i \leq y$. This operation will also be denoted as an allocation.*

It is helpful to require that a few basic properties hold for narrow-sense disaggregation operators. An allocation is *feasible* if it leads to non-negative allotments, i.e. if $x_i \geq 0$ for all i. An allocation is *efficient* (or *tidy*, see Demski (1981)) if the entire aggregate quantity y is allocated, i.e. if the following condition holds:

$$\sum_{i=1}^{n} x_i = y. \tag{3.12}$$

The simplest allocation operator which satisfies the above properties is the equal split, i.e.

$$x_i = \frac{y}{n}, \quad \forall i = 1, \ldots, n. \tag{3.13}$$

Now consider the case of non-summable quantities such as prices or margins. If an aggregate value y is available, recall that by definition this value is representative for all objects at the lower level. Hence, this value can also be used to describe each object at the next lower level. An example of a possible disaggregation is therefore simply

$$x_i = y, \quad \forall i = 1, \ldots, n. \tag{3.14}$$

[3] Furthermore, quantity-weighted arithmetic averages are also used in econometric applications to determine price indices for inflation measurement, see Afriat (2005) for more details.

[4] The use of other aggregation operators for non-summable quantities may constitute an option for follow-on work.

It becomes clear that in the case of non-summable figures such as unit prices or unit profits, there is always a trivial disaggregation, but the (narrow-sense) aggregation is usually more involved. The reverse is true for summable values such as product quantities. In that case, there is always a trivial aggregation via the summation, but the (narrow-sense) disaggregation process constitutes a more challenging task. Essentially, one has to re-introduce detailed lower-level information which was lost in the prior aggregation step.

Later, in Section 4.2, a comprehensive overview of existing allocation schemes for summable quantities will be provided. As these schemes are primarily employed in multi-entity, hierarchical settings, with distributed rather than centralized decisions, it is necessary to first advance the scope of this discussion from purely formal hierarchies to multi-entity organizations which possess a hierarchical structure. This step will be explored in the following section.

3.2. Decision-Making in Hierarchies

While the previous section has exclusively addressed hierarchies from a formal, i.e. from a mathematical perspective, the focus will shift now to hierarchical structures which are employed for decision-making, more precisely for *distributed decision-making*. It is the purpose of this section to present a framework allowing for an analytic differentiation between four basic types of hierarchies which are characterized by distributed decision-making. Actual real-world situations, for example customer hierarchies, are hierarchies of mixed types as they typically exhibit the characteristics of more than one of these four basic types. Understanding these basic types will therefore enhance the understanding of the more complex real-world hierarchies with distributed decision-making. Following Schneeweiß (2003, p. 7), a hierarchy with distributed decision-making can be defined as follows:

Definition 9. *In a hierarchy with distributed decision-making, at least two entities have different levels of discretionary power, have asymmetric information statuses, or simply make their decisions at different points in time. These entities are then positioned at different hierarchical levels.*

As pointed out by Schneeweiß, this definition corresponds to the microeconomic *Stackelberg* property which establishes a leader-follower relationship between two entities, one at an upper top-level, one at a lower base-level (see Varian, 1992). A number of different hierarchical situations exist which are characterized by such a leader-follower relationship between two entities.

In a first step, two major cases need to be distinguished.

1. A hierarchical structure can (intentionally) be imposed on a **single decision** problem. This means that multiple entities will be created for the sole purpose of facilitating the process of finding a solution to the single problem.

2. If a multi-entity situation already exists in the first place, it is usually employed continuously to make **multiple decisions** on a decentral basis. These existing

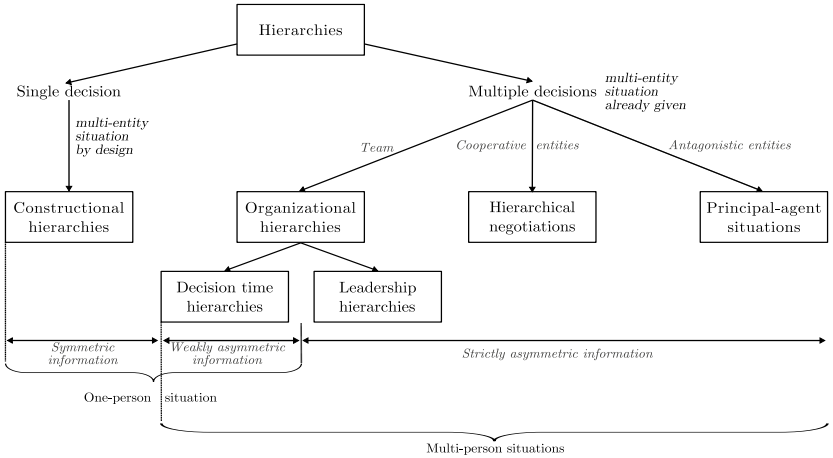


Figure 3.2. – Prototypical hierarchies (adapted from Schneeweiß (2003, Fig. 1.4, p. 10))

entities can be characterized further by their degree of mutual interdependence, i.e. by the degree to which they have similar information endowments, objectives or decision rights.

Based on these two categories, Schneeweiß (2003) has introduced a general analytical framework and identified four prototypical hierarchy types which are characterized by distributed decision-making. These four are *constructional* hierarchies, *organizational* hierarchies, *hierarchical negotiations* and *principal-agent* situations.

A graphical representation of the framework is depicted in Figure 3.2 and will be explained in the following paragraphs. As will become obvious, these hierarchy prototypes are well-suited to illustrate different aspects of decentral planning and decision-making in the DMC problem.

Constructional Hierarchies

A constructional hierarchy corresponds to a situation in which a single decision is split into dependent subproblems without any information asymmetries. It can be seen as an *artificial* hierarchy "which results from imposing a (...) [distributed decision-making] structure on a non-structured system" (Schneeweiß, 2003, p. 7). Constructional hierarchies emerge when breaking a complex problem into a number of related disaggregate subproblems, e.g. to simplify (or even enable) the process of finding a solution to the aggregate problem by reducing the computational or conceptual complexity (see Schneeweiß (2003, p. 73), similarly Okuda (2001, p. 217)). In a strict sense, constructional hierarchies involve only a single decision-maker. The upper and the lower entity, which form the hierarchical relationship, are the result of the specific modeling approach chosen by the single planner. In practical implementations, multiple decision makers may exist who

make decentral decisions. But they fully support a common objective so that an optimal solution to the overall problem can be found.

A typical area of application of constructional hierarchies exists in *mathematical programming*. For example, in the context of splitting a large linear optimization problem into smaller parts, both the so-called price-directive *Dantzig-Wolfe* and the resource-directive *Benders decomposition* are well known. Both approaches split an overall LP problem into a number of simpler lower-level subproblems. These subproblems are chosen such that interdependencies between them will be minimized. The remaining interdependencies (which can often be interpreted as mutually shared resources) are captured in a higher-level master problem, hence the hierarchical relationship.

While in the Dantzig-Wolfe decomposition coordination between the subproblems is achieved by centrally provided (shadow) prices for the shared resources, the Benders approach is based on a centralized, direct allocation of the resource quantities. More precisely, in a price-directive decomposition, the central decision-maker uses quantity proposals from the lower-level entities to determine prices for the shared resources. These prices enter the objective function of each lower-level subproblem. As the lower-level decision-makers equate their marginal benefits from using the resource with the centrally given shadow prices, this leads to an efficient utilization of the scarce resource by the lower-level entities. In a resource-directive scheme, often termed *bi-level planning*, the center itself determines resource allocation suggestions for each lower-level entity. Each of the latter in turn responds by reporting the marginal value which it can generate with the suggested resource allocation (see Kornai and Liptak, 1965). The center thereupon adjusts the resource allocation in order to maximize the overall value creation. Finite convergence of this algorithm has been shown by Freeland (1975).

It has been pointed out already in the early literature that such (mathematical) decompositions in constructional hierarchies can also be interpreted as decentral decision-making processes, e.g. by Dantzig (1963) or Baumol and Fabian (1964, Footnote 1). Moreover, the similarity to allocation processes has been stressed by Burton et al. (1974). A number of authors have focused on the analogy to decision-making in multi-divisional organizations. For example, Jörnsten and Leisten (1994) studied decentralization in an organizational context. They presented an aggregated LP which corresponds to the management problem at the central unit of providing a suitable allocation of a common resource to the lower-level divisions. The latter create their production plan in response to the signals from the center by solving a disaggregated divisional subproblem.

Overall, these analogies show that the resource allocation problem does not necessarily require a central planner with full oversight. Rather, decomposition methods based on constructional hierarchies can ensure that a decentralized allocation approach will also lead to an optimal allocation solution. An essential assumption in this context is that decentral decision makers are honest and that they do not bias any reports, as they are assumed to fully support the common objective at the higher level.

The following paragraphs will focus on the second group of hierarchies with multiple decisions. These have been grouped under the right branch in Figure 3.2. In these sit-

uations, hierarchical relationships between two or more entities are a given fact, rather than being imposed to simplify the solution finding process. The permanent nature of the hierarchies implies that the individual entities encounter distributed decision-making processes on a repeated basis, i.e. multiple decisions are usually made over time. Schneeweiß (1998, p. 548) identified three degrees of mutual interdependence which may exist between such entities at different hierarchical levels:[5]

- In a *team*[6] setting, both the top and the base entity accept and mutually support either common or each other's goals. Team settings are equivalent to various forms of organizational hierarchies.
- On the contrary, in an *antagonistic* situation, both entities pursue their own goals in an opportunistic manner. This setting is frequently studied in principal-agency theory.
- Between these two extreme points, a continuum of intermediate positions exists. For example, a *cooperative* situation is characterized by both levels behaving like a team, provided that some private aspiration levels are fulfilled. This situation may be encountered in *hierarchical negotiations*.

Organizational Hierarchies

First, assume that both entities pursue similar objectives, either as a real team with truly congruent objectives, or as an enforced team with the upper level exerting managerial authority over the lower level. Using the terminology of Schneeweiß, these settings are referred to as *organizational hierarchies*. However, as will be clear shortly, organizational hierarchies may also involve the case of only a single decision-maker.

Two sub-types of organizational hierarchies can be distinguished by considering two kinds of information asymmetry which may exist between the entities. A *decision time hierarchy* is formed if information asymmetries between the entities are primarily due to time delays. Essentially, one decision simply has to be made at an earlier point in time (and usually based on less reliable information) while another decision can be made later, possibly based on more precise information. This temporal gap creates a rather implicit hierarchical relationship between the first (higher-level entity) and the second decision (lower-level entity). Note that decision time hierarchies not necessarily require the decision makers at both entities to be strictly different persons, although the term 'team' may be misleading in that context. A decision time hierarchy can simply result if one decision-maker has to make two decisions at different points in time.

Decision time hierarchies can be found in most production planning and SCP settings which use a hierarchical planning approach (see Section 2.1.2). For example, long-term decisions such as strategic network planning need to be made before mid-term master

[5] For simplicity, it will be assumed that there is a single entity at the top level and a single entity at the base level, unless noted otherwise.

[6] For a groundbreaking team decision model where the team members share the team head's preferences, see Groves (1973).

planning decisions. Scheduling decisions are usually made last as they have to be based on the most recent demand information possible. However, all information asymmetries will eventually be resolved in hierarchical planning situations, at least partially. Resolution of information asymmetries may occur either via the many feedback loops (see Section 2.1.3) or as part of the rolling horizon planning approach which triggers information updates at all levels of planning at regular intervals.

In contrast to decision time hierarchies, strict information asymmetry is typically present in *leadership hierarchies*. Leadership hierarchies are genuine multi-person situations with each entity being associated with a different decision-maker. These decision makers usually possess strictly different types of information. Such differences in information status will usually not be resolved over time, at least not completely. Leadership hierarchy settings are typical of confined organizations where formal hierarchies exist based on discretionary power (superior-subordinate relationships), e.g. as in most companies. Despite any differences in detail, most employees of a company generally support the overall objective of the organization (non-antagonistic situation), either voluntarily (team situation) or due to the authority relationship (enforced team).

Leadership hierarchies are often an appropriate setting to solve large-scale problems via decentralization, i.e. by employing a number of individual agents to perform specific sub-tasks, as done in supply chain planning. Van Zandt (2003) distinguished between two key forms of decentralization, *decentralized information processing* and *decentralized decision-making*. He studied both forms using a formal model for a recurring resource allocation problem. Demand for the resource constitutes local information and originates from the leaf nodes of the leadership hierarchy. In line with the leadership hierarchy model, all entities are assumed to work towards a common goal. In a first step, this distributed demand information has to be gathered and aggregated; in a second step, the allocations need to be calculated. Unfortunately, in the meantime, the local information may have changed. The problematic trade-off arises between gathering as much information as possible to make an all-embracing decision and between using the most current information, although the latter might only be available locally.

One solution approach is to employ a central planner to first gather a comprehensive set of demand information for all leaf nodes. He may then determine the proper allocations. By contrast, van Zandt showed the benefits of a decentralized approach: As demand information is only available locally, i.e. at the leaf nodes, it is beneficial to parallelize the demand collection at these leaf nodes and the aggregation process via the intermediate levels (decentralized information processing). Subsequently, the calculation of the disaggregated allocations for lower levels may be performed concurrently by multiple decision makers in the hierarchy (decentralized decision-making). Both parallelizations speed up the overall process, hence the total delay between information gathering and allocation is shorter and resources are allocated based on more recent information. In sum, decentralized decision-making allows allocating some resources among a small group of entities at lower levels in the hierarchy using more recent demand information. Higher tiers, by contrast, have to rely on older, but more comprehensive demand information.

Hierarchical Negotiations

The assumption made in the case of organizational hierarchies that the objectives of the higher and the lower-level entity are fully congruent is not always realistic. In some situations, these differences need to be accounted for if they have a significant impact on the overall outcome of the distributed decision-making process.

For example, the higher-level entity often has to cope with the fact that the lower-level entity has different preferences or pursues different objectives. Additionally, the lower-level entity may also enjoy an information advantage over the higher-level entity. This does not necessarily imply a conflict between both tiers. Often, the higher level has some minimum requirements or objectives, but once those are fulfilled, concerns of the lower level may be respected, too. According to Schneeweiß (2003), one way of coordinating the subordinate level under such *cooperative* circumstances is via *hierarchical negotiations*. This leadership activity essentially consists of a sequence of bargaining cycles where both levels repetitively make suggestions and provide mutual feedback. In particular, the lower level may explicitly utter dissenting views and may actively influence the overall solution via counter-proposals. However, in contrast to a negotiation on equal terms, the top-level may ultimately end the negotiation cycle by issuing an instruction if no mutual agreement can be reached.

A typical example in practice are budgeting processes in organizations (for a critical comment, see especially Jensen (2003)). More generally, almost all allocation problems in organizations involve negotiation cycles. An example has already been given in Section 1.1 where the quota allocation example in the oil industry case study was described. Since the lower-level entities provide essential input information and give feedback on allocation proposals in these negotiation cycles, they exercise an active influence on the final allocation decision.

However, there is a typical problem in practice once the higher-level entity needs to conduct multiple negotiations in parallel. In the absence of proper decision support systems, the allocation decision will not be made based on objective criteria (i.e. to give preferential allocations to the entity with the highest need or to the most profitable entity). Rather, the entity with the best (personal) relationship to the top-level will be preferred, or a simple "the squeaky wheel gets the grease"-rule is being followed in practice (Weisenborn and McCright, 1999).

Principal-Agent Situations

While organizational hierarchies and hierarchical negotiations are characterized by a certain level of congruence between the higher and lower entities, some situations with decentralized decisions may incur the risk of antagonistic behavior by either party. If the entities at the higher and at the lower level have divergent objectives, a number of problems may arise. Zimmer (2001, p. 25) summarized the following aspects:

- **Independent optimization decisions** often lead to sub-optimal overall results. A typical example is the problem of double marginalization in the channel coordination

problem (see Section 2.1.1), e.g. in wholesaler-retailer relationships. Here, independent price setting decisions both by the wholesaler and by the retailer to maximize their own individual profits not only jeopardize overall supply chain profits, but also lead to individually disadvantageous results.

- **Uncertainty**, e.g. of demand, prices or production output may lead to inefficient decisions at both levels. A fundamental question is which party will bear which share of the risk.

- Given asymmetric information between higher and lower levels, **agency problems** may arise if decision makers opportunistically exploit this discrepancy to their advantage (see also Harris et al., 1982).

The *principal-agent* setting allows studying such problems. More precisely, in principal-agent relationships, the top-level principal (referred to as a she) and the base-level agent (he) have different information endowments and generally follow different goals. Some information is unavailable to the principal and she cannot control the actions of the agent. Therefore, she strives to conclude a contract with the agent, incentivizing him to act at least partially in her interests. Usually, these contracts involve certain types of payments which remove the incentives for the agent to behave against the interests of the principal. An overview of such compensation schemes will be provided in Section 3.4.

In organizations such as corporations, principal-agent problems often cannot be neglected if the information advantage of the lower-level agent is material. The disciplinary power of the principal may actually prove insufficient to enforce full team behavior if the agent can gain a large advantage by pursuing his own objectives. Such situations are predominantly discussed in the literature in the context of budgeting problems.

In budgeting in hierarchical organizations, a given budget needs to be allocated in a top-down manner such that only the most profitable projects get funded. If the upper tier (principal) cannot adequately assess the profitability of certain projects, the lower-level decentral decision maker (agent) has an incentive to exploit his superior knowledge. He may report biased reports about the profitability or funding needs of certain projects. For example, by reporting a higher-than-actual investment need, a particular agent may receive a surplus allocation of funds termed 'slack' (see Antle and Eppen, 1985). Such managerial slack corresponds to an economic rent which is earned by the better informed agent. Slack either directly increases the income of the agent, or, more typically, can be spent on managerial perks.

Overall, a clear divergence of preferences arises in such budgeting situations: While the objective of the principal (i.e. corporate management) consists of finding an optimal investment program, the agent (i.e. the local manager) seeks to maximize his direct or indirect compensation. However, slack is not entirely bad. Empirical studies (for an overview, see Arya et al., 2000) have shown that slack also encourages innovation in companies and serves as an important motivational tool.

Similar diverging interests may also arise in the DMC problem. Assume each lower-level agent l privately observes an actual demand level d_l, but submits a demand report $\hat{d}_l$ to

the principal. The principal, lacking any ability to determine the true demands by herself, has to determine the allocation based on these reported values. There are clear incentives for the agents to bias their reports upward: If supply is scarce, a higher reported demand will lead to a higher allocation, e.g. under a proportional allocation scheme (more on this in Section 3.4). Lying increases the chance for the agents to obtain sufficient or even surplus allocations, permitting them to ensure a particularly high service level to their own customer accounts.

Before these aspects of principal-agent situations will be investigated in more detail in the following sections, a number of definitions need to be introduced to characterize the games which may be played between the top and the base level. Clearly, the outcome of such a principal-agent game depends on two aspects:

- The actual **allocation rule** employed by the principal as well as
- the **messages** submitted by the agents.

Focusing on the latter first, the case $\hat{d}_l = d_l$ will be referred to as a *truthful report*. Since the principal cannot verify the agents' reports immediately,[7] she has to motivate the agents to report truthfully.

For a given allocation rule, the allocation to an agent l usually depends on two factors, his own report $\hat{d}_l$ as well as the set of messages which have been submitted by all other agents except agent l. The symbol $\hat{d}_{-l}$ will be used to refer to these other messages. The function which determines the allocation to agent l can thus be written as $x_l(\hat{d}_l; \hat{d}_{-l})$. Such an allocation function is *individually responsive* if a higher report $\hat{d}'_l > \hat{d}_l$ leads to a strictly higher allocation to agent l unless the available supply $\bar{x}$ at the level of the principal is tight (Cachon and Lariviere, 1999c, p. 1095). Formally,

$$\hat{d}'_l > \hat{d}_l \Rightarrow x_l(\hat{d}'_l; \hat{d}_{-l}) > x_l(\hat{d}_l; \hat{d}_{-l}) \quad \text{if no shortage} \tag{3.15}$$

If the second $>$ sign is replaced by $\geq$, the allocation function is *weakly individually responsive*. Note that any individually responsive scheme provides incentives to the agents to misrepresent their messages so that personal objectives may be satisfied.

Standard game-theoretic solution concepts can be used to assess the overall outcome of a particular allocation function (for the following concepts, see the overview in Hall and Liu, 2008). Assume that the objective of each agent is maximizing his allocation to be able to provide a high service level to his customers.

If agent l has a strategy which allows him to maximize his allocation *independent* of the messages $\hat{d}_{-l}$ of all other agents, such a strategy is referred to as a *dominant strategy*. A weaker concept is that of a *Nash equilibrium*. An allocation game is said to be in a Nash equilibrium if no agent can gain a higher allocation by changing his report if everybody else sticks to his message. This means that given a message vector $\hat{d}_{-l}$, there is a message

[7] Yet often, an ex-post verification is possible and the situation more closely resembles a decision time hierarchy.

$\hat{d}_l$, not necessarily a truthful report, which must be submitted by agent l to maximize his allocation. In particular, it may turn out to be beneficial for agent l to lie when becoming aware of other agents lying. A Nash equilibrium does not have to be unique and with multiple equilibria, the outcome of a certain game may not be obvious. The existence of a Nash equilibrium is only an indication about the stability of a particular solution, but not necessarily implicates a certain desirable outcome. For example, while one Nash equilibrium may be achieved by a truthful-reporting strategy, a collusive solution may also be possible in situations where the agents coordinate their messages. An example of the latter case will be given in Section 3.4.3 when discussing the profit sharing compensation scheme.

Hardly any real-world hierarchical situation exhibits the characteristics of only one of these four prototypical hierarchies for distributed decision-making. Rather, mixed forms prevail. An illustrative example is the DMC problem where aspects of all four types of distributed decision-making in a hierarchy can be observed. Since only a brief characterization of multi-stage customer hierarchies has yet been provided in Chapter 1, a more comprehensive definition and characterization will follow in the next section.

3.3. Customer Hierarchies and Hierarchical Sales Organizations

The basic definitions of formal hierarchies (Section 3.1) and of the different aspects of distributed decision-making in hierarchical situations (Section 3.2) constitute essential preliminary work. Based on this, the following four sections allow for an introduction and characterization of customer hierarchies.

In Section 3.3.1, different approaches for customer segmentation will be discussed and a formal definition of customer profitability will be provided. This allows giving a formal definition of customer hierarchies in Section 3.3.2. As stressed before, there is a close relationship between a customer hierarchy in terms of the constituting customer segments and a hierarchical sales organization. A more detailed characterization of this relationship will be provided in Section 3.3.3. Finally, Section 3.3.4 will show that customer hierarchies exhibit characteristics of all four prototypical hierarchy types according to the framework by Schneeweiß (2003) which was introduced in Section 3.2.

3.3.1. Customer Segmentation and Customer Profitability

Traditionally, customer segmentation has aimed at identifying sub-markets with similar characteristics and thus similar needs. On the one hand, this has resulted in consumer markets often being clustered according to geographical, demographic, psychographic and behavioral characteristics. Business markets, on the other hand, were typically clustered according to industry sector, buying process characteristics, procurement structure or buyer-seller relationship (Helgesen, 2006).

Such traditional approaches to customer segmentation have focused on non-financial figures. Kalchschmidt et al. (2006, p. 621) have summarized three general dimensions which may be used for segmentation purposes independent of the market type:

- Inner characteristics: Buying patterns, loyalty, customer utility
- Sensitivity to exogenous environmental variables: Socio-cultural and macroeconomic aspects, weather conditions, etc.
- Decisional factors: Sensitivity to typical marketing-mix variables beyond pricing, e.g. reaction to promotion activities, service-levels, distribution policies.

While several of these aspects can be measured and used for comparisons, Kalchschmidt et al. (2006, p. 636) observed that there seems to be a lack of formal measures of heterogeneity to compare cases where heterogeneity occurs along several of these dimensions. As noted by Bock and Uncles (2002), the costs of building and utilizing effective market models which consider the aspects listed above will often be prohibitively high. Rather, for many practical applications, the criterion of *customer profitability* will be sufficient and can be used as a proxy for many of these non-financial criteria.

Above all, an orientation towards customer profitability ensures that a company strives towards what is best for its own long-term survival. Based on Mulhern (1999), the following definition of customer profitability will be employed in the course of this thesis:

Definition 10. *Customer profitability is defined as the "net dollar contribution made by individual customers to an organization" (Mulhern, 1999, p. 26).*

This definition calls for a proper measuring method to establish this net contribution. Shapiro et al. (1987) suggested analyzing the profitability dispersion of a customer base by examining both the associated revenues and the costs to serve a particular customer. This view corresponds to an accounting perspective in which customer profitability is the difference between revenues and costs over a certain time period. While revenues are relatively easy to track at a customer level, the allocation of costs to an individual customer (or even order) is often more difficult. Therefore, an explicit consideration of profitability aspects for segmentation purposes has only emerged once sufficient data on the purchase behavior of individual customers became available and could be utilized. A decisive factor was the introduction of customer relationship management (CRM) software in connection with the availability of point-of-sales data (see Blattberg and Deighton, 1991).

A comprehensive measurement of customer profitability requires the setup and management of dedicated customer cost accounts to track all direct and indirect costs which are incurred at several different levels (e.g. business unit, market, customer, order, see Helgesen (2006, p. 228)). However, many companies still tend to allocate central expenses such as marketing or sales costs to products instead of to customers (Howell and Soucy, 1990). In the context of profitability differences in the DMC problem, it typically suffices to focus on the key cost components where significant cost differences exist among the customer base.

Another critical aspect in accounting for customer-specific revenues and costs is the choice of the time-frame to consider. Customer profitability has both a historical and a future-oriented perspective. While historical accounts are typically easier to establish, forward-looking approaches are more appropriate for decision-making. The latter are also closely linked to the concepts of 'customer lifetime value' and 'customer equity' (for a broader overview, see McManus and Guilding, 2008). Profitability assessments based on historic customer transactions are myopic in only considering the direct dollar contribution of each customer. Such approaches disregard prospective customers who only have the potential to contribute (possibly significantly) to future profits. In a similar manner, some current customers may be undervalued if they are only evaluated based on their own purchasing behavior. However, if such customers act as opinion leaders and if their buying behavior influences a number of other customers, their indirect contribution to revenues and often to profits is much higher (Mulhern, 1999).

Overall, it has been observed in a number of industries that there is often significant *profitability dispersion* within the customer base of a firm. A limited number of customers usually accounts for a large share of overall profits. Cooper and Kaplan (1991) have reported that the 80-20 pareto rule-of-thumb known in sales (80% of sales are attributable to 20% of the customers) corresponds to a 225-20 rule when considering the profit dispersion in a customer base. While some customers in the example cited by Cooper and Kaplan had strictly positive profit contributions, most customers hardly generated any profits at all. Serving the remaining few customers actually implied heavy losses for the firm. The share of customers with strictly positive profit contributions was found to correspond to only 20%. The sum of the positive and negative profit contributions can be normalized, putting the aggregate profits of the entire customer base to a value of 100%. The cumulative profits of those 20% of customers with strictly positive profit contributions were found to amount to 225% of the aggregate profits of the entire customer base. This phenomenon of profitability dispersion can be depicted graphically with the so-called *Stobachoff curve* which will be introduced in Section 3.5.2 with the help of Figure 3.10.

The above discussion of customer profitability has neglected that customers are rarely managed at the individual level, but rather in the form of larger customer segments. If individual customers have already been assessed according to their profitability, larger customer segments can be defined by grouping customers with similar levels of profitability. Two basic approaches exist (Storbacka, 1997, p. 483): The first approach focuses on relative profitability. The first customer group may contain the 20% most profitable customers, the second group may contain the next 30% and so on. The second approach assigns individual customers to customer segments according to the customers' absolute level of profitability (e.g. those customers with an annual contribution between \$1,000–2,000, between \$2,000–5,000 etc.).

In both cases, the number of groups is a matter of judgment. Storbacka (1997, p. 484) has suggested refraining from using regular interval sizes and proposed choosing smaller segments once absolute segment profitability gets closer to zero. Indeed, determining the 'right' number of customer segments for a given customer base, or equivalently, setting

the size of a particular customer segment are difficult tasks in practice. Serving a high number of customers via shared resources (e.g. a single sales agent) involves less direct expenses for these shared resources. Yet, indirect costs will be higher since each of the resulting large customer segments will still exhibit a high level of customer heterogeneity which will not be exploited. However, the smaller the segment, the sparser is usually the amount of available planning data (Jiang and Tuzhilin, 2006). In the most extreme case, a 1-on-1 marketing approach may be only desirable for large, very important customers (see the comprehensive literature on key account management, e.g. McDonald (1997)).

In many practical situations, customer segments cannot be defined purely from a profitability point of view. Rather, customer segments in practice usually arise as by-products of other management decisions and thus often also follow pragmatic considerations. Typical examples are customer segments which correspond to the geographical location of the individual customers. Consider industries which are characterized by MTS production, with fixed sales prices and with constant production costs. In this setting, there will still be significant differences in the cost of order fulfillment simply due to the geographical segmentation of the customer base. If additional local taxes, different transportation costs or exchange rate effects are born by the company, as it is often the case in practice, significant differences in customer profitability may result.[8]

A number of other popular dimensions have been discussed in Section 2.2 in the context of demand planning and hierarchical forecasting. As discussed extensively in Chapter 1, the resulting partitioning of the customer base is not flat in practice. Instead, it exhibits a hierarchical structure. The following section finally provides a formal definition of such customer hierarchies.

3.3.2. A Formal Model of Customer Hierarchies

First, a common terminology as well as a few simplifying assumptions need to be introduced for a formal model to describe customer hierarchies. Finding the adequate level of detail in modeling is always a challenge. The dilemma of capturing the essence of the problem while leaving out unnecessary aspects is often referred to as the 'art of modeling' (see Williams, 2000; Voß and Woodruff, 2006).

In this thesis, a particularly simple representation of a customer hierarchy is helpful. An important building block of such a hierarchy is the individual *customer segment*. Each customer segment i is characterized by a particular size, which corresponds to the demand volume d_i and by a particular customer profitability value as measured by the average unit profitability p_i.

Customer segments exist at different levels of aggregation. The most disaggregate type of customer segment will be referred to as a *base customer segment*. For simplicity, it will be assumed that there is no variation regarding the customer profitability within any particular base customer segment l; i.e. p_l is assumed to be constant and identical

[8] Additional aspects such as multiple transport modes in settings with geographically spread multi-site networks are considered in the overview of demand fulfillment in network structures in Nguyen et al. (2012).

for each of the d_l individual demand units. This assumption is not restricting since any (aggregate) customer segment with heterogeneous profitabilities can easily be subdivided into finer sub-segments until the within-segment heterogeneity approaches zero.

The most aggregate customer segment comprises the entire *customer base* with an overall demand d_0 and average profitability p_0. Between these two extremes, there are segments at intermediate levels of aggregation. These different types of customer segments constitute the building blocks of a customer hierarchy.[9] Focusing on the hierarchical relationship between the individual customer segments at all levels, a customer hierarchy will be defined as follows for the scope of this thesis:

Definition 11. *A customer hierarchy is a particular multi-level segmentation of a given customer base with three key properties:*

- *All leaf nodes correspond to the set of base customer segments which are homogeneous with respect to profitability.*
- *All leaf nodes with the same immediate parent node k strictly differ in terms of profitability, i.e. $p_l \neq p_{l'}$ for all $l \neq l'$ and $l, l' \in \mathcal{D}_k$.*
- *Each level of aggregation describes a particular partitioning of the entire customer base.*

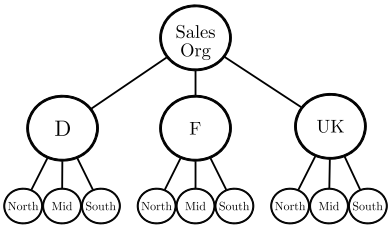
In addition to the above properties, the characteristics of basic hierarchies as defined in Section 3.1 continue to hold, implying that each customer hierarchy has a strictly convergent structure since there is a unique parent node to each successor node. In particular, there is a unique path connecting each leaf node (base customer segment) with the root node (entire customer base).

For a given customer base, different customer hierarchies may exist. They differ with respect to the aggregation structure, i.e. the choice of aggregation criteria used at each stage. Figure 3.3 depicts examples of three different customer hierarchies which can be defined for the same customer base (and the same base customer segments) by using different combinations of geography and type of sales channel as aggregation criteria at the different hierarchical levels. In Figure 3.3a, only the geography criterion is used. The first level consists of a split per country and the second, i.e. lowest level distinguishes within each country between a northern, middle and southern region. In Figure 3.3b, also a split into national sales organizations is used. However, the lowest level consists of a split along the three distribution channels retail, online and wholesale. Lastly, in Figure 3.3c, the superior criterion is the distribution channel. Each channel is managed across all countries.

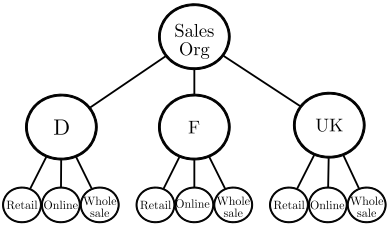
Choosing a suitable customer hierarchy structure is often a difficult task. As outlined in Section 2.2.5, the approach based on demand planning paths may provide a starting

[9] The convention from Section 3.1.1 regarding the use of indices to describe the nodes in a hierarchy will also be kept for the case of segments in customer hierarchies: The index i is used for general segments, index k is reserved for aggregate segments (i.e. intermediate nodes) and index l denotes base customer segments or leaf nodes.

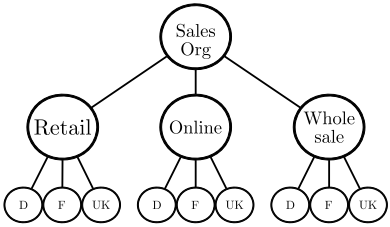
point, but it is crucial to account for the constraints and interdependencies imposed by the design of the overall planning system.



(a) Level 1: country; level 2: region



(b) Level 1: country; level 2: sales channel



(c) Level 1: sales channel; level 2: country

Figure 3.3. – Customer hierarchy variants due to different aggregation structures

As stated before, a geography-based customer hierarchy will be assumed to simplify the presentation in this thesis. This does not constitute a limitation and the subsequently presented results will also apply to other types of customer hierarchies with different aggregation structures.

In the model of customer hierarchies employed here, the values of the demands and unit profitabilities at the more aggregate customer segments can be obtained by simple recursive calculations, starting at the lowest level $M-1$: As it is a summable quantity, aggregate demand at node i at level $M-2$ corresponds to the sum of the demands of the leaf nodes in $\mathcal{D}_i$. Similarly, aggregate demand of a node k at level l with $l < M-2$ refers to the sum of the demands at all direct successor nodes in $\mathcal{D}_k$. More generally, the demand at each intermediate node is the derived demand of the immediately following nodes at the lower level. It is calculated by

$$d_k = \sum_{i \in \mathcal{D}_k} d_i \quad \forall k \notin \mathcal{L}. \tag{3.16}$$

Profitability, however, is a non-summable quantity. Several possible aggregation schemes are possible (see the discussion in Section 3.1.2). The most intuitive choice is to use the demand-weighted arithmetic average to determine the aggregate unit profit at intermediate node k. This value can also be calculated in a recursive manner via the demands and

unit profits at the next lower level, i.e.

$$p_k = \frac{\sum\limits_{i \in \mathcal{D}_k} p_i \cdot d_i}{d_k} \quad \forall k \notin \mathcal{L}. \tag{3.17}$$

As will become clear in Section 3.5.3, this demand-weighted aggregation of unit profits is also required for the measurement of heterogeneity in customer hierarchies via the Theil index.

A graphical representation of a customer hierarchy, also depicting the key notation, is given in Figure 3.4. The hierarchy consists of $|\mathcal{N}|$ nodes. Nodes at the upper levels $1, \cdots, M-2$ are intermediate nodes which possess a unique parent node and have links to an integer number of successor nodes. They correspond to coarser customer segments where unit profits are no longer uniform. For simplicity, in most examples in this thesis, balanced and uniform hierarchies will be used, i.e. the direct path from the root node to each leaf node passes through the same number of intermediate nodes. As a consequence, all leaf nodes in such hierarchies are positioned at the lowest level $M-1$.

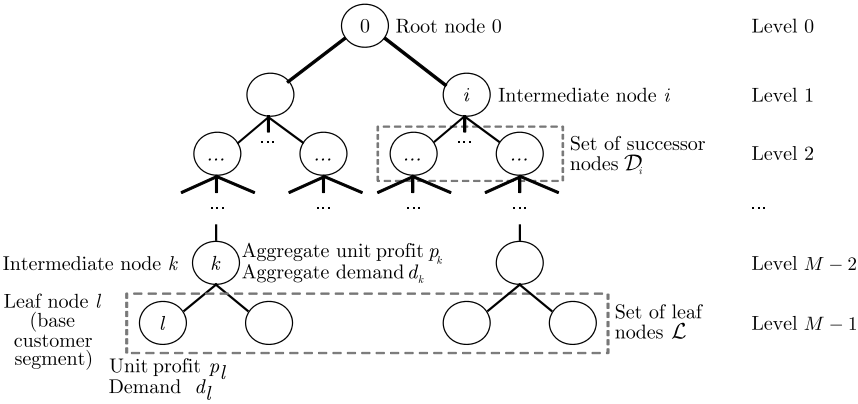


Figure 3.4. – Naming conventions for customer hierarchies

Nevertheless, unbalanced and ragged hierarchies often exist in practice. Imagine a four-level geography-based customer hierarchy in which the aggregation structure of the two intermediate levels (between root node and leaf nodes) corresponds to the country and (federal) state or province level. While this setup applies well to countries like France or Germany, the intermediate level of (federal) states may not exist in some countries like Finland. This leads to a ragged hierarchy, illustrated in Figure 3.5, since the immediately superior level above the base customer segments has to be the national (country) level. Many global companies have policies which delegate authority for certain decisions to a particular hierarchy level. For example, national sales plans may be prepared at the country level whereas the assignment of sales territories to agents may be performed by

regional managers operating at the province level. One option is to shift the responsibility usually associated with the province level to the next higher level. This effectively turns the ragged hierarchy into one without gaps or jumps. As long as the higher administrative layer is not burdened with rather ancillary problems, such a practice is unproblematic and commonplace in sales organizations in smaller countries such as Finland where a less granular structure is sufficient.

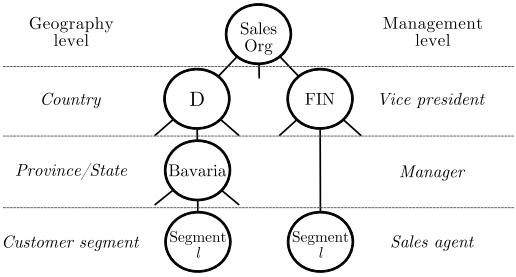


Figure 3.5. – Ragged hierarchy example: Geography and management levels

The practical problems associated with ragged hierarchies are more of a psychological nature. For example, in organizational charts, the country-level vice presidents of sales in Finland will still be placed at the same hierarchical level as their peers in other countries, although other countries such as Germany have an additional layer of management at the province level. The Finnish sales agents will enjoy a direct line of reporting to a country-level vice president, but are aligned with their peers in other countries at the lower hierarchy level (see right-hand side of Figure 3.5).

3.3.3. Hierarchical Sales Organizations

The above discussion of ragged hierarchies has already indicated that there is a close relationship between a customer hierarchy and the design of the sales organization of a company. In fact, one structure usually matches the other as there is often a direct relationship between a certain sales manager and a customer segment (see also Section 2.2.3). A hierarchical sales organization (refer again to Figure 2.10) is particularly prevalent in business markets where the sales force composite method is used for demand planning, forecasting and sales management.

The model of a customer hierarchy which is employed in this thesis builds on the assumption that each base customer segment is managed by a particular sales agent. Accordingly, there is also a sales manager who bears responsibility for a particular aggregate customer segment at the next higher level. At the same time, this sales manager is the direct superior of the lower-level sales agents. Similar relationships exist at higher hierarchical levels between higher-level and lower-level sales managers.

The sales agents have two primary tasks: On the one hand, from an external perspective, they conduct sales activities by informing potential customers, closing sales contracts and

performing after-sales service. On the other hand, from an internal perspective, they gather market knowledge and prepare forecasts of future demand within their customer segment. The tasks of the sales managers are more biased towards internal activities. They act as superiors for the sales agents (or lower-level sales managers), they supervise and monitor both the sales activities and market reports of the agents, and they are also involved in sales and demand planning. In particular, two of their major tasks are the *aggregation of lower-level forecasts* and the *disaggregation of sales quotas* received from higher hierarchical levels. This broad spectrum of activities can explain the need for a hierarchical structure in practice.

Starting with the seminal works of Williamson (1967), the existence of multi-level hierarchical structures in organizations has primarily been rationalized with bounded information processing capabilities (e.g. McAfee and McMillan, 1995).[10] Put differently, a supervisor or manager is only capable of coordinating a limited number of subordinates in an effective manner. An implication of a limited span of control is the existence of multiple levels of management.

Assume a company requires 25 sales agents to sell a product in a larger geographical area. Their sales activities need to be aligned among each other (e.g. to ensure that the individual sales territories are mutually exclusive and collectively exhaustive) and with the production capabilities of the organization (to match supply with overall demand). For these coordination activities, the firm would like to employ a coordinator as a sales manager. However, assume that a single coordinator can only efficiently coordinate five individual agents (span of control of five) due to bounded information processing capabilities. This means that in a first step five sales managers are required in addition to the 25 sales agents. However, this is not yet sufficient; the interface to production still needs to be coordinated and the problem of coordinating the 25 sales agents has been replaced by the problem of coordinating five sales managers. Hence, an additional, second layer of management is required. It consists of an additional, central sales manager who coordinates the five lower-level managers and also manages the interface to production. As a result, a hierarchical, arborescent sales organization with three levels and 31 individuals has been formed.

Besides the limited cognitive abilities of humans, a number of other reasons let many practitioners favor a hierarchical structure over a centrally administered organization. Summarizing prior publications, Gijsbrechts (1985) compiled an overview of typical motives. In particular, he stated that hierarchies

- allow for easier managerial interactions,
- facilitate the recognition of success and failure of top managers,
- induce a more judicious allocation of responsibility and
- require fewer information transmissions.

[10] The motive of a limited cognitive ability goes back to Hayek (1944).

As will be argued in the next chapter, this last aspect is still highly relevant in practice despite major advances in information technology deployments. In addition to the aspects listed above, Biddle and Steinberg (1984, p. 5) noted that a decentral, hierarchical structure often improves the motivation of managers, allows for faster reaction times and provides training opportunities for managers.

These clear advantages of hierarchical organizational structures need to be contrasted by a number of typical disadvantages. The most notable ones are agency problems due to asymmetric or decentral information and opportunistically behaving agents. Like all planning hierarchies, also hierarchical sales organizations have to find a balance between improving the forecast quality and reducing administrative expenses (see Osband, 1989, p. 1108).[11] However, the prevalence of large multi-tier sales hierarchies in practice suggests that the advantages of hierarchies may often justify their costs.

The most obvious advantage is that hierarchies are an effective way to decompose a complex task. The following section will highlight the different aspects of distributed decision-making which can be attributed to customer hierarchies, using the framework which has been introduced in Section 3.2.

3.3.4. Customer Hierarchies in the Context of the Schneeweiß Framework

As discussed extensively, the key difference between a flat partitioning of customer segments and a multi-stage customer hierarchy is that the latter case usually requires a decentral planning approach. The peculiarities of the distributed decision-making process in customer hierarchies can aptly be characterized with the help of the framework of Schneeweiß (2003) which was introduced in Section 3.2. As will be shown in the following paragraphs, multi-stage customer hierarchies exhibit the features of all four fundamental hierarchy types.

Constructional Hierarchies: Most hierarchical structures like customer hierarchies have been established to simplify a number of rather complex tasks. In practice, these tasks typically cannot be performed by a single, central planner alone. In that sense, customer hierarchies exhibit aspects of *constructional hierarchies*. The hierarchical structure—both in form of the customer segments and in form of the sales organization—constitutes a means to reduce complexity as no single planner may simultaneously serve all customer segments, perform forecasting and allocate sales quotas and product quantities. Cognitive

[11] When moving from a simple two-level hierarchy to a larger multi-tier hierarchy, these adverse effects may be exacerbated due to similar game playing at intermediate levels. Models investigating multi-level principal-agent settings can be found in Demski and Sappington (1987), Melumad and Mookherjee (1995), McAfee and McMillan (1995) or Mookherjee and Reichelstein (1997). The exact costs of the hierarchy depend largely on the extent of information asymmetry (more precisely, on the number of levels with private information), on decision rights (i.e. who designs the contracts) or on budget constraints. Furthermore, transmitting information (e.g. forecasts) and instructions (e.g. allocation decisions) over a longer cascade of individual agents may result in (random) errors creeping in since information is reproduced serially several times (Williamson, 1967, p. 126).

limitations of human planners necessitate a decentralized approach with a limited span of control, with distributed decision-making and with distributed processing of information.

Organizational Hierarchies: From a macroscopic perspective of the overall firm, it is reasonable to assume that all decentral decision makers pursue the same objective, i.e. of satisfying the needs of the customers in a way which maximizes profits to the firm. In a sales organization, this is often equivalent to maximizing the sales volume in a joint effort by the individual sales agents, supervisors and managers.[12]

In pursuing this objective, customer hierarchies exhibit the properties of both types of *organizational hierarchies*—decision time hierarchies as well as leadership hierarchies.

- On the one hand, characteristics of *decision time hierarchies* can be found as some decisions are made earlier than others. Typical strategic and tactical managerial tasks such as sales force design, demand planning and quota assignment are performed by higher hierarchical levels. These decisions are typically based on an early information state (i.e. mid- and long-term forecasts). At an operational level, the sales agents make demand fulfillment decisions based on very current, i.e. actual demand information. Eventually, the realized demand will become common knowledge to all hierarchy levels and this type of information asymmetry is thus resolved over time.

- On the other hand, also characteristics of *leadership hierarchies* are present. Each decentral sales agent, interacting closely with the customers in his segment, possesses superior knowledge regarding his own customer segment both with respect to his peer sales agents and compared to the higher-level sales managers. Primarily, this private information comprises quantitative information such as the actual demand volume in the market (especially if lost orders are not recorded centrally). Furthermore, many qualitative properties of the individual customers may neither be obtained nor verified by higher-level managers in the customer hierarchy. It is generally only the local sales agents who have precise information regarding the customer preferences. This type of information asymmetry is typically not resolved over time.

Hierarchical Negotiations and Principal-Agent Situations: As sales managers need to coordinate their subordinates, often *hierarchical negotiations* take place in customer hierarchies. An example is the iterative quota-setting process to establish or re-negotiate sales targets. Sales agents and lower-level sales managers provide valuable input which helps directing scarce product quantities to their most profitable use. However, given the information differential in the hierarchy regarding the true demands of the individual customers, this negotiation process is prone to manipulation by the lower-level agents. Hence, the assumption of an actual or enforced team setting may not always hold.

[12] Additional ancillary sales force objectives may include contributing to particularly profitable sales and to enable continuous growth.

If sales agents or lower-level sales managers have an incentive to exploit their superior information status to pursue own objectives, the customer hierarchy will exhibit *principal-agent* properties. As a result, a certain amount of antagonistic behavior may exist in customer hierarchies given the control issues and the multitude of incentives for the individual agent. Nevertheless, such agency and game-playing behavior, especially in the context of forecasting, can be mitigated or even entirely suppressed with the help of properly set incentives and other means such as monitoring. This problem area of customer hierarchies will be addressed in the following section.

3.4. Forecast Misrepresentation in Customer Hierarchies

In many customer hierarchies, sales agents and sales managers jointly work towards fulfilling the same objectives. However, it has already been pointed out that significant information asymmetries may result from using the salesforce composite forecasting method. Occasionally, these asymmetries can give rise to agency behavior, i.e. they may lead to a principal-agent situation and resulting forecast misrepresentations.

With respect to forecasting in such hierarchical settings, there are two categories of game-playing behavior which can compromise the overall sales force composite forecast (see McCarthy Byrne et al., 2011, p. 130):

- On the one hand, if decentral forecast reports are used to set quotas or sales targets, agents will **underestimate** future demand to secure a target which is easy to reach. This problem is likely to be more prevalent on a mid-term horizon, e.g. as part of the internal budgeting process.

- On the other hand, if the sales organization is facing potential product shortages and agents are put on allocation, they will **exaggerate** their demand forecast reports to secure higher allocations. This problem will typically be more prevalent on a more short-term horizon, e.g. as part of demand fulfillment activities.

Taking up a differentiation introduced by Kilger and Meyr (2008), it will be argued in the following that these two categories of game playing correspond to a supply chain which is either *constrained with respect to demand* or which is *constrained with respect to supply*. In the former case, the supply chain is able to meet all customer demands (supply $>$ demand) whereas in the latter case supply needs to be rationed among the customers (supply $<$ demand). In a general supply chain setting, the nature of the supply depends on the production environment and on the location of the CODP (see Section 2.1.4). Put differently, supply may either correspond to production capacity (MTO environment), both assembly capacity and material/component inventory (ATO environment) or final product inventory (MTS environment). In the course of the following discussion, however, the focus will be placed on MTS environments.

The two types of game-playing introduced above may exist at all hierarchical levels, not just between the lower-level sales agents (leaf nodes) and their immediate superior

sales managers. Sales managers at higher hierarchical levels may have similar incentives to bias their aggregated forecast reports. In sum, the cumulative distortion may even be magnified due to the hierarchical structure, possibly leading to distinctly false demand and profitability forecasts[13] at the root node.

Such forecast misrepresentation problems and possible solution approaches will be addressed in the following sections. They have been organized as follows:

- First, Section 3.4.1 will use the principal-agent framework to characterize the relationship between a sales agent and his superior sales manager. In particular, agency problems will be introduced which result from information asymmetries between both parties.
- Section 3.4.2 contains an overview of compensation schemes which have primarily been discussed in the salesforce management literature. The setting in this stream of literature fits a demand-constrained supply chain and a mid-term planning horizon.
- Section 3.4.3 then focuses on supply-constrained supply chains where an allocation problem must be solved in the short-term. Corresponding compensation schemes have been derived in the microeconomics and game theory literature.
- Stylized practical and experimental experiences with the various compensation-schemes for both types of supply chain environments will be presented in Section 3.4.4.
- Ultimately, Section 3.4.5 concludes this review. The relevance and practical applicability of the inspected compensation schemes will be discussed in the context of the DMC problem.

3.4.1. Sales Forecasting and Agency Problems

As outlined in Section 2.2, obtaining reliable demand forecasts is crucial for almost all planning processes in a firm. Many companies rely on their own sales force, primarily for forecasts covering a medium time horizon, e.g. the next quarter or next year (White, 1984, p. 37). Most forecasting of short- and medium-term demand is usually done internally[14] and typically based—to different degrees—on inputs from the firm's sales force.

The use of these sales force composite forecasts is particularly popular for forecasting in business and industrial markets: As the buying behavior in such markets differs significantly from consumer markets (see Section 2.2.3), the importance of each single customer is much greater in a business-to-business (B2B) environment. The individual customer is no longer anonymous, and there is usually a close relationship with the sales agent.

[13] Similar to the demand figures, there is the risk that profitability forecasts can become biased at any hierarchy level. Even if all leaf-node profitability figures are unbiased, aggregate profitability figures are still adversely affected by the aggregation process as the biased demand forecasts enter the formula for the demand-weighted arithmetic average.

[14] The use of outside services is more common for long term trend forecasts, see Davidson and Prusak (1987) or when new products are being introduced, see White (1984).

Besides selling, salespeople typically spend a significant amount of their time and effort on conducting market research. This intimate exposure to market insights gives salespeople information advantages and makes it reasonable for a firm to exploit their knowledge for forecasting. This is a clear contrast to a central planner who is significantly more distant to the individual markets and has fewer means to make accurate forecasts.

These sales force composite forecasts have a tremendous economic importance. They affect a wide number of decisions in a company either directly or indirectly, ranging from production and inventory planning, new product development to the setting of the compensation levels of the salesforce (Chen, 2005, p. 60). Moreover, sales force compensation is a significant expenditure for most firms as observed by Zoltners et al. (2008): With more than 3.6 mn salespeople only in B2B selling, 4.3 mn sales agents in retail sales and more than 15 mn employed in direct-to-customer sales, Zoltners et al. conservatively estimated that the total sales force spending in the US amounts to $800 bn annually. Put differently, this expenditure is equivalent to 10% of annual sales for an average American company, but can in some instances be as high as 40%. This importance of salespeople is reflected in the rich literature on sales force management. For a comprehensive research framework and a crisp overview of key problem areas, see e.g. Mantrala et al. (2010).

Despite the obvious virtue of relying on its own sales force for demand forecasts, a firm needs to handle the associated **hazards** as sales force composite forecasts may result in erroneous forecasts.[15] In particular, firms incur significant costs if actual demand turns out to be different from the forecast: Once sales fall short of the forecast, costly inventories remain and overhead costs of production may not be covered. If actual demand was underestimated, production resources may be strained due to rush orders or even stockouts may result (see Mantrala and Raman, 1990). Obviously, forecasting proficiency is a key driver of accuracy. In many situations, additional intentional and unintentional bias may be present. Forecasts may be misrepresented due to diverging interests in the customer hierarchy, leading to over- and under-budgeting.

The focus in this section lies primarily on the (hierarchical) relationship between an individual sales agent and his immediate superior, the sales manager. As discussed above, both have dual roles. The role of a sales agent, who both conducts selling activities and prepares forecasts, has been summarized aptly by Ijiri et al. (1968) as that of a "forecast-operator". The manager, in turn, while supervising and monitoring the agent, not only aggregates the agent's forecast report, but also allocates sales quotas and product quantities based on these reports. Similar relationships exist between lower-level and higher-level sales managers. The main difference is that the lower-level sales manager is not directly involved in sales activities. Nevertheless, she performs forecasting duties in the sense that she aggregates forecast reports of her subordinates. This step is similarly susceptible to agency behavior. Therefore, the argumentation and the results presented in the following also apply to principal-agent settings at higher hierarchical levels.

[15] A short overview of typical problems associated with judgmental forecasts has already been given in Section 2.2.3.

To simplify the following analysis, the objectives of the sales manager (principal) are assumed to be in line with those of corporate management, but not necessarily have to coincide with those of the agent. There are three different types of information asymmetries which may occur in customer hierarchies (see Varian (1992, Ch. 25) and Schneeweiß (2003, pp. 126–129)):

- The obvious type of asymmetric information is referred to as *hidden information*. In a customer hierarchy, there is usually a large amount of market-related knowledge which is not known to the principal, but only to the agent, e.g. regarding the market size (demand), profitability or regarding the form and shape of the sales response function. In many principal-agent models, it is assumed that at the time a contract is concluded between the principal and the agent, both have the same information status, but that the agent receives additional market information until the time he has to make a decision or needs to commit to an activity.

- A second type of information asymmetry relates to specific characteristics of the sales agent, e.g. his capabilities or preference for leisure. To some extent, these *hidden characteristics* are revealed over time. Therefore, the problem of making an *adverse selection* primarily emerges if the principal has to select a sales agent from a larger pool, e.g. when hiring. To overcome information asymmetry caused by hidden characteristics, usually three methods (with increasing levels of sophistication) have been discussed in the literature (e.g. see Varian (1992) or Fudenberg and Tirole (1991)). First, the principal may rely on a simple *screening* of the candidates, e.g. by interviewing and checking formal qualifications. An alternative is *signaling* by which the agent actively discloses some private information in an attempt to be selected. Third, *self-selection* corresponds to a screening process by which the principal offers a menu of different contracts. By choosing a particular contract, the agent inadvertently reveals some of his properties to the principal.

- The third type of asymmetric information is referred to as *hidden action*: These can arise in situations in which the principal cannot monitor the actions of the agent. Instead, she can only observe the results of the activities of the agent. Exercising effort is usually costly to the agent and thus avoided where possible. The principal cannot conclusively infer whether the agent was negligent or was withholding efforts intentionally (Osband, 1989, p. 1091). In stochastic environments, she is also unable to distinguish between deceiving actions of the agent and unfavorable environmental conditions. Hidden actions may thus lead to the problem of *moral hazard* if the agent's objective differs from that of the principal.

The problem of hidden characteristics is usually relevant for the hiring decision, but of a lower importance once the individual sales agents have already been selected. Given that the focus of this thesis lies on operational demand fulfillment decisions in existing customer hierarchies rather than on design issues, it is reasonable to assume that sales agents have already been selected. Hence, the problem of hidden characteristics will be

ignored in the following. The more important problems relate to hidden information and to hidden actions. Both types of asymmetric information can constitute the root causes of the two types of game-playing introduced earlier, underreporting to obtain a lower sales target and exaggerating to ensure a higher allocation. The purpose of the following presentation is to discuss possibilities for the principal (and ultimately for the company) to incentivize the agent to behave in a non-antagonistic manner. This will allow the principal to better match supply with demand and to make more informed production and allocation decisions.

Depending on the state of the supply chain, the two key types of asymmetric information existing in customer hierarchies—hidden information and hidden action—will usually differ in their impact. For the following analyses, it is helpful to distinguish between two fundamentally different supply chain states. On the one hand, the supply chain can be in a *demand-constrained mode*; on the other hand, a *supply-constrained mode* can be encountered (see Kilger and Meyr (2008, pp. 188–189)).

- In the first case, the supply chain is generally able to generate ample supply to match any level of demand. From a sales perspective, this implies that sales activities are push-based and require efforts, for example calls and visits to customers or even promotional activities. In practice, the principal has only limited abilities to monitor the extent to which the agent exercises effort; hence the problem of hidden actions arises. In a demand-constrained supply chain, truthfully disclosing market information and exercising effort are hardly compatible goals from the point of view of the sales agent (Chen, 2005, p. 60). Both hidden information and hidden actions may give rise to agency problems in demand-constrained supply chains.

  Such a situation is more likely on a mid-term planning horizon where supply volumes can still be adjusted.

- Things are different in the supply-constrained case: Selling is comparably easy since customers 'pull' the limited amounts of supply. The principal still cannot monitor the level of effort which is exercised by the agent. However, it is reasonable to assume that hidden actions are less of a problem because the sales agent will have less difficulty in closing a sale and will also spend less time searching for customers. Rather, his role will center more on market monitoring (for internal purposes such as forecasting) and on providing customer services. Therefore, the problem of hidden information is likely to be of a far greater importance. If the agent either has an (intrinsic) motivation to serve his customer segment particularly well or if he is paid according to actual sales, he has an incentive to overstate his demand forecast.

  This situation can often be encountered in short-term demand fulfillment where supply quantities are no longer adjustable. If actual demand has been underestimated at a mid-term planning level, the available supply quantities may be insufficient to serve all customers.

In short, the key point here is that the demand-constrained supply chain is more likely to lead to the underreporting game play (both due to hidden actions and hidden information) whereas the supply-constrained supply chain appears to be more susceptible to exaggeration (primarily due to hidden information). A particular supply chain may experience both types of game play at different stages in the planning cycle, as it may be constrained with respect to demand at a mid-term level while supply may turn out to be tight on a short-term horizon.

In the following, a number of popular compensation schemes will be presented. It will be analyzed to what extent they can mitigate or even solve the incentive problems in customer hierarchies. Section 3.4.2 starts with schemes for the demand-constrained supply chain. The situation with a supply-constrained supply chain will be addressed in Section 3.4.3.

3.4.2. Compensation Schemes for Demand-Constrained Supply Chains

In a deterministic demand-constrained supply chain environment, the problem of the principal consists of motivating the agent to conclude as many sales as possible in a given amount of time. The time horizon of this problem is more likely to be on a mid-term level. In the absence of the forecasting problem, she will pay the agent directly according to the observed sales in a certain period, assuming that there is a direct relationship between the activities of the agent in period t and the total output y_t.

In general, this relationship is not perfect. Rather, it is reasonable to assume that overall output is related at least to two components:

- The level of effort e_t exercised by the agent and
- a random component ϕ_t which reflects certain characteristics of the market environment in period t.

The level of effort constitutes private information of the agent and cannot be observed by the principal (hidden action). Furthermore, the agent will usually be able to make a more accurate forecast regarding the market characteristics ϕ_t than the principal (hidden information).

In such a basic market model, total sales in a particular period t amount to $y_t = e_t + \phi_t$. A principal-agent situation arises because the principal cannot distinguish between a bad market environment (ϕ_t) and a deceptive agent (e_t). Since production is subject to a lead time, the principal has to elicit an accurate demand forecast significantly ahead of the selling season. Additionally, she must incentivize him to exercise effort during the actual sales period. As will be seen shortly, this crucial upfront forecasting step is rarely addressed by the existing schemes.

The marketing and salesforce management literature contains a wealth of models which illustrate ways to solve or mitigate the above principal agent problem. Most contributions follow a standard setup (see Coughlan, 1993): The principal is assumed to have the power

to suggest a particular contract. First, she determines the form of the compensation plan and sets the level of the incentives while anticipating certain behavioral characteristics of the agent. The latter accepts the contract (otherwise nothing happens) and chooses his actions (e.g. level of effort) based on the incentives present and prevailing environmental conditions. Finally, the agent receives his compensation according to the pre-announced plan, based on the sales volume which the principal can observe.

There are two fundamental approaches to analyze this basic principal-agent setting; examples of both will be given in the following paragraphs: On the one hand, one may assume a particular contract type and its parameters as exogenously given. From an ex-post perspective, an analysis can then be conducted to identify the incentives which have been created by this particular compensation scheme. On the other hand, given the sales environment, the behavioral characteristics of the agent as well as the preferences of the principal, one may use the agency framework to endogenously derive a suitable compensation scheme. As a result of such tailored schemes, the principal effectively 'bribes' the agent to induce a certain favorable behavior (e.g. to exercise effort).

It will be shown in the following that these schemes primarily address the moral hazard problem and can incentivize the agent to work hard. First, standard quota-based schemes with bonus payments and commissions will be analyzed. Then, a brief overview of endogenously derived compensation schemes will be given. Unfortunately, neither of these models from the salesforce management literature provides adequate incentives for truthful forecasting. The only reward system for the demand-constrained supply chain which encourages exercising effort and elicits truthful forecasts is the so-called Gonik scheme. It will be discussed separately in the last paragraph of this section.

Quota-Based Schemes with Bonus Payments and Commissions

A *sales quota* is the basis of most basic compensation schemes. It represents a specific volume or revenue objective which an individual sales agent must achieve over a period of time, typically one year (Good and Stone, 1991, p. 51). Two basic approaches are distinguished in practice and in the literature, the *top-down* and the *participative quota setting* process. In a top-down approach, the individual agent has no influence on his target. Targets for sales areas are broken down from aggregate production and marketing plans, with area-specific forecasts only being used for minor reconciliations (Fildes and Hastings, 1994, p. 2). When using a top-down quota setting process, the principal does not need to elicit a truthful demand forecast from the agent. Rather, her problem is reduced to incentivizing the agent to exercise effort to fulfill the quota.

Top-Down Quota Setting: How does the principal set the quota? Assume she knows nothing about the true potential of a particular market and picks a rather low target q_t for period t. To incentivize the agent to work hard, she uses a *commission*-based compensation. Denote the entire compensation by B_t which the principal will pay to the agent subsequent to period t. One component of the compensation is the *base salary* which is paid independent of the actual sales volume y_t. The other component is a variable,

volume-dependent part. In the absence of other income alternatives for the agent, the base salary can be normalized to zero without loss of generality to simplify the analysis.[16] The remaining variable part is a simple linear function of the difference between actual sales and quota once the quota has been surpassed. Denoting the commission rate by b, this gives the following total compensation:

$$B_t = \begin{cases} 0 & \text{if } y_t < q_t, \\ b \cdot (y_t - q_t) & \text{if } y_t \geq q_t. \end{cases} \tag{3.18}$$

After actual sales y_t are known, the principal usually avoids picking a new target for the agent which lies *below* the just *achieved* sales in an attempt to keep 'raising the bar'. Hence, she will choose some $q_{t+1} \geq y_t$ for the next-period target. Weitzman (1980) has analyzed an updating rule which is frequently used in practice. This updating rule consists of a constant increment δ_t as well as a proportional adjustment λ_t to account for any over- or under-fulfillment observed in the current period t:

$$q_{t+1} = q_t + \delta_t + \lambda_t \cdot (y_t - q_t), \quad \delta_t, \lambda_t > 0. \tag{3.19}$$

The sequence of resulting quotas according to (3.19) is characterized by what is commonly termed the 'ratchet effect'.[17] In this dynamic setting, the agent needs to find a trade-off between exercising more effort today (since over-fulfillment means a higher current compensation) versus having to cope with an increased target in the following period. Weitzman (1980) derived a simple, myopic solution for the optimal allocation of effort of the agent (who is assumed to have a certain time-preference). He showed that the ratchet effect induces the agent to deliberately under-perform and to withhold efforts in every period. This odd incentive to withhold efforts is frequently termed a 'sandbagging' reaction (see, e.g., Vergés, 2010).

However, given the information asymmetries present, the above quota-setting rule is the only way for the principal to learn over time about the true characteristics of a particular sales territory. Justification for Weitzman's simple quota updating rule was provided in a more comprehensive theoretical analysis by Mantrala et al. (1997). They showed that from the perspective of the principal, the optimal path of quotas corresponds to a simple, myopic updating rule in the form of (3.19), i.e. the optimal quota is indeed linearly related to the observed sales of the previous period.

Leone et al. (2004) provided one of the few empirical studies of top-down quota setting. They reported strong evidence for *asymmetric ratcheting* in a Fortune500 company. In this company, next-period quotas increase more in response to favorable differences between sales and the current quota than they decrease for unfavorable gaps of the same magnitude. There appears to be a high degree of unobserved heterogeneity of the individual markets, and the authors' explanation for asymmetric ratcheting rests to some extent

[16] A base salary is often required in practice to ensure that the agent is willing to work at all (participation constraint). In the absence of income alternatives, however, a zero base salary does not change his decision calculus.

[17] As indicated by Weitzman (1980), this term was coined by Berliner (1957).

on the assumption that the firm sets quotas to account for these unobservable factors. A disproportionally higher next-period quota reflects the belief of the managers that the positive gap in the current period is due to an increased level of effort which is to be sustained. A shortfall, on the other side, particularly in markets with a low intensity of competition, is assumed to be due to poor performance and should be punished.[18] The empirical results of Leone et al. showed that not only the level of competition but also the tenure of the sales agent both dampen the asymmetric ratcheting effect. The reverse held for sales agents with high within-year sales volatility.

Overall, since the learning process for the principal is typically rather lengthy, the top-down approach is less appropriate in a setting with significant demand variability over time. Therefore, the top-down approach will be disregarded in the following.

Participative Quota-Setting: In a participative quota-setting process, it is the sales agent who provides essential input via his demand forecast. The actual quota is usually set equal to or slightly above the agent-provided sales forecast as it is normal management practice to encourage sales agents to "stretch their performance rather than restrain their output" (Mantrala and Raman, 1990, p. 190). In giving the agents a say, a participative quota-setting process may improve the motivation of the sales force,[19] but obviously opens the door for game-playing behavior, as will be shown in the following.

For simplicity, assume that the principal sets the sales quota q_t for the agent for period t exactly equal to his forecast report $\hat{d}_t$.[20] In the following, the time index t will be dropped if only a single period is considered. An even simpler compensation scheme than the previously introduced commission-based scheme (3.18) is to pay a fixed *bonus* B^0. It is granted in full once actual sales at least amount to the previously set quota q, i.e.

$$B = \begin{cases} 0 & \text{if } y < q, \\ B^0 & \text{if } y \geq q. \end{cases} \tag{3.20}$$

A risk-neutral agent will choose his level of effort to maximize his net utility, i.e. the difference between expected compensation and cost of effort. As long as the quota q is set equal to the demand forecast $\hat{d}$, there is a clear incentive for the agent to understate his sales forecast. However, the wide-spread use of such ill-designed schemes has been aptly summarized by Jensen (2003) as "paying people to lie".

In practice, this problem can be attenuated to some extent if the principal uses historic sales information to evaluate the forecast report of the agent. Any severe deviations from historic values will at least warrant a more in-depth investigation by the principal.

[18] The authors offered a second explanation why quotas are seldom reduced: Regional managers regularly balance their sales territory by ensuring a similar expected workload for all agents. Agents who have seen their sales surge in the previous period will be assigned less customers in the following period to allow them to focus on these high-volume clients. Agents with disappointing results usually receive additional target clients.

[19] See e.g. Wotruba and Thurlow (1976); a more general perspective on sales force motivation was given in McCarthy Byrne et al. (2011).

[20] Note that an additional index, which would indicate the period in which the quota was set, has been omitted here to keep the notation simple.

However, a more problematic drawback of a bonus-based scheme is that it does not provide any motivation for over-fulfillment since the agent does not receive any additional compensation once he has reached the target (Weitzman, 1976).

A commission-based scheme in the form of (3.18)[21] is a similar way to pay agents to lie. With costly effort, the motive to understate the forecast is even higher. Over-fulfillment of an intentionally too low quota will be rewarded. To mitigate such excesses in practice, the principal is likely to place a cap on the total compensation paid out. However, this has the effect of limiting the efforts of the agent if he is already very productive.

Given these rather deceptive results, a principal may be tempted to directly incentivize the agent to submit a particularly high forecast report. This can be done by paying the agent at a linear rate a per unit of the self-selected quota q. Additionally, a commission may be used to punish any under-achievement and to reward over-achievement of the quota. Such a scheme can have the following form:

$$B = a \cdot q + b(y - q) \quad \text{with } a, b > 0. \tag{3.21}$$

However, this scheme is not attractive, either. On the one hand, if $a < b$, the agent will submit a too low target and will be paid at a rate of $(b - a)$ for every unit of actual sales above the quota. On the other hand, if $a > b$, the scheme creates an incentive to rather *overstate* the sales target (see Mantrala and Raman, 1990, p. 191).

Like top-down quota setting processes, participative approaches can also give rise to unfavorable inter-temporal phenomena. For example, assume that a simple bonus-based scheme in the form of (3.20) is used. Over time, it will be observed that an agent exercises (extra) sales efforts primarily towards the end of a multi-period interval, shortly before his bonus will be paid out. For example, if bonus payments are made annually, the marginal benefit of effort might not be known in early periods of the year when sales prospects are still uncertain. Usually, the outlook will become clearer towards year-end. If 'making quota' appears realistic, the agent will exercise additional efforts, leading to spikes in sales volumes at year-end. This has been termed the 'hockey stick phenomenon' (see e.g. Lee et al. (1997b); Sohoni et al. (2010) provided a more in-depth discussion). This accumulation of orders has a particularly adverse effect on the planning quality in the supply chain. In fact, it corresponds to one form of *order batching* which is one of the key sources of the bullwhip effect in supply chains (see Section 2.1.1).

[21] Obviously, also non-linear commissions are possible. For example, a close relative to the linear commission is the so-called *stair-step* incentive which is primarily used in the automotive industry. Sohoni and Mohan (2005) provided a comprehensive analysis: Under the stair-step incentive, the sales agent receives a constant payment per additional vehicle sold until total sales exceed a certain threshold value. Higher sales qualify for an even higher per-vehicle commission which is granted for *all* sales. Total compensation is thus a piece-wise linear, convex function of sales. In practice, the scheme turned out to have some serious flaws. While intended to boost sales at Chrysler, the scheme rather led to decreasing sales as agents restricted their sales efforts when realizing that they were unlikely to make sufficient sales in the current month to qualify for the next commission level. Sales efforts were then postponed to improve chances in the following month.

Misra and Nair (2011) reported clear empirical evidence from the direct sales force of a US contact lens manufacturer whose sales agents indeed adjust their levels of effort depending on their current sales position relative to the period goals.[22] To mitigate these sales spikes, overlapping periods may be used for the assessment of the agent. Evaluating the agent for the current and the past L periods using a moving time window can improve his motivation to continuously work hard and to smooth his output, but such a scheme may be very expensive to administer (see Chen, 2000).

Concluding this brief discussion of bonus- and commission-based schemes, it can be noted that neither of these commonly used simple compensation schemes provides sufficient incentives to the agent to work hard *and* to submit truthful forecasts. More complex schemes will be analyzed briefly in the following section.

Endogenously-Derived Compensation Schemes

A number of microeconomic models have been proposed since the mid-1980s employing the principal-agent framework to derive compensation schemes. By explicitly observing the objectives of the principal and of the agent, these models aim at providing *tailored incentives* to the agent. As before, the principal is modeled as a Stackelberg leader who first determines the form of the compensation plan and sets the level of the incentives. The sales agent observes his private information and chooses his level of effort according to the incentives provided. In following the groundbreaking works of Harris and Raviv (1978, 1979), these models primarily focus on solving the moral hazard problem.

The first contribution to endogenously derive an optimal sales force incentive contract was the seminal paper by Basu et al. (1985) (subsequently referred to as BLSS). In the absence of information asymmetries, the authors derived an optimal compensation plan for an agent in an uncertain selling environment. The agent suffers a disutility from exercising effort and has a desire to work less. The compensation plan shall ensure that the agent works sufficiently hard. The optimal pay function turned out to be a convex, increasing function of sales, but is rather difficult to implement in practice. In follow-up papers, the BLSS framework was extended to situations with asymmetric information and with heterogeneous sales agents, i.e. different abilities and skill levels (Lal and Staelin, 1986; Rao, 1990). For example, in the Rao (1990) model, the principal presents a menu of plans to the agent. The chosen contract allowed the identification of low-skill salespeople and thus also solved the hidden characteristics problem.

While all compensation plans which have been inspired by the BLSS paper set matching incentives to solve the sales effort problem, they still failed to elicit truthful upfront demand forecasts from the agents. As with the basic quota-based schemes (3.18), (3.20) or (3.21), the agents are not punished effectively for forecast misrepresentations.

A second major drawback of compensation plans of the BLSS type is the large number of plan parameters which need to be determined, particularly in a heterogeneous sales environment with many sales agents (each potentially with different productivities levels

[22] Contact lenses are a very convenient product for such a study as the demand curve experiences hardly any seasonality and prices rarely change.

of risk aversion, disutilities of effort and alternative employment opportunities) and multiple sales territories (with different sizes, sales response functions). Under the BLSS-type plans, the commission structure of each agent has to be adjusted if the characteristics of a single sales territory change or if a single agent is relocated. Raju and Srinivasan (1996) discussed the problem of implementing the BLSS plan in a heterogeneous sales environment, showing that a commission-based plan in the form of (3.18) actually constitutes a sufficient approximation to the non-linear BLSS-type compensation schemes. They proved that only slight optimality losses will result compared to the BLSS plan if only quota adjustments are used to adapt to changes in structural parameters while holding salaries and the commission structure fixed. This is consistent with observations from practice where firms rather change quotas than the salary structure (Leone et al., 2004). Furthermore, firms seem to prefer setting compensation plan parameters such as commission rates uniformly to all sales representatives to avoid conflicts and morale problems (Darmon, 1979). Overall, the results of Raju and Srinivasan (1996) may help explain the prevalence of commission-based schemes in practice. However, as discussed above, the latter cannot solve the forecast misrepresentation problem and neither do the endogenously-derived contracts. An incentive scheme without such a flaw is the Gonik scheme which will be discussed next.

The Gonik Scheme

The fundamental problem of the compensation schemes outlined above is that neither of them provides incentives to the agent to provide a truthful forecast *and* to work hard. In their attempt to characterize compensation schemes which encourage the agent to fulfill both requirements at the same time, Mantrala and Raman (1990, p. 191) have stated the following four key requirements:

- Choosing and fulfilling a high quota must be more favorable than choosing and fulfilling a lower quota.
- Choosing and meeting a high quota must be preferred to over-fulfilling a low quota.
- Given any chosen quota, rewards must be higher for over-fulfillment than for rather meeting the quota.
- Similarly, given any chosen quota, meeting the quota must provide higher rewards than under-fulfillment.

The only incentive scheme which simultaneously fulfills all these requirements is the OFA model (objective, forecast, actual). It is known more commonly as the Gonik scheme after Gonik (1978) who reported about his experiences with the scheme at IBM Brazil.[23]

[23] As discussed below in more detail, this scheme has been known before in the economics literature. In a business context, a similar scheme has previously been described by Ijiri et al. (1968). Thomson (1979) provided a formal analysis of the entire class of incentive schemes which encourage truthful revelation of private information for which he coined the term *elicitation scheme*. Comprehensive analyses of such schemes with a focus on Gonik's variant have appeared in Mantrala and Raman (1990) and in Chen (2005).

The interaction between the principal and the agent consists of three steps: In a first step, the principal communicates both a tentative sales target q' and the parameters of the scheme (objective). Then, the agent submits his quota suggestion $\hat{d} = q$ (forecast) and subsequently exercises efforts to realize sales. In a last step, the compensation will be paid to the agent after the actual revenues in the sales period have been observed by the principal (actual). The key idea behind OFA or the Gonik scheme is that any increase in compensation should be in line with the level of ambition of the self-selected quota, with guards against over- and under-fulfillment. The generalized form of the Gonik scheme corresponds to a piece-wise linear curve which has a discontinuity at the self-selected quota (see Mantrala and Raman, 1990, p. 192):

$$B = \begin{cases} a(q - q') + c(y - q), & y < q \\ a(q - q') + b(y - q), & y \geq q \end{cases}, \qquad c > a > b > 0. \tag{3.22}$$

The agent receives a reward of a for each unit by which his self-imposed target q surpasses the principal's quota suggestion q'. Under-fulfillment of the self-selected quota q is penalized at a rate of c whereas the reward for over-fulfillment corresponds to a commission of b per unit above the quota. Since $c > a > b$, there is an a-posteriori penalization for forecast errors (see Vergés, 2010) and the agent is induced to submit a truthful forecast. Moreover, by carefully selecting the parameters of the scheme, the principal is able to directly control the ex-post probability of plan fulfillment, $P(y \geq q)$.[24] If the agent's level of effort can be assumed to be fixed, this probability is given by the following simple ratio (see Mantrala and Raman, 1990, p. 194)

$$P(y \geq q) = \frac{c - a}{c - b}. \tag{3.23}$$

This means, for example, that the principal can induce the agent to report a forecast $\hat{d} = q$ which corresponds to the median point of the probability distribution of the actual sales. By setting $a = 2, b = 1, c = 3$, the probabilities that actual demand will be higher or lower than the forecast report will both equal 50%. In other words, the forecast report will on average correspond to the actual demand.

The attractiveness of the Gonik scheme stems from the fact that it maintains the property to elicit truthful forecasts even if the level of effort is a variable which is controlled by the agent (see Miller and Thornton, 1978). If the level of effort is chosen by the agent, the principal can still control the accuracy of the forecast report; Mantrala and Raman (1990) derived an equation similar to (3.23) for the case with costly effort.

The Gonik scheme has a long tradition in the economics literature where it is usually termed the *New Soviet Incentive Scheme* (Weitzman, 1976), referring to reforms in the planned economy of the Soviet Union during the 1960s. The 'new' scheme was introduced to incentivize the executives of the state-run companies to send accurate forecasts of their production capabilities to the central planning bureau. In the context of a centrally

[24] Recall that it has been assumed that the market environment is stochastic and unobservable to the principal.

planned economy, Miller and Thornton (1978, p. 433) have pointed out that property (3.23) is particularly relevant for firms producing intermediate products. In planned economies, output forecasts are usually made for multiple periods, e.g. in the form of 5-year plans. An under-fulfillment of the forecast for such intermediate goods may put the overall output of the entire economy at risk.

In the economics literature, a few refinements of the Gonik model have been discussed: Risk aversion by the agent was introduced in Snowberger (1977). Osband (1989) presented a model where an agent with an expertise level which is unknown to the principal can refine the accuracy of his forecast at constant marginal costs per unit of precision, but his efforts do not affect overall sales. The principal's objective is then twofold: to induce both a truthful revelation of market knowledge and to encourage an appropriate degree of learning. Thus, the problem has an additional adverse selection component as more expert forecasters have cheaper costs per unit of precision.

In an attempt to extend the Gonik scheme to a multi-period game, Murrell (1979) allowed the agent to build inventories. The inventory levels are assumed to be unobservable for the principal. Surplus may be kept for later periods to ensure making plan in the current period and to reduce the need to exercise effort in the following period. In this setting, the Gonik scheme maintains its properties if the discount rate of the agent is high, otherwise, large inventories may result.

To summarize: The above discussion has illustrated that under the Gonik scheme, each individual sales agent has a strong incentive to submit unbiased forecasts, even if effort is costly to him. More specifically, truth-telling corresponds to a dominant strategy equilibrium, i.e. each agent is better off by submitting truthful forecasts, independent of what all other stakeholders in the sales organization do. This allows using the Gonik scheme to compensate sales agents on a mid-term planning horizon (e.g. regarding the annual sales forecast). Unfortunately, the Gonik scheme no longer enforces truth-telling when applied in a supply-constrained supply chain where the forecasts are used for product allocations. For this short-term planning problem alternative schemes have to be used if truthful forecasting must be enforced. Such schemes for supply-constrained supply chains will be explored in the following section. Note that this need for differentiated incentives and thus different compensation schemes suggests that it may be advantageous to employ a combination of schemes in practice. This aspect will be addressed briefly in Section 3.4.5 when discussing applications to customer hierarchies.

3.4.3. Compensation Schemes for Supply-Constrained Supply Chains

The setting with a supply-constrained supply chain involves short-term allocation and rationing problems such as the DMC problem. As discussed, sales efforts matter less in supply-constrained supply chains. It will therefore be assumed that the problem of hidden actions, or shirking, is of a minor importance. As a result, the objective of the agent no longer consists of minimizing efforts while attaining an acceptable income. Rather, he will

be assumed to focus on maximizing his (financial) reward, giving him an incentive to ask for the highest possible allocation. The discussion in Section 3.2 has already shown that it is primarily the class of individually responsive allocation schemes whose members allow the agents to manipulate the overall product allocation. Proportional allocation and the turn-and-earn scheme appear to be used most widely, while non-responsive schemes like the uniform allocation seem to be irrelevant in practice (see, e.g. Cachon and Lariviere (1999a, p. 840), Cachon and Lariviere (1999c, p. 1091), Furuhata and Zhang (2006, p. 31) or Lariviere (2011, pp. 567–568)). Detailed characterizations of the proportional allocation and other popular schemes will be provided later in Section 4.2.

Supply-constrained supply chains are characterized by the fact that the allocation to agent l, x_l, will usually be less than his forecast report, i.e. $x_l < \hat{d}_l$. This will lead to problems if the compensation remains linked to his turnover y_l in the *current period*. Assume that agent l receives a commission which is proportional to actual sales y_l in the form of

$$B_l = b \cdot y_l, \quad b > 0. \tag{3.24}$$

Obviously, as the agent's earning potential is limited by the actual allocation x_l rather than by the market potential, he will exaggerate his forecast report to obtain a high amount x_l, so $\hat{d}_l \geq d_l$. Conversely, if a fixed bonus in the form of Equation (3.20) is used, the agent may even understate his forecast report to obtain a target which is particularly easy to reach. Overall, it is therefore useless to pay the agent according to actually achieved sales y_l, neither via (3.20) nor via (3.24), as effort is of no importance.

Assume now that the Gonik scheme is being used in the presence of the rationing problem.[25] With an individually-responsive allocation scheme, the allocation (and thus the earning potential) to a particular agent will depend on the forecasts submitted by all other agents. As before, truth-telling is no longer rational for any of the agents. Moreover, truth-telling does not even constitute a Nash equilibrium under the Gonik scheme with allocations. Each agent can benefit by changing his forecast unilaterally: Given all other agents hold to their strategy of sending truthful forecasts, each agent can gain a higher allocation (and thus higher compensation) by sending a biased forecast. In the presence of resource allocation decisions, no elicitation scheme of the Gonik type can provide optimal incentives (Conn, 1979, p. 271, Proposition 2).

Therefore, linking the compensation of the agent either to his achieved *sales in the current period* or to his *forecast report* is problematic. In the absence of hidden actions, it is reasonable to decouple the agent's income from these two quantities. One option is to have the principal simply pay a constant wage. This leaves the agent indifferent between reporting the truth and any other action (see Levinthal, 1988).[26]

[25] For example, Chen (2005) discussed the Gonik scheme in the context of the principal making production/inventory planning decisions.

[26] A second alternative is to link the compensation of the agent to his sales in one or several previous (not the current!) periods. This approach corresponds to the so-called turn-and-earn allocation scheme. It will be discussed in more detail in Section 4.2.

However, this may become problematic over time if the agent wants to maximize his compensation in the long-term. The logic here works as follows: A high customer service, e.g. as measured by a high service rate (i.e. no stockouts) has been shown to increase customer loyalty in the long run (e.g. see Reichheld, 1993). High loyalty, in turn, makes it easier for the agent to sell in later periods. Not only will commanding over a loyal (and thus ultimately more profitable) customer base raise the profile of the agent within the sales organization, it will also mean that selling will require significantly fewer efforts in later periods when supply is no longer scarce and the agent is rather facing a demand-constrained supply chain. In sum, while a fixed compensation is a simple solution which at least does not set wrong incentives, it neither provides the right incentives to the agent to behave in favor of the principal, i.e. truth-telling is by no means a dominant strategy. In the following, two alternative compensation schemes for supply-constrained supply chains will be discussed, profit sharing and the Groves scheme.

Profit Sharing

The key problem of the principal-agent setting with asymmetric information is that the individual agents do not have an incentive to behave in a manner which is consistent with the overall objective of the principal whose objective is maximizing overall (firm) profit. In the absence of explicitly set incentives, opportunistic behavior of the agents will pay off. A drastic solution to this problem consists of simply enforcing the principal's objective directly upon the agents by letting them partake in the firm's total profits.

For a simplified presentation, it will be assumed in the following that the agents submit a report of the profit contribution which results from the selling activities in their sales area, not just demand forecasts as was the case before. In line with the previous notation, the *function of overall actual profits* which are generated by the sales activities of agent l are given by $\pi_l(x_l)$. These profits clearly depend on the allocation x_l which will be sold in its entirety. It is reasonable to assume that π_l is well behaved, i.e. that a higher allocation leads to a higher actual return ($\frac{\partial \pi_l}{\partial x_l} > 0$). However, since the actual profit function constitutes the private knowledge of the agent, it may differ from the *reported profit function* which will be denoted by $\hat{\pi}_l(x_l)$.

For the allocation step, the principal asks all agents to submit a forecast of their profit functions $\hat{\pi}_l(\cdot)$. This function gives the principal the reported profits for any level of allocation x_i. Given the constrained supplies S, her profit maximization problem can be described by the following program:[27]

$$\text{Max} \sum_l \hat{\pi}_l(x_l) \tag{3.25a}$$

subject to

$$\sum_l x_l \leq S \tag{3.25b}$$

[27] This problem will be discussed in more detail in Chapter 4; finding a solution will actually turn out to be difficult in the case of multi-stage hierarchies.

The solution to this problem consists of the optimal allocations x_l. They will be used by the agents to generate profits. To determine the compensation to be paid to the agents, the principal simply monitors the actual total profits of all agents after the selling season. Such an ex-post monitoring is possible since the actual sales volumes are known then. The overall total profits of all agents are given by

$$\Pi = \sum_l \pi_l(x_l). \tag{3.26}$$

The principal uses the following profit sharing scheme to reward each agent l:

$$B_l = \alpha_l \cdot \Pi, \quad 0 < \alpha_l \leq 1. \tag{3.27}$$

Here, the value α_l constitutes a fixed and ex-ante set parameter which describes the share of the overall profit agent l will receive. Hence, agent l can only affect his compensation B_l by making appropriate allocation decisions which affect the value of Π via π_l. Furthermore, also $\sum_l \alpha_l < 1$ will hold (more on this below). For a given allocation mechanism, the total profits Π, from the perspective of agent l, are a function of his own profit forecast report $\hat{\pi}_l$ as well as of the reports of all other agents, $\hat{\pi}_{-l}$. In formal notation, this will be expressed as $\Pi(\hat{\pi}_l; \hat{\pi}_{-l})$.

The compensation scheme (3.27) gives each agent an incentive to report truthfully. Reporting an exaggerated profit function $\hat{\pi}_l(\cdot) > \pi_l(\cdot)$ will result in agent l receiving a too high allocation $x_l' > x_l$. The excess allocation should have better been given to another agent with a higher actual profit function. Hence, any exaggeration will result in lower overall profits $\Pi(\hat{\pi}_l(x_l'); \hat{\pi}_{-l}) < \Pi(\pi_l(x_l); \hat{\pi}_{-l})$ for given forecast reports of all other agents. Conversely, deliberate underreporting of the own profit function is not a reasonable strategy either. The other agents will receive a too high allocation which they either cannot sell completely or only for a lower unit profit. Total profits will be lower than under truthful reporting by all parties.

As a result, no agent has an incentive to depart from telling the truth if all other agents do the same. Sending truthful forecasts thus constitutes a Nash equilibrium (see Loeb and Magat, 1978a). Furthermore, each agent has an incentive to actually fulfill his sales quota (i.e. allocation) after the allocated quantities x_l are known.

As an intermediate remark, note that the objective function (3.25a) only contains the *gross* profits as the costs of the profit sharing scheme are not included. However, this is not a problem. The *net* profit of the principal, after accounting for the costs of the compensation scheme, corresponds to $(1 - \sum_l \alpha_l) \cdot \Pi$. This term differs from the gross profit only by a proportional factor. Hence, any allocation $x_l^\star$ which maximizes gross profits simultaneously maximizes net profits. By adjusting the α_l, the principal can directly control the costs of the compensation scheme. To avoid bankruptcy, she will at least ensure that $\sum_l \alpha_l < 1$.

In connection with profit sharing schemes, three key problems have been discussed in the literature:

- First, the Nash equilibrium reached by truth-telling is not unique if the optimal firm profit can also be attained by other forecast reports.[28] Assume a firm has two agents. Let agent 1 become aware of the fact that agent 2 has submitted a biased forecast report $\hat{\pi}_2(\cdot) = \gamma \cdot \pi_2(\cdot)$, $\gamma > 1$, which will distort the optimal product allocation. Under a proportional allocation scheme, it is in the best interest of agent 1 (and also of the firm!) if he, agent 1, sends a "compensating biased forecast" (Loeb and Magat, 1978a, p. 113) $\hat{\pi}_1(\cdot) = \gamma \cdot \pi_1(\cdot)$. The subsequent proportional rationing step will then result in the same allocations as if both agents had submitted truthful forecasts. Therefore, submitting these biased reports also corresponds to a Nash equilibrium!

  However, such additional equilibria must be considered as mere theoretical oddities. All agents must be able to detect biased forecast reports of the other agents in order to adjust their own reports accordingly to obtain a firm-wide optimum. This means that they need perfect information regarding the marginal profits and true demands of each fellow agent. This is unrealistic in practice.[29]

- The second problem relates to the presence of hidden actions. If the agent incurs a disutility from exerting (unobservable) effort, Cohen and Loeb (1984) have shown that a profit sharing scheme where effort is costly generally does not possess any Nash equilibria at all. The underlying cause is the presence of a *freerider phenomenon*. When working particularly hard, agent i will only receive a fraction of his entire profit contribution. In the same manner, he does not have to bear the full consequences of his under-performance. However, as discussed before, the problem of hidden costly efforts is likely to be less critical in the case of the supply-constrained supply chain where information asymmetry is the main concern.

- Profit sharing is often opposed in theoretical research since the rule violates the *controllability* principle in management accounting (see e.g. Waller and Bishop (1990)). This principle states that agents should only be evaluated based on performance indicators which they can control themselves (e.g. accuracy of forecast reports, the level of effort). In practice, this principle often leads to an evaluation based on profit centers. This means agents are assessed only against key performance indicators (KPIs) which they can control themselves. The downside is that such an evaluation based on profit centers no longer has a firm-wide perspective and thus leads to incentive misalignments.

  In a true profit sharing scheme, the compensation of each agent is directly dependent on the accuracy of all other forecast reports and on each fellow agent's level of effort, via Π. In other words, each agent will directly bear the consequences of all

[28] This has been pointed out by Loeb and Magat (1978a); see the original paper for more details on the following argumentation.

[29] Furthermore, such additional equilibria are impossible in the customer hierarchy model which is employed in this thesis. One major assumption is that the market information of each sales agent is strictly his private information. Hence, the other agents simply cannot assess to which extent forecast reports need to be biased.

other agents' actions. However, a strict application of the controllability principle is rarely possible in practice. Effectively, this would require each agent becoming an independent entity which is not subject to external influences. Many researchers therefore relax the controllability principle and allow the evaluation measure of a particular agent to also depend on the *reports* of the other agents, but not on other agents' effort levels (see e.g. Groves and Loeb (1979, p. 225)).[30] Once the level of effort does not matter (as assumed in the context of selling in supply-constrained supply chains), also the profit sharing scheme fulfills this relaxed controllability principle.

To summarize: A profit sharing scheme provides incentives to make truth-telling a rational decision for a profit-maximizing agent and thus constitutes a Nash equilibrium. Unfortunately, the incentive compatibility is only weak since truth-telling is only rational if every agent is honest. Truth-telling is therefore by no means a dominant strategy, as the decisions of the other agents matter. The following section will discuss the so-called Groves scheme which does not suffer from this defect.

The Groves Scheme

There is indeed a compensation scheme under which truth-telling constitutes a dominant strategy, even if the forecast reports of the agents are used by the principal to allocate a scarce common resource. This scheme is commonly referred to as the *Groves scheme*. Its basis is the famous Vickrey-Clarke-Groves (VCG) mechanism (Vickrey (1961), Clarke (1971) and Groves (1973)) which was introduced as an incentive-compatible means to allocate public goods.[31] The Groves scheme encourages agents to truthfully reveal their private information, putting the principal in a position to use these reports for an optimal allocation of a scarce resource.

The Groves mechanism was initially presented as a coordination scheme for centrally planned economies (e.g. Loeb and Magat, 1978b). Similarities to problems in divisionalized firms were quickly pointed out, particularly with respect to budgeting (e.g. Loeb and Magat, 1978a; Groves and Loeb, 1979). Feldmann and Müller (2003) gave a review of further contributions which address the Groves scheme and also discussed its application in the context of supply chain coordination problems.

An overview of the scheme will be given in the following paragraphs. The basic setting is again a supply-constrained supply chain, as in the case of the profit sharing contract. Overall supply is limited to S product units which have to be allocated to the individual agents. As before, each sales agent generates profits $\pi_l(\cdot)$, but this actual profit function constitutes private information of the agent. Each agent has to send a (potentially biased) *reported* profit function $\hat{\pi}_l(\cdot)$. Then, the principal solves the allocation problem (3.25) based on all messages $\hat{\pi}_l$ to determine the allocation for each individual agent x_l.

[30] For example, this relaxed controllability principle is fulfilled by the Groves scheme which will be discussed in the next section.

[31] The VCG mechanism is closely related to second-price sealed-bid auctions, modified versions of which are frequently employed in electronic market places such as eBay.com (see Lucking-Reiley, 2000).

First, take the perspective of agent i and consider the factors which influence his allocation x_l. Given an individually responsive allocation scheme, it is a function of his profit report $\hat{\pi}_l$ and of those of his peers $\hat{\pi}_{-l}$. This will be denoted as

$$x_l(\hat{\pi}_l; \hat{\pi}_{-l}). \tag{3.28}$$

As a consequence, the *realized* profit contribution of agent l is also determined by both his own and all other profit reports. This will be indicated by

$$\pi_l(\hat{\pi}_l; \hat{\pi}_{-l}). \tag{3.29}$$

Now consider the determinants of the *realized overall* profits $\Pi = \sum_l \pi_l(x_l)$: When aggregating the individual profit functions (3.29), the overall profits are determined on the one hand by *all actual profit functions* (π_l and π_{-l}) and on the other hand by *all reported profit functions* ($\hat{\pi}_l$ and $\hat{\pi}_{-l}$). Hence, these four components should also be used in determining the compensation of agent l. This can be expressed formally as

$$B_l(\pi_l; \pi_{-l}; \hat{\pi}_l; \hat{\pi}_{-l}). \tag{3.30}$$

For such a compensation scheme B_l to be an optimal one, two elementary properties need to be fulfilled (see Loeb and Magat, 1978b, p. 178):

- *Operational desirability:* For any π_l, π_l' with $\pi_l' > \pi_l$, it is required that $B_l(\pi_l'; \cdot; \cdot; \cdot) > B_l(\pi_l; \cdot; \cdot; \cdot)$. Rewards must increase with the actual profit function.
- *Message desirability:* $B_l(\cdot; \cdot; \pi_l; \cdot) \geq B_l(\cdot; \cdot; \hat{\pi}_l; \cdot)$. Truthful reporting must be a dominant strategy, independent of the actions of the other agents.

As illustrated earlier, neither the Gonik elicitation scheme, the fixed compensation nor the profit sharing contract are message-desirable. Deviating from a truthful report will pay off, depending on the reports of the other agents. As an alternative, consider the following modified profit sharing compensation scheme (see Loeb and Magat, 1978a,b):

$$\begin{aligned} B_l &= \alpha_l \cdot \left(\pi_l(x_l) + \sum_{l' \neq l} \hat{\pi}_{l'}(x_{l'}) - A_{-l} \right) \\ &= \alpha_l \cdot \left(\pi_l(\hat{\pi}_l; \hat{\pi}_{-l}) + \sum_{l' \neq l} \hat{\pi}_{l'}(\hat{\pi}_l; \hat{\pi}_{-l}) - A_{-l}(\hat{\pi}_{-l}) \right). \end{aligned} \tag{3.31}$$

α_l is again a scaling factor to control the magnitude of the compensation given to agent l. The first term in the bracket in (3.31) corresponds to agent l's *realized* profits given the allocated quantities x_l which in turn depend on the reported profit functions. The second term represents the *reported* profit contributions of all other agents; these reports are common knowledge, both for the principal and the fellow agents. This second component depends on the profit reports of all agents, including those of agent l. Essentially, it

captures the effect which agent l's report has on all other agents. Lastly, A_{-l} is a function which is *independent* of agent l's profit message $\hat{\pi}_l$, but may depend on the messages $\hat{\pi}_{-l}$ of all other agents. This is indicated by the representation $A_{-l}(\hat{\pi}_{-l})$. Intuitively, under the Groves scheme, each agent is *rewarded* for his contributions to total profit, but he is also *penalized* for any erroneous forecast reports which result in a distortion from the optimal allocation of the resource.

It can easily be verified that this scheme indeed induces truthful forecasts: Assume that agent l submits an exaggerated (understated) profit forecast. He will then receive a too high (too low) allocation x_l which he cannot sell as profitably (could have sold more profitably) as some other agent. Hence, due to (3.29), his realized profit will be smaller than under truthful forecasting. Note that this first part of the argumentation already held in the case of the profit sharing scheme. What is different now is the presence of the second term, which penalizes agent l on an ex-post basis for the negative impact his false report has on all other agents. This penalty is evaluated using the reported profit functions. As the third term is independent of the report of agent l, it does not affect his optimization calculus. Overall, the scheme ensures that truth-telling is *strictly* the best strategy for agent l, no matter what the other agents report. A concrete example illustrating that the Groves scheme in fact leads to truthful forecasting will be given in Section 3.4.5.

Regarding the form of the independent component $A_{-l}(\hat{\pi}_{-l})$, Loeb and Magat (1978a,b) have made an appealing suggestion. They proposed calculating quantities $\bar{x}_{l'}$ according to the following program:

$$\text{Max} \sum_{l' \neq l} \hat{\pi}_{l'}(\bar{x}_{l'}) \tag{3.32a}$$

subject to

$$\sum_{l' \neq l} \bar{x}_{l'} \leq S \tag{3.32b}$$

The resulting form of the Groves scheme becomes (Loeb and Magat, 1978a):

$$B_l = \alpha_l \cdot \left(\pi_l(\hat{\pi}_l; \hat{\pi}_{-l}) + \sum_{l' \neq l} \hat{\pi}_{l'}(\hat{\pi}_l; \hat{\pi}_{-l}) - \sum_{l' \neq l} \hat{\pi}_{l'}(\bar{x}_{l'}) \right). \tag{3.33}$$

The above substitute for the last term A_{-l} represents the total profit which can be generated by all other agents if agent l will be left out of the allocation game. Therefore, the form (3.33) of the Groves scheme pays each agent exactly according to his individual contribution to total profits. If all agents submit a truthful forecast, (3.33) corresponds to the difference between overall profits with agent l and overall profits if agent l is not served. In other words, Equation (3.33) captures the "opportunity cost" of not serving agent l (Loeb and Magat, 1978b, p. 180).

Alternative definitions of A_{-l} can be used to ensure that no agent receives a negative compensation. Furthermore, appropriately chosen A_{-l} can be used to fix the sum of all expected compensation payments *at any arbitrary* level (Conn, 1979, p. 274). This

last point implies that the overall costs of the compensation scheme can be ignored in the profit maximization of the principal. This is similar to the situation with the profit sharing scheme; recall that (3.25a) only contains the gross profits.

Overall, the Groves mechanism guarantees that truth-telling is a dominant strategy equilibrium for each agent: Each agent's reward is increasing in his own realized profits, the reward is independent of the profits generated by all other agents, and no agent needs to know the forecasts of the others to determine his own best forecast report. Moreover, Green and Laffont (1977) have shown that the Groves mechanism is in fact the *only procedure* with this property if effort is of no concern.

Notwithstanding the theoretical advantages of the Groves scheme, there are a few problematic aspects to be considered. An obvious major drawback is its complexity which may help explain why it is not used in practice at all (Arnold et al., 2008) and why it is also rarely applied in the theoretical supply chain literature (for an exception, see Garg et al., 2005). Another problem is that the scheme is not resistant to collusive behavior of the agents. As the compensation of each agent depends on the reports of his peers, by coordinating their messages and jointly biasing their reported profits upwards, all agents can benefit from such a collusion, provided the coalition is stable (e.g. see Arnold et al., 2008, p. 58). This is primarily relevant in multi-period settings where the agents may communicate with each other.[32]

What happens if hidden actions are brought back into the game? Cohen and Loeb (1984) have presented an extended model which includes a moral hazard problem. If the agents need to choose a level of (sales) effort, but dislike working hard, the principal has to compensate the agents for their disutility of work. This has the effect that the overall costs of the compensation scheme can no longer be controlled directly. However, if the costs of the compensation scheme do not matter, i.e. as long as the principal only maximizes gross profits, the Groves scheme still maintains its beneficial properties. This is a difference to profit sharing. If agents have a disutility from exercising unobservable effort, truth-telling no longer constitutes an equilibrium strategy under profit sharing.

Unfortunately, if the principal maximizes overall *net* profits and thus has to consider the compensation paid to the agents as costs, the Groves scheme loses its attractive incentive properties if effort of the agents matters (as shown by Miller and Murrell, 1981). The problem is that the principal needs to encourage her agents to work hard to maximize output and, at the same time, to reveal their true output forecasts. Two types of conflict are at work here: Effort exercised by the agent increases overall output of the principal, but decreases the agent's utility or welfare. Paying an agent more will indeed encourage him to work harder and will incentivize him to truthfully reveal his private information. But these payments at the same time will lead to a decrease of the net surplus of the principal. This conflict cannot be resolved. More precisely, Miller and Murrell (1981,

[32] When no communication is possible, several experimental studies have shown that agents who submit biased forecasts in an attempt to collude tacitly are rarely successful with this strategy. The key studies in this respect are due to Waller and Bishop (1990) and Chow et al. (1994, 2000) and will be discussed in the next section.

Theorem 2) proved that there is no compensation function which—if employed alone—leads to a simultaneous maximization of net outputs *and* to an accurate revelation of market information if managerial or sales efforts matter. To attain both objectives at the same time, the principal always has to rely on additional means. One possibility is to invest in a close monitoring or auditing of the agent to be able to detect shirking (Miller and Murrell, 1981, p. 270).

This last aspect is typical of practical applications. Companies rarely use only a compensation scheme. Usually a mixed strategy is employed to mitigate agency problems. Besides incentive payments, such mixed strategies may include investments in monitoring and a number of other 'softer' factors such as corporate culture. Nevertheless, compensation schemes still play an important role in practice. The following section will give an overview whether the properties of the compensation schemes discussed above also hold in practical settings.

3.4.4. Forecast Misrepresentations: Empirical and Experimental Evidence

Unfortunately, only a limited amount of empirical research exists which gives insights into the use of the different compensation schemes in practice. Even fewer studies analyze the incentive properties of the different compensation schemes in experimental settings. In the following, some empirical observations will be summarized which have been made both for demand- and for supply-constrained supply chains.

Demand-Constrained Supply Chains with Costly Sales Effort

Most empirical literature contributions address demand-constrained supply chains and fall into one of two categories: fairly broad surveys of (industrial) managers or case studies with a narrow focus. In the latter case, often just the experience of one company over one or a few years is considered.

Based on a comprehensive survey, Joseph and Kalwani (1998) concluded that rather simple compensation schemes prevail in practice. They found that the overwhelming majority of their respondents either use a bonus (72%) and/or a commission payment (59%) to reward their sales force. Given these prevailing compensation schemes, some indications suggest that agency and opportunistic behavior of the sales agents may not be as severe as assumed in the discussion in Section 3.4.2. In an empirical study involving the sales force of a manufacturer of electronic devices, Winer (1973) has arrived at the rather surprising result that the salespeople in scope of his study were rather quota achievers than income maximizers. The results of Good and Stone (1991) point in a similar direction. They surveyed industrial sales managers regarding their perceived importance of various factors in the sales quota development and implementation process. Their respondents generally rated issues outside their control as most important, e.g. the sales territory and the product sold. Interestingly, very little importance was placed on organizational requirements of the quota setting process and on the consideration of past sales forecasts

when determining new quotas. This suggests, for example, that ratcheting appears to be less of an issue in practice.

In a case study of a durable goods company, Steenburgh (2008) observed that sales agents do not tend to reduce effort in response to lump-sum bonuses. Rather surprisingly, he found proof for the opposite: that agents actually work harder in response to these fixed bonus payments.

Chow and Cooper (1991) gave an account of an experimental evaluation of truth-inducing compensation schemes such as the Gonik scheme in a participative quota setting process. As expected, the use of a truth-inducing scheme resulted in significantly less biasing behavior once effort levels were unobservable to the principal. However, Chow and Cooper managed to limit the biasing behavior similarly effectively by using a simple ratcheting scheme. In contrast to expectation, they did not encounter any preemptive forecast biasing before the ratcheting feedback mechanism became effective. As a result, the authors suggested that a sole focus on the truth-inducing property may be too narrow. In fact, the principal will often have other means to ensure her agents behave honestly, for example by closely monitoring the past forecasting performance.

Supply-Constrained Supply chains with Allocation Problems

In contrast to the demand-constrained supply chain setting, only few experimental results have yet been reported for supply-constrained settings. Most of the empirical results discussed in the following have been obtained in a series of experiments in which primarily business major students were involved. All studies used very similar experimental setups. Running multiple experiments in parallel, the researchers formed groups of three people. Two participants were asked to assume the role of an agent while a third person represented the principal. The task of the agents was to submit unit profit reports to the principal, with the objective of obtaining a sufficient allocation to satisfy a given demand. The reports of the agents were evaluated by the principal who strictly allocated the scarce resource based on the forecast reports. This means she was applying a simple rank-based allocation rule, with the reported unit profits determining the priority order in which each agent was served, without questioning the reports.[33] Upon receiving their allocation, the agents were paid according to a pre-announced compensation scheme. Usually, several rounds with multiple participants and several different compensation schemes were conducted.

The first such study has been conducted by Waller and Bishop (1990). In their setting, each agent had a fixed need of 80 units of a scarce resource, but overall supply was limited to 100 units. The agent with the highest unit profit report was allocated 80 units while the remaining 20 units were given to the other agent. Both agents were aware of the nature of the allocation rule (Waller and Bishop, 1990, p. 817). Physical separation of the agents precluded any explicit collusive behavior among the agents.

The study compared the outcomes using three different compensation schemes. In the first scheme, the compensation of each agent l depended linearly only on his own generated

[33] This simple profit-based allocation rule will be discussed further in Section 4.3.

profit (i.e. compensation is only linked to the performance of the profit center which agent l controls). This *profit participation* scheme differs from the previously discussed profit sharing in that the results of all other agents have no influence on the compensation. For agent l, it has the following form:

$$B_l = \alpha_l \cdot \pi_l. \tag{3.34}$$

Here, α_l reflects the share of the profit *only* from profit center l which is given to agent l as his compensation (note the difference to 3.27). The second scheme used in the study by Waller and Bishop was a variant of this profit participation scheme. Any ex-post deviation between reported and actual profits is penalized heavily and results in an income of zero, i.e.

$$B_l = \begin{cases} \alpha_l \cdot \pi_l & \text{if } \hat{p}_l = p_l, \\ 0 & \text{otherwise.} \end{cases} \tag{3.35}$$

The third scheme tested was the Groves scheme.

As expected, Waller and Bishop (1990) found that the frequency of profit misrepresentations was highest for the profit participation scheme. The profit participation scheme with ex-post penalties led to the least amount of misrepresentations, even fewer than under the Groves scheme. However, the misrepresentations under the Groves scheme did not result in higher costs to the principal (i.e. no deviations in terms of total profits) as the misrepresentations usually did not affect the allocation decision. The authors noted that this is likely to be different once the number of agents will be increased to more than two.

Waller and Bishop (1990) gave a number of explanations for the observed outcomes. First, under the profit participation scheme with penalties, truthful reporting is always the best strategy as any deviation can directly be punished by the principal. Under the Groves scheme, each agent actually has two strategies: Truth-telling or overstatement and collusion. Recall that truth-telling is a dominant strategy under the Groves scheme, but that mutually arranged forecast exaggerations may pay off if the coalition of agents is stable.

A simple explanation for the many misrepresentations observed in the experiment when using the Groves scheme is that the mechanics of this scheme may have not been understood properly by the participants. An alternative explanation is that the many misrepresentations can be interpreted as (unsuccessful) attempts to gain from tacit collusions. In other words, the agents were hoping that the other agent would exaggerate as well for their mutual benefit. While this experiment precluded direct interactions between the agents, agents typically will have multiple opportunities in practice to engage in mutual discussions. Waller and Bishop (1990) suspected that their experiment underestimated the probability of (explicit) collusion under the Groves scheme.

In response to the experiment by Waller and Bishop (1990), Chow et al. (1994) pointed towards a possible explanation for the strong dominance of the profit participation scheme with penalties. Their main argument was that the deterministic, linear sales response function which was employed in the first study did not allow for any state uncertainty, neither for the agents nor for the principal. As a result, the profit participation scheme

with penalties can strictly enforce truth-telling. In many practical situations, however, the actual profit is stochastic and unknown to the agent at the time he submits his forecast report. In practice, the information asymmetry between principal and agent exists with respect to the probability distribution of the unit profit parameter. Each agent has a more accurate perspective than the principal, although not a perfect one. In such situations, the profit participation with penalty appears to be a too strict benchmark.

Acknowledging that both the principal and the agent can learn from past observations to make better decisions in the present, Chow et al. (2000) analyzed the performance of several different compensation schemes in a multi-period setup. In addition to the Groves and the simple profit participation scheme, they also introduced a proxy scheme to account for a typical situation in practice: Under the proxy scheme, a subset of the principals (recall that several experiments were conducted in parallel) was allowed to allocate the scarce quantity in a *non-mechanistic* manner. Rather than giving an allocation purely based on the agents' reports, selected principals were allowed to conduct costly and imperfect audits of the unit profit reports of the agents to detect misrepresentations. This subset of the principals could base their allocation decisions on the audit findings, on subjective judgment and on the record of the historic forecast accuracy of their agents.

Chow et al. (2000) found that their proxy scheme was more effective than any other scheme, including the Groves scheme, in reducing both the frequency and magnitude of forecast misrepresentations. However, the costs to conduct the audits were larger than the additional profits from avoided misallocations. Unfortunately, the study did not incorporate any implementation costs of the other schemes, putting this proxy scheme at a disadvantage. While a significant number of misrepresentations occurred under all other schemes including the Groves scheme, many of those misrepresentations were either not severe or actually offset each other. In the end, only a few misrepresentation cases led to a deviation from the optimal allocation. This may be interpreted as an indication that misrepresentations do not have too severe consequences in practical settings.

Overall, the study by Chow et al. (2000) suggested that forecast misrepresentation can indeed be mitigated effectively by the Groves scheme. Yet the same effect can be achieved by a number of other means such as audits by superiors and an observation of the historic forecast performance. This leads again to the conclusion that a sole focus on the compensation scheme may not be an appropriate strategy to ensure truthful forecasts.

Moreover, many features of incentive systems found in practice such as fairness, morale, equity, trust or culture are hard to explain by traditional economic reasoning when employing a purely pecuniary perspective (see Baker et al., 1988). In sum, a rich tool kit seems to exist, consisting of financial and non-financial means to mitigate the principal-agent problems in distributed decision-making. In the following section, the above discussion of forecast misrepresentation and incentive schemes will finally be applied to the case of customer hierarchies and to the DMC problem.

3.4.5. Application to Customer Hierarchies

Multi-stage customer hierarchies are characterized by a number of different planning problems which require accurate forecasts. In a mid-term planning horizon, accurate demand forecasts are important to drive production planning. This situation is a typical example of a demand-constrained supply chain as the supply quantities are still adjustable. As a conclusion from Section 3.4.2, an attractive option to elicit truthful demand forecasts is to apply the Gonik scheme.

However, this scheme fails in the short-term when there is a demand fulfillment problem. As supply is no longer adjustable, the given supply resources should be allocated in a manner which leads to maximum profits for the overall company. The general discussion in Section 3.4.3 has suggested that both the profit sharing and the Groves scheme have attractive incentive properties, but neither are free from disadvantages.

Given these two different problem types, it may be advantageous in practice to rely on a combination of different schemes when determining the compensation to be paid to a particular sales agent. This would permit providing tailored incentives both for the mid- and for the short-term. Such a strategy can exploit that mid-term and short-term forecasts typically involve separated formal processes in most sales organizations. For example, mid-term sales forecasting is often part of the company-wide (bi-)annual budgeting process whereas short-term demand forecast updates may feed into operational S&OP activities. This separation of the formal internal processes should allow using a Gonik-type scheme for the mid-term forecast and a second scheme which better suits the supply-constrained supply chain in the short-term. A more detailed investigation of such a combined compensation scheme, with a particular focus on mutual interdependencies, may be the subject of follow-up research.

Focusing primarily on the short-term DMC problem, the first part of the discussion in this section takes a closer look at the different forecast reports which will be exchanged in a customer hierarchy as introduced in Section 3.3. Afterwards, the applicability of the different compensation schemes for supply-constrained supply chains of Section 3.4.3 will be discussed, followed by a numerical example of a simple principal-agent hierarchy. Lastly, conclusions will be drawn to simplify the model assumptions in the remainder of this thesis.

Forecast Reports in Customer Hierarchies

Once a customer hierarchy faces a supply shortage, the scarce resources need to be allocated so that overall company profits are maximized. If such allocation decisions are made on a decentral basis, the individual planners require profit function reports from the lower levels of the hierarchy. While the overview of compensation schemes for general supply-constrained environments in Section 3.4.3 was based on general actual and reported profit functions π_l and $\hat{\pi}_l$, the setting is simpler in a customer hierarchy as modeled in Section 3.3. Here, the profit functions have a particular functional form.

Under the condition that the functional form of the profit function is common knowledge across all agents and principals, forecast misrepresentation problems in customer

hierarchies primarily exist with respect to the actual values of the profit function parameters, rather than with respect to entire profit functions. It will be argued in the following that the principal can monitor at least some of these parameters.

For a start, recall that in each customer hierarchy, each base customer segment l is characterized by a demand d_l and an associated unit profit p_l. Initially, both these parameters are assumed to be private information of the sales agent who is responsible for base customer segment l. In actual customer hierarchies, usually multiple sales agents (i.e. ≥ 2) report to the same superior sales manager (principal). The index l will be used to refer to a particular agent and '$-l$' will refer to all other agents excluding l. The principal allocates a scarce supply $\bar{x}$. She uses a (weakly) individually responsive allocation scheme to determine allocations x_l which depend on the reports $\hat{d}_l$ and $\hat{p}_l$ of all agents. Ceteris paribus, the allocation rule satisfies

$$\frac{\partial x_l}{\partial \hat{d}_l} \geq 0, \quad \frac{\partial x_l}{\partial \hat{p}_l} \geq 0, \tag{3.36}$$

i.e. the allocation to agent l increases—or at least stays constant—for higher reports of p or d. By contrast, deliberate under-reporting will lead to a lower allocation with certainty (unless the allocation is already zero).

The profit contribution of each agent is

$$\pi_l = p_l \cdot \min\left(d_l; x_l\right) \tag{3.37}$$

and can be monitored by the principal. The objective of the principal is maximizing overall profits.

In the presence of asymmetric information, the agents may cheat with respect to both parameters of the profit function. It is helpful to first discuss to what extent these reports can be verified by their principal in this customer hierarchy setting.

Provided that supply $\bar{x}$ is scarce, at least one of the agents, say agent l, will receive an allocation which is smaller than his report, i.e. $x_l < \hat{d}_l$. Assume that lost sales will not be recorded, i.e. the actual demand in each customer segment is unknown to the principal. If agent l has exaggerated his demand and $x_l < d_l < \hat{d}_l$ holds, the principal cannot detect this type of misrepresentation because the entire allocation will still be sold. Only if $d_l < x_l \leq \hat{d}$ holds, excess quantities will remain as leftovers $\Delta_{d_l} = x_l - d_l > 0$ after the sales period. It is realistic to assume that the principal can monitor this value of Δ_{d_l}. If leftovers remain, the principal can infer that the agent has exaggerated his demand forecast (in the absence of stochastic influences).

Things are different with respect to the unit profits. Recall that the principal can observe the total output, i.e. total profit contribution (3.37) of each agent. Furthermore, for each agent, both the actual allocation x_l as well as any potential leftovers Δ_{d_l} (from lying with respect to the demand) are also observable. Hence, Equation (3.37) allows determining the true p_l after each period. Any misrepresentation with respect to p will be detected on an ex-post basis.

This assumption of observable unit profits can also be defended with practical experiences. Often, differences in unit profits among the individual customer segments can be related to a few single factors, for example distribution costs, taxes, tariffs, or (contractual or perceived) penalties. Many of these influencing factors either do not fluctuate significantly in the short run, or are directly observable by the principal.[34] In the first case, the principal can learn the true unit profits over time; in the second case, she can monitor the actual unit profits at least in an indirect manner. Using the terminology introduced in Section 3.2, the customer hierarchies considered here are characterized by a weak information asymmetry. In other words, there is only a decision time hierarchy with respect to unit profits, while stronger, potentially lasting, information asymmetry may exist with respect to demand.

However, some allocation rules allow for a better (indirect) monitoring of the demand forecasts via leftover inventories. As will be shown in the next chapter, a simple rank-based allocation rule, which is popular in practice (IDA, see Section 4.3.3), allocates the scarce product on an *all-or-nothing* basis in decreasing order of the aggregated unit profits. This implies that demand reports of the agents are either fulfilled in full or not at all (with the exception of at most one agent who will receive only a certain share of his forecast report). This type of allocation rule makes it comparably easy for the principal to register forecast exaggerations via the leftover quantities.

Note that the above argument suggesting that unit profits merely involve a decision time hierarchy not only holds for principal-agent situations involving a sales manager k at the lowest intermediate node and one or several of her associated sales agents l at the leaf nodes of the customer hierarchy, i.e. $l \in \mathcal{L}$ and $k \in \mathcal{N} \setminus \mathcal{L}$. In fact, some of these influencing factors may also be observed at more aggregate levels. Consider the principal-agent situation at the next higher level, involving the superior i of the aforementioned sales manager k, now in the role of the principal, with $k \in \mathcal{N} \setminus \mathcal{L}$. In a geography-based sales hierarchy, the associated formal position in the sales organization may correspond to a regional manager responsible for multiple countries. Even at this level, it will be possible in many situations in practice to observe selected influencing factors directly which induce changes in unit profits.

- On the one hand, some of these factors affect multiple sales districts simultaneously and in the same manner. For example, exchange rate effects will have the same impact on all sales districts which lie in the same currency area such as the Euro zone while tax and tariff changes will at least affect all sales districts within the same country. A regional manager will be able to monitor such developments.

- On the other hand, even if individual sales districts at the disaggregate level are affected to a different extent by such observable factors, often an aggregated figure for many of such factors can be monitored. Examples include many raw material and commodity price indices or the Baltic dry index for freight rates of bulk carrier ships.

[34] Information on some of these factors may also be acquired from neutral third parties, e.g. raw material prices.

The above discussion has focused on a simple rank-based allocation scheme for customer hierarchies where two parameters—demand d_l and unit profit p_l—need to be exchanged between an agent l and his principal k. As discussed before, the compensation schemes may also be applied to incentivize truthful reports from a sales manager at an intermediate node k of the customer hierarchy. In the next chapter, a new allocation scheme will be introduced. Its application requires that the sales managers at the intermediate nodes also submit a third parameter T_k, in addition to p_k and d_k. This additional parameter can be interpreted as a measure of customer heterogeneity in terms of Theil's index (see Section 3.5). A brief analysis is provided in Appendix A.5, showing that this three-parameter scheme poses no additional difficulties. Like the two-parameter scheme, it is weakly individually responsive with respect to the reported demand and the reported unit profit, i.e. expressions in the same manner as in (3.37) hold. However, no such unambiguous relationship exists with respect to higher reports of the level of customer heterogeneity. Only knowing his private information regarding d_k, p_k and T_k, an individual agent cannot assess the impact of a higher report of the parameter T_k. Hence, there is no incentive to benefit from misrepresentations. This in fact a useful property: Since T_k depends both on the lower-level unit profits and demands, the actual realization of T_k is unknown to the principal and cannot be monitored on an ex-post basis, neither directly nor indirectly. This would make an identification of biased reports difficult. However, with T_k not being susceptible to misrepresentations, the following discussion regarding the reporting strategy under the simpler two-parameter profit function (demand and unit profit) also applies to the three-parameter version.

General Reporting Strategy per Compensation Scheme

For a start, it is helpful to analyze to what extent the previously discussed rank-based allocation scheme (quota allotment in decreasing order of the aggregate unit profits) provides incentives for either misrepresentations or truthful reports of the parameters d_l and p_l. Table 3.2 contains an overview of the compensation schemes discussed in the following. For an easier presentation, simple values have been assumed for all terms which do not influence the decision of an agent, but rather only affect the actual numerical value of the paid compensation. In particular, the scaling factor α_l for the profit participation scheme (3.34) has been set equal to 1 so that each agent will earn the entire profit contribution of his profit center. For the profit sharing and the Groves scheme, the factor α_l has been set equal to 0.5 to prevent the principal from paying out more than the overall profit to all his agents together. Furthermore, the term A_{-l} for the Groves scheme has been assumed to equal zero, since this independent component does not influence the decision of agent l.

In the first case, under the **constant wage** scheme, each agent will receive a constant compensation independent of his reports or sales outcomes. While neither agent can gain from lying, the compensation scheme nevertheless cannot prevent any misrepresentations.

Under the **profit participation** scheme (see Equation (3.34)), lying with respect to any parameter may pay off for agent l due to (3.36). However, the other agents will anticipate

Compensation scheme	**Formula**	**Eq.**	**Resulting reporting strategy of the agent regarding d_l, p_l**
Constant wage	$B_l = B^0$	—	Indifferent (no strategy equilibrium)
Profit participation	$B_l = \pi_l = p_l \cdot x_l$	3.34	Misrepresentation (dominant strategy)
Profit sharing	$B_l = 1/2 \cdot (\pi_l + \pi_{-l})$ $= 1/2 \cdot (p_l x_l + p_{-l} x_{-l})$	3.27	Truth-telling (Nash equilibrium)
Groves scheme	$B_l = 1/2 \cdot (\pi_l + \hat{\pi}_{-l})$ $= 1/2 \cdot (p_l x_l + \hat{p}_{-l} x_{-l})$	3.31	Truth-telling (dominant strategy in the absence of collusion)

Table 3.2. – Optimal reporting strategy per compensation scheme

this behavior and have an own incentive to submit a higher-than actual forecast report to counter agent l's lie. However, this ill incentive to lie can be removed for the case of p_l with the help of an additional ex-post penalty in the form of Equation (3.35). As a consequence, truth-telling with respect to the unit profit now becomes a dominant strategy, independent of the actions of all other agents. One may extend this scheme by also including an ex-post penalty for any leftover inventories to create at least a weak incentive to also report the demand volume forecast truthfully. As discussed, this approach does not create a strict incentive for truthful reporting. Any demand exaggerations which do not result in leftover inventories will remain unpunished.

Profit sharing creates an incentive for each agent to truthfully reveal his forecasts if all other agents also tell the truth. Nevertheless, should the other agents choose to submit biased parameters and should this fact become known to agent l, agent l for his part can send a compensating biased profit forecast to restore the original order with respect to unit profits. As discussed above, this is an unrealistic scenario, as it requires each agent to have a knowledge of the private market information of his peer agents. From a practical point of view, profit sharing may suffice to encourage truthful reporting.

Finally, consider the **Groves scheme**. As discussed in Section 3.4.3, any exaggeration of agent l with respect to one of the two parameters will lead to a higher-than-optimal allocation x_l and corresponding leftovers. While this does not affect agent l's personal profit compensation (first term in the compensation formula (3.33)), it will reduce the actual allocation to the other agents. As the second term of the compensation formula corresponds to the profit generated by all other agents at their stated levels of unit profit (for the given resource allocation), any misallocation to the other agents will simultaneously lower agent l's own overall compensation. The difference to the profit sharing scheme is that the other agents no longer have any incentive to counter biased reports of agent l by submitting exaggerated forecasts on their own. On the contrary, any lie of agent l which leads to a distortion of the optimal resource allocation will be beneficial to the other agents, but will decrease agent l's own compensation. As indicated before,

this may open the door for collusion among the agents. Practical experiments, however, indicate that this is a rather theoretical problem. Such coalitions either rarely succeed in distorting the optimal allocation or are simply not stable.

From a theoretical perspective, the Groves scheme sets the strictest incentives to the sales agents and managers in a customer hierarchy to refrain from forecast misrepresentations. It has the main disadvantage of being perceived to be too complex for most practical applications. However, a profit sharing scheme is almost as good in reducing incentives for forecast misrepresentations. Its main advantage is its immediate intuitive appeal.

Numerical Example: Payout per Reporting Strategy and Compensation Scheme

These aspects will now be illustrated with a numerical example. Assume a principal-agent situation consisting of a principal, who needs to allocate a scarce supply of $\bar{x} = 15$ units, and two agents. The latter have an actual profit function of the form (3.37), but the information regarding the two parameters p_l and d_l, $l = 1, 2$ is their private knowledge. Assume that $d_1 = d_2 = 10$ units and that $p_1 = 1$ and $p_2 = 2$, implying that there is an overall supply shortage which requires an allocation. The principal uses a simple rank-based allocation mechanism and serves the agent with the higher unit profit report $\hat{p}_l$ first up to a maximum of $\hat{d}_l$ or until running out of supplies. The remaining supply quantities, if any, will be given to the other agent with the lower unit profit report. Overall, the maximum profit in this situation equals 25 units — it is no more than this amount which can be paid by the principal to both her agents per period in the long run (otherwise, she would risk bankrupcy).

The following Table 3.3 gives an overview of the compensation paid out to each agent under different compensation schemes (indicated by B_1 and B_2 per scheme). Using the compensation formulas given before in Table 3.2, the profit participation, the profit sharing and the Groves scheme will be analyzed — a constant wage is of no particular interest, as discussed above.

The first row of Table 3.3 corresponds to truth telling by both agents, the next three rows address situations in which agent 1 lies while the last three rows cover the case with agent 2 lying. Lying may occur with respect to d_i, p_i or both. As the allocation scheme is individually responsive, only exaggerations will be considered.

First, the values in the table need to be explained. Initially, consider truthful reporting, i.e. the first data row where agent 2 receives 10 units (he commands over the more profitable segment with $\hat{p}_2 = p_2 = 2$) and where agent 1 receives only the remaining 5 units. Under profit participation, each agent exactly earns the profit generated by his profit center, i.e. $B_1 = 1 \cdot 5 = 5$ and $B_2 = 2 \cdot 10 = 20$. Under profit sharing, each agent exactly earns half of the overall profit, i.e. $B_1 = B_2 = 1/2 \cdot (1 \cdot 5 + 2 \cdot 10) = 12.5$. If both agents report truthfully, the actual and the reported profit terms are equal, i.e. $\hat{p}_1 = p_1$ and $\hat{p}_2 = p_2$. This means that also the Groves scheme leads to a compensation of $B_1 = B_2 = 12.5$ units.

Reporting Strategy					Allocation (units)		Profit Participation		Profit Sharing		Groves Scheme	
	$\hat{d}_1$	$\hat{p}_1$	$\hat{d}_2$	$\hat{p}_2$	x_1	x_2	B_1	B_2	B_1	B_2	B_1	B_2
Truth	10	1	10	2	5	10	5	20	12.5	12.5	12.5	12.5
Agent 1 lies	15	1	10	2	5	10	5	20	12.5	12.5	12.5	12.5
	10	3	10	2	10	5	10	10	10	10	10	20
	15	3	10	2	15	0	10	0	5	5	5	22.5
Agent 2 lies	10	1	15	2	0	15	0	20	10	10	15	10
	10	1	10	3	5	10	5	20	12.5	12.5	12.5	12.5
	10	1	15	3	0	15	0	20	10	10	22.5	10

Table 3.3. – Example: Payout per reporting strategy under different compensation schemes

In the second row of Table 3.3, agent 1 exaggerates with respect to the demand d_1. However, since $\hat{p}_1 = p_1 < \hat{p}_2 = p_2$, this exaggeration does not affect the allocation of the scarce quantities. As the allocation is not changed and overall profit remains at 25 units, also the payouts under all three compensation schemes remain as in the case of truthful reporting.

Now consider the third data row, where agent 1 lies with respect to the value of $p_1 = 1$ and reports $\hat{p}_1 = 3$ instead. This leads to a changed allocation of $x_1 = 10$ and $x_2 = 5$ units. However, agent 1 can in fact only generate a unit profit of 1. His actual profit center contribution therefore only amounts to $\pi_1 = 1 \cdot 10 = 10$. Agent 2 remains more profitable, but only has $x_2 = 5$ units to sell, earning a profit of $\pi_2 = 2 \cdot 5 = 10$ units. As a result, the profit participation scheme actually rewards the lie of agent 1, giving him a compensation of $B_1 = 10$, while agent 2, who nevertheless reported truthfully, is only paid $B_2 = 10$.

Under profit sharing, agent 1 is punished for distorting the optimal allocation. Since overall profits are reduced from the optimal value of 25 to only 20, agent 1 also only receives a compensation of 10 (and agent 2, reporting truthfully, is also punished as a result of agent 1's lie and also only receives 10 as his compensation).

Lastly, consider the Groves scheme. The actual payout to agent 1 corresponds to one half of the sum of his own, actual profit center contribution $p_1 \cdot x_1$ and of the reported contribution of his fellow agent 2, i.e. $\hat{p}_2 \cdot x_2$. The latter reported truthfully, so overall payout to agent 1 amounts to $B_1 = 1/2 \cdot (1 \cdot 10 + 2 \cdot 5) = 10$. As in the case of profit sharing, agent 1 is punished for distorting the optimal allocation. Now consider the payout to agent 2. The first part of his compensation relates to his own *actual* profit contribution $\pi_2 = p_2 \cdot x_2 = 2 \cdot 5 = 10$. The second part of his compensation relates to the *reported* profit contribution of his fellow agent 1, who claimed to be able to generate a unit profit of 3! Therefore, agent 1 also received an allocation $x_1 = 10$ units. As a result, the *reported* profit of agent 1 amounts to $\hat{\pi}_1 = \hat{p}_1 \cdot x_1 = 3 \cdot 10 = 30$. Considering these two components, the overall compensation paid to agent 2 under the Groves scheme amounts to $B_2 = 1/2 \cdot (10 + 30) = 20$. Note that the total compensation paid to both agents, i.e.

$10 + 20 = 30$ units, is larger than the available total profit (which only amounts to 20 units, as the exaggeration by agent 1 has led to a distorted allocation). As noted before, one way to mitigate this undesired side effect of the basic Groves scheme is to set an appropriate value for the term A_{-l} to avoid bankruptcy of the principal.

The other payout values in the lower rows can be determined in the same manner. Some of the resulting values are particularly interesting. First, note the expected result that under profit participation, an agent does not have to suffer the consequences from distorting the allocation. For example, agent 1 is not punished for overstating the demand report (row 2) and will actually gain from exaggerating with respect to the unit profit (rows 3 and 4). Agent 2, shown in row 5, will receive a larger allocation due to an overstated demand forecast (increase from 10 to 15 units). However, he is not able to generate a higher actual profit as 5 units cannot be sold in the market which had better been given to agent 1 instead. As a result, while agent 2 does not benefit from his exaggeration, he is also not punished for it under profit participation. Similar results hold if agent 2 exaggerates with respect to the profit report (data rows 6 and 7) — agent 2 always keeps a compensation of 20 units.

By contrast, under profit sharing, an agent will notice a reduction in his payout if his reports leads to a deviation from the optimal allocation, as can be seen in rows 3 and 4 for misrepresented reports of agent 1 and in rows 5 and 7 for false reports of agent 2. Hence, also the other agent (who reports truthfully) will see a reduction in his payout and thus also suffers from misrepresentations of his peer under profit sharing.

This is different under the Groves scheme. As discussed above, the second part of the compensation formula (3.31) for agent l[35] corresponds to the *reported* profit earned by the other agent $-l$ at the allocated level of supply. In the simple example above, this payout component equals $\hat{p}_{-l} \cdot x_{-l}$. Recall the calculation for row 3 where the exaggerated unit profit report of agent 1 leads to a distorted allocation. Under the Groves scheme, agent 2 will actually benefit from agent 1's lies. While the first term of his compensation formula equals 10 (since agent 2 receives only a reduced allocation of 5 unit and earns an actual unit profit of 2), the second term corresponds to $3 \cdot 10$, i.e. the reported unit profit of agent $-l$ = agent 1 times the allocated level of allocation.

This case also illustrates the problem of collusion under the Groves scheme. Agent 2 actually receives a higher compensation (20) than under truth-telling (12.5). In the unlikely case that agent 2 is aware of this situation,[36] he may be tempted to offer a bribe to agent 1 so that the latter submits exaggerated reports. In this case, both agents will benefit from a collusion and may prefer it over truth-telling. This raises the costs of operating the scheme for the principal, and this adverse effect cannot even be mitigated by setting a lower payout factor α_l. Because at least in theory, agent 1 could also report even higher unit profits $\hat{p}_1$ of 10, thus driving up the compensation for agent 2.

[35] The third term, which does not affect the optimal decision of agent l, has been assumed to equal zero here for simplicity.

[36] Recall that this typically requires that information asymmetry only exists between each agent and the principal, but not between the agents.

Conclusions for Customer Hierarchies

The problem of forecast misrepresentations in customer hierarchies in the short-term when supply is no longer flexible can be handled appropriately by the Groves scheme and to a lesser extent by profit sharing. The effectiveness of profit sharing can be increased by taking complementary measures:

1. One option is to let the principal **monitor** both the reported and actual parameters. Deviations between reported and actual values of the parameters which are larger than predetermined safety margins (to account for stochastic influences) will be penalized. As discussed above, this strategy may even lead to fewer misrepresentations than a Groves scheme. Monitoring and ex-post penalization are possible for all parameters where the information asymmetry between agent and principal is due to a decision time hierarchy.

   In the customer hierarchies considered here, unit profits are clearly observable on an ex-post basis. Misrepresented demand values can only be monitored (and penalized) indirectly by observing potential leftover quantities at the end of each forecast interval (or by investing in technology to capture all lost sale cases).

2. Besides monitoring, another effective strategy to prevent forecast misrepresentations is to invest in probing the reports of the agents. The principal does not have to conduct an audit for all reports. The mere chance of being audited may already be sufficient to discourage misrepresentations. In the experiments conducted by Chow et al. (2000), this has led to similarly low levels of misrepresentations as a Groves scheme.

Overall, for truthful forecasting in customer hierarchies, a profit sharing scheme combined with penalties for observed misrepresentations and audits appears to be a viable substitute for a Groves scheme. Indeed, similar schemes are also frequently employed in practice. Often, the *variable compensation* of employees in many companies consists of two parts (see Gerhart and Trevor (2008, p. 69)):

- One part is linked directly to the overall success of the company. It can be interpreted as a profit sharing scheme.

- Another part depends on the achievement of individual objectives. For a sales agent or sales manager, such an objective can have the form of a particular forecast accuracy target. Any deviations will lower the individual component of the variable compensation.

All taken together, the above argumentation indicates that there are indeed sufficient means to manage forecast misrepresentation problems in customer hierarchies. It will therefore be assumed that agency problems do not constitute a severe problem in customer hierarchies. More clearly, in all models which will be presented in the following chapters, sales agents and managers will be assumed to submit truthful reports.

In the above discussion of forecasting and sales staff compensation schemes, one prerequisite has been assumed to hold: The individual customer segments are sufficiently heterogeneous in terms of demand and profitability. The more heterogeneous a customer organization is, the more critical it is for the principal to elicit exact forecasts to allocate the scarce product quantities in the most profitable manner. This raises the management problem of measuring the degree of heterogeneity in a given customer hierarchy. An overview of approaches for this task will be provided in the following section.

3.5. Measures of Heterogeneity in Customer Hierarchies

For any given customer hierarchy and the associated customer segmentation, it is helpful to measure to what extent the individual customer segments differ from one another with respect to profitability. If there is hardly any variation among the segments, profit-based management approaches are barely justified; but if there exist significant differences in profitability, management must pay attention to these differences. Accordingly, a measure of customer heterogeneity not only reveals to which extent an organization depends on a small set of customers for its profits (Mulhern, 1999), but also allows inter-company comparisons and gives insights whether a profitability-oriented customer management approach is worthwhile.

The subsequent discussion of heterogeneity measures has been structured as follows:

- First, in Section 3.5.1, it will be shown that there is a direct analogy between income inequality and the measurement of customer heterogeneity in customer hierarchies.
- Afterwards, in Section 3.5.2, the application of several key inequality measures in the context of heterogeneous customer hierarchies will be shown, proceeding from very simple to more advanced measures.
- A special class of measures, the so-called *Generalized Entropy* (GE) measures will be covered separately and in more detail in Section 3.5.3. Here, the focus will be placed on its most important representative, Theil's index T.
- In the final Section 3.5.4, a comparison and assessment of all the previously introduced measures will be provided.

3.5.1. Inequality and Heterogeneity Measurement

In customer hierarchies as introduced in Section 3.3, heterogeneity is due to different sizes (i.e. demands d_l) and different profitabilities (i.e. p_l) of the base customer segments, i.e. of the leaf nodes of the hierarchy.[37]

[37] Recall that the index l has been reserved to refer to these base customer segments. These base customer segments are characterized by a within-segment customer heterogeneity of zero.

This situation has an apparent similarity with the measurement of inequality in terms of income or wealth among the individuals in a given population. An inequality measure is typically defined as a 'scalar numerical representation of the interpersonal differences in income within a given population' (Cowell, 2011, p. 7). The important aspect of this definition is the aggregated, single-dimensional nature of the measure. It allows for unambiguous comparisons, either for the same population at different points in time, or between different populations.

The similarities between income inequality and customer heterogeneity in a customer hierarchy are striking: Consider the three-level hierarchy in Figure 3.6. The concepts on the left side correspond to the problem of measuring income inequality. By contrast, the concepts on the right side relate to heterogeneity in customer hierarchies. The key elements are the *income class* on the one hand and the *customer segment* on the other hand.

The number of income classes or customer segments is given by L. In income inequality, each income class $l = 1, \ldots, L$ consists of n_l individuals. Each individual has an income of y_i, $i = 1, \ldots, n_l$. Hence, the average income in class l is $\mu_l = \frac{1}{n_l} \sum_{i=1}^{n_l} y_i$.

The essential analogy is that each income class can be interpreted as a particular customer segment. The size of income class l, n_l, is equivalent to the total demand in the customer segment d_l. The average income per class μ_l coincides with the unit profitability p_l of a customer segment. While it is the *average individual* in income class l who earns an income of y_l, each individual unit of demand in customer segment l fetches a profit of p_l. This is due to the assumption of homogeneous leaf nodes in customer hierarchies. By contrast, income classes usually exhibit a certain remaining level of inequality. For example, they are often defined in terms of brackets with lower and upper annual incomes.

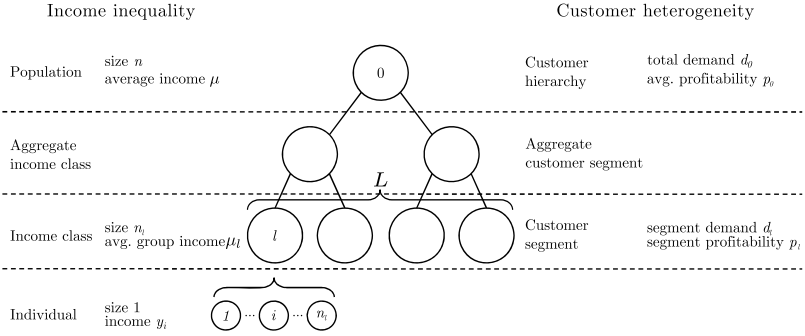


Figure 3.6. – Analogy between income inequality and customer heterogeneity measures

As illustrated in Figure 3.6, aggregate income classes and aggregate customer segments are given by the nodes in the upper two levels. At the top level of the hierarchy (node 0), the population of the n individuals has an average income μ. Similarly, the customer base has a total demand of d_0 and is characterized by an average unit profit p_0. Recall that d_0 and p_0 can be calculated by repetitive application of the summation formula (3.16) and of

the demand-weighted arithmetic mean (3.17), respectively. Alternatively, d_0 and p_0 may also be calculated directly by a central planner via

$$d_0 = \sum_{l=1}^{L} d_l, \qquad p_0 = \frac{\sum_{l=1}^{L} p_l \cdot d_l}{d_0}. \tag{3.38}$$

Sorting all individuals in ascending order of their individual incomes yields the *income distribution* of the population. In a similar manner, by sorting all base customer segments according to increasing unit profitability, the *profitability distribution* results. However, note an important difference:

- Since the income distribution is based on the incomes of the individuals in the population, it is in fact a discrete distribution. But provided that the number of individuals in the population is large, the income distribution can often be *approximated by a smooth*, i.e. *continuous curve*. The associated approximation error can usually be neglected in practice.

- By contrast, since the unit profits within each customer segment are identical, the profitability distribution is a *piece-wise linear function*.

The discrete yet approximately smooth income distribution at the level of individual incomes is represented graphically by the gray curve in Figure 3.7. After forming four income classes, the level of income inequality at the class level is represented by the black piece-wise linear curve. Note that the gray curve generally lies below the black curve, only touching each other at the limits of the income classes (i.e. at 0%, 25%, 50% and 100%, since equally sized income classes have been used in the example with $n_1 = n_2 = n_3 = n_4$). Using the analogies introduced in Figure 3.6, the black piece-wise linear curve can be interpreted as the profitability distribution among the base customer segments of the customer hierarchy. The main difference between the income inequality and the customer heterogeneity perspective is that all demand units in a particular customer segment l fetch the same unit profit p_l, i.e. there is no within-segment heterogeneity.

In the following sections, it will be shown how popular econometric measures of income inequality can be applied to the context of heterogeneous customer hierarchies by exploiting the above analogies. To improve the clarity of the presentation, the term *inequality measure* will only be used to refer to the measurement in the econometric context of income inequality. By contrast, the term *heterogeneity measure* will be reserved for customer hierarchies.

Next, a number of desirable properties of such measures will be summarized. While the following properties will be given with respect to inequality measures for an easier comparison with the literature, they can also be applied to characterize heterogeneity measures for customer hierarchies.

Assume a particular population A is given. Its inequality will be measured with a certain measure M, resulting in a particular value M_A.

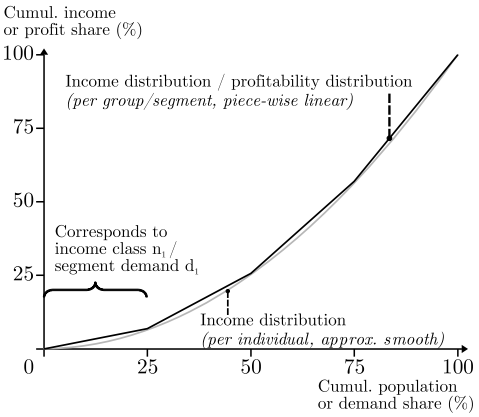


Figure 3.7. – Income and profitability distribution

Symmetry: An inequality measure is symmetric if its value is not altered under permutations of individuals, i.e. if the personality of the income earners does not affect the calculation of the metric (Bourguignon, 1979). Assume that population A consists of the individuals $a_1, a_2, \ldots, a_n$. Symmetry merely requires that $M(a_1, a_2, \ldots, a_n) = M(a_2, a_1, \ldots, a_n)$. This is usually fulfilled by most measures.

Population-Size Independence / Replication Invariance: In many situations, it is helpful to require that any proportional scaling in the population size shall not alter inequality (see Cowell, 2011, p. 63). Consider again the income distribution in population A with n individuals. Now assume that there is a second population A' which derives from A. A' simply consists of twice the number of individuals, i.e. $a_1, a_1, a_2, a_2, \ldots, a_n, a_n$. If $M_A = M_{A'}$ holds, i.e. if the inequality measure for the larger population is the same as for the smaller one, the measure is *population size independent*.

Scale Invariance: Consider two populations A and B which are identical except that all incomes in population B are scaled by a constant multiplier α compared to the incomes in A. An inequality measure possesses the property of being *invariant to multiplicative scale changes* if $M_B = M_A$ holds (e.g. see Cowell, 2011, p. 63). Now consider only population A with individuals $a_1, \ldots, a_n$. Assume that the income y_{a_i} of each individual a_i in A is raised by the same absolute amount β. Denote this resulting population as population C with $c_1, \ldots, c_n$. As a result, $y_{c_i} = y_{a_i} + \beta$. An inequality measure for which $M_C = M_A$ holds is said to be *invariant to additive scale changes*.

Sensitivity to Transfer Effects: In the context of inequality measures, an important property is related to the so-called *transfer effect*. Assume that in population A a small amount of income Δy is taken from a high-income individual a_i and given to a low-income

individual a_j (e.g. via the taxation system). Let the transferred amount be small enough so that the relative order of the two individuals is not changed, so $y_{a_i} - \Delta y > y_{a_j} + \Delta y$. Note that the average income in A will remain unaffected by this transfer. The population after this transfer will be referred to as A'.

The transfer has reduced the level of inequality in the population. This decrease should also be reflected in the inequality measure, so $M_A > M_{A'}$ should hold. This property is also referred to as the *Pigou-Dalton principle*, see Sen (1973, p. 27), or weak principle of transfers, see Cowell (2011, p. 62).

Many inequality measures which fulfill the Pigou-Dalton principle are sensitive to the position in the income distribution where such a transfer takes place. Consider two individuals a_k, a_l from population A with $y_{a_k} > y_{a_l}$ and assume that both individuals have higher incomes than individuals a_i and a_j, so $y_{a_k} > y_{a_l} > y_{a_i} > y_{a_j}$. Now transfer the *same absolute amount* of income Δy as before, but now from individual a_k to a_l. As in the previous transfer, the relative order of a_k and a_l will remain unaffected. Denote the population after this second transfer by A''. An inequality measure which fulfills $M_{A'} > M_{A''}$ is said to be more sensitive to transfers in the upper end of the income distribution. As will be shown in Section 3.5.4, popular inequality measures differ significantly in terms of this sensitivity. Choosing a particular inequality measure is therefore implicitly an expression of what the user perceives to be a societally desirable transfer.[38] In the context of measuring customer heterogeneity, it is desirable that the impact of such transfers shall only depend on the size of the transfer, but not on the position where it occurs. This special property will be referred to in the course of this thesis as the *constant transfer effect sensitivity*.

Additive Decomposability: Assume now that L mutually exclusive, collectively exhaustive income classes g_l with $l = 1, \ldots, L$ have been defined in population A. The inequality within each class g_l corresponds to M_{g_l}. For an application with hierarchically structured data, it is helpful if the overall inequality M_A can be expressed, using data at the income class level, as the sum of two components,

$$M_A = M_A^W + M_A^B. \tag{3.39}$$

The first component, within-group inequality M_A^W, captures the contribution to overall inequality which is introduced by the income inequality *within* each income class g_l. The second component, between-group inequality M_A^B, is the inequality *between* the L income classes, i.e. at an aggregate level. Therefore, M_A^B is independent of the actual income distribution within each class g_l. M_A^B is positive if the L classes have different average incomes μ_l. If the μ_l are all identical, this between-group inequality equals zero.

An inequality measure for which (3.39) holds is said to be *additively decomposable* if the within-component M_A^W can be expressed further as a weighted sum of the inequality contributions M_{g_l} of the L income classes g_l, i.e. if the following holds (Bourguignon (1979,

[38] For a more comprehensive discussion of the related concepts of inequality aversion and social welfare functions, see Cowell (2011, Ch. 3).

p. 905), Shorrocks (1980, p. 1370)):

$$M_A^W = \sum_{l=1}^{L} w_l \cdot M_{g_l}. \tag{3.40}$$

If the additional condition $\sum_{l=1}^{L} w_l = 1$ holds, the within-inequality component is a *true weighted average* of the L different sub-group inequality contributions. This property will be referred to in the following as being *strictly additively decomposable*. Shorrocks (1980) has shown that only two inequality measures exist which fulfill this strong property. They will be introduced later in Expressions (3.51) and (3.52).

3.5.2. Standard Heterogeneity Measures

In this and the following section, a number of heterogeneity measures will be presented for multi-stage customer hierarchies. All measures originate from the measurement of income inequality by observing the analogies between income inequality and customer heterogeneity as depicted in Figure 3.6. Recall that the overall demand in the customer hierarchy d_0 corresponds to the summation of all leaf node demands via repetitive application of (3.16) whereas the unit profit at the root node p_0 is the demand-weighted arithmetic mean calculated via (3.17). Alternatively, both figures can be calculated by a central planner using(3.38).

The desirable properties of inequality measures, which have been discussed in the previous section, can be used to also characterize heterogeneity measures. A comparison of the following measures and an application to an example hierarchy will be provided later in Section 3.5.4.

Relative Range: For a start, a very simplistic approach to measure heterogeneity consists of sorting all leaf nodes of the hierarchy in order of descending profitability. The relative range of profitability among the leaf nodes of a customer hierarchy will be defined as

$$RR = \frac{\max_l p_l - \min_l p_l}{p_0}, \tag{3.41}$$

i.e. by normalizing the absolute range of the unit profit values at the leaf nodes by the overall demand-weighted average unit profit in the customer hierarchy p_0 as defined in (3.38). A number of related measures can be obtained by considering the distance between symmetric percentile values instead of the extreme points. Since the relative range divides the distance between the extreme points by the average, this measure is invariant to multiplicative scale changes. However, two main disadvantages are apparent: First, no consideration is given to the different sizes of the customer segments in terms of demand d_l. While this issue does not arise in an econometric context where the relative range is usually calculated at the level of individuals (i.e. each individual equals one 'unit'), the leaf nodes of a customer hierarchy actually differ in terms of demand per customer segment. A second critical point is that the relative range and its related measures are

based on only two single (extreme) data points, i.e. the distribution of the remaining values in between is not considered.

Standard Deviation and Coefficient of Variation: A better measure can be obtained by comparing the unit profitability of *each* individual customer segment with the overall demand-weighted average profitability p_0 in the hierarchy, over all L segments. To incorporate the effect of different segment sizes, the contribution of each segment to overall heterogeneity will be weighted with its relative size $\frac{d_l}{d_0}$. The resulting inequality measure is the demand-weighted standard deviation σ, defined as

$$\sigma = \sqrt{\sum_{l=1}^{L} \frac{d_l}{d_0} (p_l - p_0)^2}. \tag{3.42}$$

While σ is invariant to additive scale changes, it is affected by multiplicative scale changes. In most practical situations, however, multiplicative scale invariance is the more important property, e.g. because the heterogeneity measure should not change when measuring customer heterogeneity over time in the presence of price inflation. This deficiency can be cured by normalizing σ, i.e. by dividing by the average unit profit p_0. This leads to the coefficient of variation $CV = \frac{\sigma}{p_0}$. CV is invariant to multiplicative, but no longer to additive scale changes.[39]

Lorenz Curve, Gini Coefficient and Stobachoff Index: The following heterogeneity measures have been obtained from graphical representations of the profitability distribution. As observed by Storbacka (1994), histograms of profitability dispersions in practice tend to be heavily skewed. Therefore, ordered distributions are better suited for a graphical representation of the heterogeneity of a customer base in terms of profitability. This is already a long-standing practice in the context of income distributions. The so-called Lorenz curve (Lorenz, 1905) is a popular graphical concept which is used in the presence of skewed distributions. A Lorenz curve $\eta(\phi)$ plots cumulative income share η over the cumulative population share ϕ. To determine ϕ, the population has been sorted in order of ascending individual income levels (see Figure 3.8).[40]

Any Lorenz curve is convex and fulfills the following properties (Chotikapanich, 1993, p. 129):

$$\eta(0) = 0, \quad \eta(1) = 1, \quad 0 < \eta(\phi) < \phi < 1, \quad \frac{\partial \eta}{\partial \phi} \geq 0, \quad \frac{\partial^2 \eta}{\partial \phi^2} > 0. \tag{3.43}$$

The Lorenz curve shows the proportion of total societal income which is earned per share of the population. In a homogeneous population where all individuals receive identical

[39] Furthermore, the CV is an instructive example of a heterogeneity measure which is very sensitive to transfers at the higher end of the profitability distribution. More details on this property will be given in Section 3.5.4.

[40] Technically, this definition implies that the resulting Lorenz curve is a discrete rather than a continuous representation of income inequality. However, in most practical applications, the number of individuals is very large which usually justifies the continuous approximation.

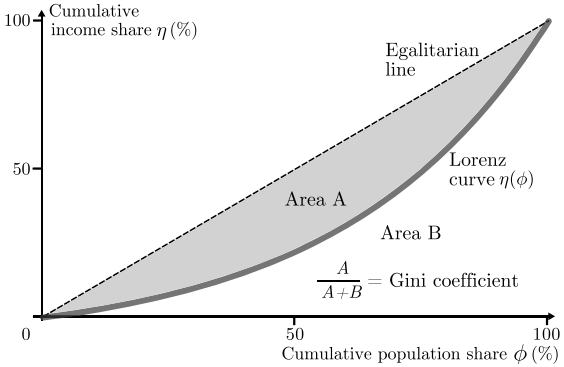


Figure 3.8. – Lorenz curve depicting the income inequality in a population

incomes $y_i = y$ for all $i = 1, \ldots, n$, the Lorenz curve corresponds to a straight line from the origin $(0; 0)$ to the point $(1; 1)$. This line is the *egalitarian line*. In a population with income inequality, the deviation of the Lorenz curve from the egalitarian line depicts the level of income concentration at the high end of the population. To transfer this graphical representation into a scalar representation for the aggregate level of income inequality, a number of summary statistics have been proposed which derive directly from the Lorenz curve (see Arnold, 2008, p. 18):

- *Kakwani index*: Length of the Lorenz curve (values are between $\sqrt{2}$ and 2).
- *Pietra index*: Maximum vertical distance between the Lorenz curve and the egalitarian line (between 0 and 1).
- *Gini coefficient*: Twice the size of area A (see Figure 3.8) between the Lorenz curve and the egalitarian line (between 0 and 1).

The Gini coefficient G is the most popular of these summary statistics. The geometric definition of G is often stated in the form $G = A/(A + B)$, where B corresponds to the area below the Lorenz curve in Figure 3.8. Realizing that $A + B = 1/2$, this gives $G = A/(A+B) = 2A$. Using $A = 1/2 - B$ in the previous expression yields another form of the Gini coefficient,

$$G = 2A = 2(\frac{1}{2} - B) = 1 - 2B. \tag{3.44}$$

This last form can be used if the continuous Lorenz curve $\eta(\phi)$ of a particular income distribution is given. G can then be calculated via (see Cowell, 2011, p. 157)

$$G = 1 - 2\int_0^1 \eta(\phi)d\phi. \tag{3.45}$$

A simpler calculation of G is possible if discrete values for the individual incomes are available (cf. Cowell, 2011, p. 155). Then,

$$G = \frac{1}{2n^2\mu} \sum_{i=1}^{n} \sum_{j=1}^{n} |y_i - y_j|. \tag{3.46}$$

Here, recall that μ represents the average income in the entire population (see Figure 3.6). If all individuals per income class $l = 1, \ldots, L$ have identical incomes (i.e. if the income classes are homogeneous), Equation (3.46) can also be expressed directly at the level of the L income classes using the average income group incomes μ_l:

$$G = \frac{1}{2n^2\mu} \sum_{l=1}^{L} \left(n_l \cdot \sum_{l'=1}^{L} n_{l'} \cdot |\mu_{l'} - \mu_l| \right). \tag{3.47}$$

This expression can be transferred directly to the measurement of customer heterogeneity by observing the established analogies between income inequality and customer heterogeneity (see Figure 3.6, and recall that there is no customer heterogeneity within each base customer segment by definition):

- The size n_l of income class l in terms of individuals corresponds to the total demand in leaf node (or base customer segment) l.
- The average income μ_l of income class l matches the unit profit p_l of base customer segment l.
- The overall population size n is equivalent to the total demand in the customer hierarchy d_0 (summed via repetitive application of (3.16) or using the left part of (3.38)).
- The average income in the population μ corresponds to the demand-weighted average unit profit in the customer hierarchy p_0 (calculated iteratively via (3.17) or using the right part of (3.38)).

Hence, the Gini coefficient measuring the level of customer heterogeneity at the level of the L *base customer segments* equals

$$G = \frac{1}{2d_0^2 p_0} \sum_{l=1}^{L} \left(d_l \cdot \sum_{l'=1}^{L} d_{l'} \cdot |p_{l'} - p_l| \right). \tag{3.48}$$

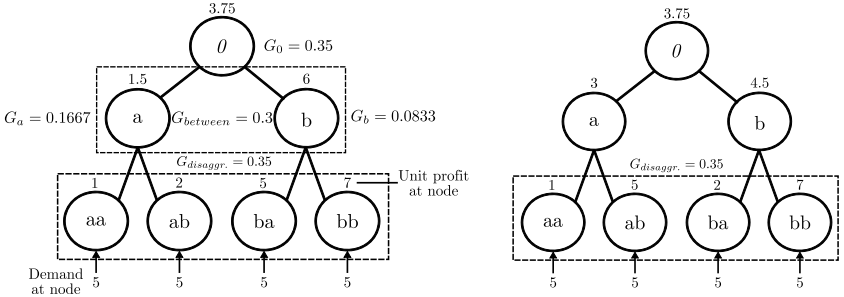
Unfortunately, a decomposed calculation of G for profitability data at the level of *aggregate customer segments* is only possible if the constituent base customer segments of the customer hierarchy do not overlap. This means that the profitability distribution at the aggregate customer segment level has to equal the profitability distribution at the level of the base customer segments. If this property holds, the overall Gini coefficient G_0 at the root node 0 can be calculated via a weighted sum of the Gini coefficients G_k within

each aggregate segment k plus an additional component which captures the heterogeneity between the K aggregate segments. According to Cowell (2011, p. 165), this relationship is given by:

$$G_0 = \sum_{k=1}^{K} \frac{d_k^2 \cdot p_k}{d_0^2 \cdot p_0} G_k + G_{between}. \tag{3.49}$$

An example of a customer hierarchy without such an overlap is depicted in Figure 3.9a. When sorting the leaf nodes in order of increasing profitability, the same order also results at the level of the aggregate customer segments, as the nodes aa and ab in the aggregate segment a both have a strictly smaller profitability than the nodes in the aggregate segment b. For the given data, the Gini coefficient among the leaf nodes (i.e. base customer segments) corresponds to 0.35, via (3.48). The Gini coefficients at the level of the two aggregate nodes $k = a$ and $k = b$ can be calculated in a similar manner and correspond to $G_a = 0.1667$ and $G_b = 0.0833$. The between-component $G_{between}$ can be calculated by assuming that each demand unit in the aggregate segments a and b fetches a unit profit of p_a and p_b, respectively. With $G_{between}$ corresponding to 0.3 (in an analog application of (3.48), the overall value of the Gini coefficient in the example amounts to $G_0 = 0.35$. This decomposed calculation via (3.49) has led to the same result as a direct calculation across all leaf nodes by a central planner via (3.48).

Such a decomposed calculation is no longer possible if there is an overlap among the leaf node segments, as indicated in Figure 3.9b. An application of (3.49) fails as it leads to a different value than the direct calculation via (3.48). As a consequence, the Gini coefficient is not decomposable in the general case and hence not suitable for application in multi-stage customer hierarchies.



(a) Non-overlapping base customer segments

(b) Overlapping base customer segments

Figure 3.9. – Decomposability of the Gini coefficient: Example hierarchies

In the marketing literature, concepts closely related to the Lorenz curve and the Gini coefficient are used.[41] The *Stobachoff curve* (Storbacka, 1994, 1997) is obtained by first

[41] Bartezzaghi et al. (1999) gave an application of the Gini coefficient in a demand planning and forecasting setting. They use G to measure the level of heterogeneity in terms of customer order size.

ordering all customers of a firm from highest to lowest absolute profitability and then plotting cumulative profits against cumulative customers, as shown in Figure 3.10. Essentially, this graphical representation corresponds to a Lorenz curve which has been flipped across the egalitarian line. As some customers can be associated with negative profits (i.e. the firm is losing money in making business with them), the Stobachoff curve may rise above 100% and then fall back to reach 100% again once all customers have been accounted for.[42]

The maximum of a Stobachoff curve is particularly revealing. All customers to the left of this point generate positive profits. These profits are required, at least in part, to subsidize all other customers who lie to the right of this point as those are associated with negative profits, i.e. losses. A summary statistic for this distribution is the Stobachoff index (see Storbacka, 1994, p. 142), which is closely related to the Gini coefficient. Denote the size of the area between the Stobachoff curve and the egalitarian line with A and the entire area between the Stobachoff curve and the horizontal axis with $F = A + C$. The Stobachoff index ST is simply the ratio between A and F, i.e. $ST = \frac{A}{F} = \frac{A}{A+1/2} = \frac{2A}{2A+1} = \frac{G}{G+1}$. The last step follows from the definition of the Gini coefficient. Hence the relationship between the Gini coefficient and the Stobachoff index is $G = \frac{ST}{1-ST}$.

Essentially, the Stobachoff index measures the deviation of the profitability distribution of a given customer base from an 'ideal', i.e. homogeneous, customer base (Storbacka, 1994, p. 143). A value of zero implies all that profitability is equally distributed and that all customers have a positive profit contribution. This is usually a desirable situation for many companies. Positive values of the Stobachoff index imply an unequal profitability distribution, with some customers even having a negative profitability (i.e. they lead to losses for the company). Overall, the Stobachoff index is not particularly helpful in the case of multi-stage customer hierarchies. Like the Gini coefficient, it is not additively decomposable in the general case.

3.5.3. Generalized Entropy Inequality Measures and Theil's Index

All inequality measures presented so far do not fulfill the property of strict additive decomposability. However, in case of multi-level customer hierarchies, it is desirable to calculate a measure M_0 of the overall level of customer heterogeneity in an iterative fashion, only based on local, aggregate data. Bourguignon (1979) and Shorrocks (1980) have proven that the so-called class of *Generalized Entropy* measures GE^c with parameter c is the only class of inequality measures whose members are additively decomposable. Only two of them, with parameters $c = 0$ and $c = 1$, are also strictly additively decomposable.

In the case of inequality measurement involving a population of $i = 1, \ldots, n$ individuals with incomes y_i and an average income of μ, the general form of the GE measures is given by (cf. Shorrocks, 1984, p. 1370):

$$GE^I(c) = \frac{1}{c(c-1)} \frac{1}{n} \sum_{i=1}^{n} \left[\left(\frac{y_i}{\mu} \right)^c - 1 \right], \quad c \neq 0 \text{ and } c \neq 1. \tag{3.50}$$

[42] The same phenomenon has been referred to in Section 3.3.1 with the 225–20 rule.

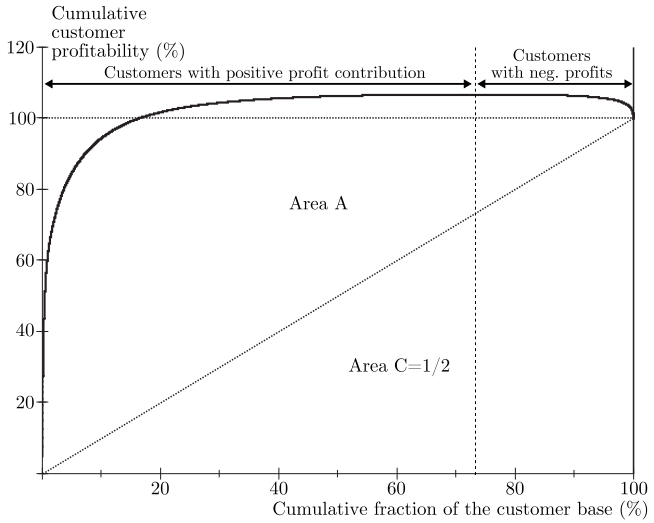


Figure 3.10. – Stobachoff curve (van Raaij et al., 2003, Fig. 3)

Note that the superscript I will be used to indicate that the measure refers to income inequality rather than customer heterogeneity. Taking limits via L'Hôpital's rule yields special forms for the cases $c = 0$ and $c = 1$. These correspond to the original entropy-based indices which have initially been suggested by Theil (1967):

$$c = 0: \quad GE^I(0) = \frac{1}{n} \sum_{i=1}^{n} \ln \frac{\mu}{y_i} \tag{3.51}$$

$$c = 1: \quad GE^I(1) = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\mu} \ln \frac{y_i}{\mu} \tag{3.52}$$

Now consider a situation where income data is only available at the level of $l = 1, \ldots, L$ income classes, each consisting of n_l individuals, with $n_1 + \cdots + n_l + \cdots + n_L = n$. As introduced with the help of Figure 3.6, the average income in class l is $\mu_l = \frac{1}{n_l} \sum_{i=1}^{n_l} y_i$. For the average income of the entire population, the following holds:

$$\mu = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{l=1}^{L} \sum_{i=1}^{n_l} y_i}{\sum_{l=1}^{L} n_l} = \frac{\sum_{l=1}^{L} \mu_l \cdot n_l}{\sum_{l=1}^{L} n_l}. \tag{3.53}$$

Essentially, the right part of (3.53) implies that μ can be interpreted as the population-weighted arithmetic mean income at the aggregate level of income classes. Measuring income inequality at this aggregate level will ignore the income inequality contributions within each income class, i.e. it will be assumed that $y_i = \mu_l$ for all i in income class l. In

this case, expression (3.50) becomes

$$GE^I(c) = \frac{1}{c(c-1)} \frac{1}{n} \sum_{l=1}^{L} \left(n_l \cdot \left[\left(\frac{\mu_l}{\mu} \right)^c - 1 \right] \right), \quad c \neq 0 \text{ and } c \neq 1 \tag{3.54}$$

while (3.51) and (3.52) turn into

$$GE^I(0) = \sum_{l=1}^{L} \frac{n_l}{n} \cdot \ln \frac{\mu}{\mu_l} \tag{3.55}$$

$$GE^I(1) = \sum_{l=1}^{L} \frac{n_l}{n} \cdot \frac{\mu_l}{\mu} \ln \frac{\mu_l}{\mu}. \tag{3.56}$$

Note that these forms of $GE^I(c)$ ignore any inequality contributions *within* each population group n_l.

However, these representations can be applied directly to the case of a customer hierarchy with L base customer segments by observing the analogies introduced with the help of Figure 3.6. In particular, only the substitutions $n \to d_0$, $\mu_l \to p_l$ and $\mu \to p_0$ are required. Note that p_0 corresponds to the demand-weighted arithmetic mean of the unit profits (see also (3.38)), in the same manner as μ can be seen as the population-weighted arithmetic mean income (see above). These substitutions lead to:

$$GE(c) = \frac{1}{c(c-1)} \frac{1}{d_0} \sum_{l=1}^{L} d_l \cdot \left[\left(\frac{p_l}{p_0} \right)^c - 1 \right], \quad c \neq 0 \text{ and } c \neq 1. \tag{3.57}$$

The cases $c = 0$ and $c = 1$ become

$$GE(0) = MLD = \frac{1}{d_0} \sum_{l=1}^{L} d_l \cdot \ln \frac{p_0}{p_l} \tag{3.58}$$

$$GE(1) = T = \frac{1}{d_0} \sum_{l=1}^{L} \frac{p_l \cdot d_l}{p_0} \ln \frac{p_l}{p_0} \tag{3.59}$$

These latter two forms also allow for an intuitive interpretation of the weight in front of each logarithmic term:

- Equation (3.58), also known as the *mean logarithmic deviation* (MLD), weighs all customer segments l in proportion to their demand, i.e. d_l.

- Equation (3.59) is commonly known as *Theil's index* T and weighs all customer segments l with their demand-weighted arithmetic unit profit $\frac{p_l \cdot d_l}{p_0}$. Employing the Theil index to measure customer heterogeneity is thus implicitly linked to the use of the demand-weighted arithmetic mean as the aggregation operator for the unit profits per customer segment.

Theil's Index

The remainder of this discussion will focus on Theil's index T. In the following, the background of Theil's index, its general properties as well as its important strict additive decomposition property in hierarchies will be investigated in more detail.

Historical Background: In deriving the inequality index T, Theil (1967) relied heavily on Claude Shannon's famous theory of information (see Shannon, 1948): An almost certain event has only little (additional) information value whereas an event with low probability of occurrence is associated with a particularly high information content. Shannon required that the combined information content of independent events should be additive. This means that the information content of a number of independent events must equal the sum of the individual information content contributions (Conceição and Galbraith, 2000, p. 62). To transform the probability of an event A, denoted by $p(A)$, into the information content or Shannon measure $S(A)$, Shannon suggested using the logarithm of the inverse of the probability of occurrence, so $S(A) = \ln \frac{1}{p(A)}$.

As a consequence, no information is obtained from sure events: With $p(A) = 1$, $S(A) = \ln 1 = 0$ holds. By contrast, the highest information content is associated with a situation where n outcomes $A_1, \ldots, A_n$ are possible and equally likely. Hence, the probability of each event A_i equals $p(A_i) = \frac{1}{n}$. In this case, each event A_i is associated with a (maximum) Shannon measure of $S(A_i) = \ln \frac{1}{1/n} = \ln n$. This situation is characterized by maximum disorder or *maximum entropy*, hence the name for the class of inequality measures defined by Equation (3.57).

General Properties: Theil's index has been shown to be monotonically increasing, differentiable and invariant to multiplicative scale changes (e.g. see Bourguignon (1979) or Shorrocks (1984)). Like many other measures, it fulfills the Pigou-Dalton principle, but also has an invariant transfer effect sensitivity. The change in T as a reaction to a given marginal transfer of profitability from one node to any other node always has the same magnitude, independent of the position in the profitability distribution where this transfer takes place (Cowell, 2011, p. 155). This property is not fulfilled by any other (strictly) additively decomposable heterogeneity measure. In Section 3.5.4, this will be illustrated with a numerical example.

The following lemma states the extreme values which will be taken by T in a customer hierarchy, i.e. the minimum value of T for a customer hierarchy without any customer heterogeneity and the maximum value of T if customer profitability is distributed in the most unequal manner among the base customer segments.

Lemma 1. *Assuming each base customer segment l in the customer hierarchy has a demand of at least $d_l \geq 1$ for $l = 1, \ldots, L$ and total demand equals $d_0 = \sum_{l=1}^{L} d_l$, the value of T will lie in the interval $[0; \ln(d_0)]$.*

Proof. A lower bound of T can be derived with the help of Jensen's inequality: Let f be a convex function on the open interval $I \in \mathbb{R}$. For $x_1, x_2, \ldots, x_L \in I$ and factors

$\lambda_1, \lambda_2, \ldots, \lambda_L \in [0;1]$ such that $\sum_{l=1}^{L} \lambda_l = 1$, it holds that (see Yeh, 2006)

$$\sum_{l=1}^{L} (\lambda_l \cdot f(x_l)) \geq f\left(\sum_{l=1}^{L} (\lambda_l \cdot x_l)\right). \tag{3.60}$$

Now, note that $f = -\ln x$ is convex, since its second derivative $\frac{1}{x^2}$ is strictly positive for all x. The following inequality can be derived from Eq. (3.59) by applying (3.60):

$$T = \sum_{l=1}^{L} \frac{p_l \cdot d_l}{p_0 \cdot d_0} \ln \frac{p_l}{p_0} = \sum_{l=1}^{L} \frac{p_l \cdot d_l}{p_0 \cdot d_0} \cdot \left(-\ln \frac{p_0}{p_l}\right) \tag{3.61}$$

$$\geq -\ln\left(\sum_{l=1}^{L} \frac{p_l \cdot d_l}{p_0 \cdot d_0} \frac{p_0}{p_l}\right) = -\ln\left(\sum_{l=1}^{L} \frac{d_l}{d_0}\right) \tag{3.62}$$

$$= -\ln\left(\frac{d_0}{d_0}\right) = -\ln 1 = 0. \tag{3.63}$$

This confirms that the values of T are non-negative. The case of the maximum value of T is more complicated. A maximum concentration of profitability implies that this (positive) profit contribution must be attributable to a single base customer segment whose demand is as small as possible. The profit contribution from all other base customer segments must be as small as possible, approaching zero in the limit.

Therefore, assume that the first $1, \ldots, L-1$ base customer segments have positive demands $d_l \geq 1$ and are associated in the limit with a unit profit $p_l \to 0$, for all $l = 1, \ldots, L-1$. The last base customer segment L is assumed to have the minimum demand of $d_L = 1$ and is associated with a strictly positive unit profit of $p_L > 0$. Initially, recall from Expression (3.38) that the aggregate demand and aggregate unit profit in this customer hierarchy can be calculated over the leaf nodes and use the above assumptions to obtain in the limit

$$d_0 = \sum_{l=1}^{L} d_l, \qquad p_0 = \frac{\sum_{l=1}^{L} p_l \cdot d_l}{d_0} = \frac{p_L \cdot d_L}{d_0}. \tag{3.64}$$

Now consider the first $L-1$ terms of the sum in (3.59), i.e. the expression $\frac{p_L \cdot d_l}{p_0 \cdot d_0} \cdot \ln\left(\frac{p_L}{p_0}\right)$ for $l = 1, \ldots, L-1$. In the limit, with $p_l \to 0$, one obtains

$$\begin{aligned}
\lim_{p_l \to 0} \left(\frac{p_l \cdot d_l}{p_0 \cdot d_0} \ln\left(\frac{p_l}{p_0}\right)\right) &= \frac{d_l}{p_0 \cdot d_0} \cdot \lim_{p_l \to 0} \left(\frac{\ln\left(\frac{p_L}{p_0}\right)}{\frac{1}{p_l}}\right) \\
&= \frac{d_l}{p_0 \cdot d_0} \cdot \lim_{p_l \to 0} \left(\frac{\frac{p_0}{p_l}}{\frac{-1}{p_l^2}}\right) \\
&= \frac{d_l}{p_0 \cdot d_0} \cdot \lim_{p_l \to 0} (-p_0 \cdot p_l) = 0, \qquad l = 1, \ldots, L-1.
\end{aligned} \tag{3.65}$$

In the above formula, L'Hôpital's rule has been applied to move from the first to the second row. Additionally, the case $l = L$ must be investigated. The corresponding term in (3.59) can be simplified by using the expression for p_0 from (3.64):

$$\frac{p_L \cdot d_L}{p_0 \cdot d_0} \cdot \ln\left(\frac{p_L}{p_0}\right) = \frac{p_L \cdot d_L \cdot d_0}{p_L \cdot d_L \cdot d_0} \cdot \ln\left(\frac{p_L \cdot d_0}{p_L \cdot d_L}\right) = 1 \cdot \ln\left(\frac{d_0}{d_L}\right). \tag{3.66}$$

Now use (3.65), (3.66) and the assumption that $d_L = 1$ in (3.59). This yields $T = \ln(d_0)$ for the maximum value of customer heterogeneity if total profits are concentrated in a single base customer segment as described above. □

Note that in practice the values of T are usually significantly below this upper bound, as will be illustrated in Section 3.5.4 and in Section 4.5.

Hierarchical Decomposition: An attractive property of the Theil index in income inequality measurement is that it is strictly additively decomposable, i.e. that the index allows for a decentral and iterative calculation. It will be shown in this paragraph that this property also holds when applied in the context of heterogeneous multi-stage customer hierarchies.

To facilitate the subsequent presentation, the symbol T_i will be used in the following when referring to the level of heterogeneity measured by the Theil index *downstream* of a particular node i in a customer hierarchy. In other words, T_i indicates the level of customer heterogeneity among all leaf nodes which are either direct or indirect successor nodes of (intermediate) node i. As a consequence, T_0 will denote the level of heterogeneity in the entire customer hierarchy, i.e. measured from the root node.

With the help of Figure 3.11—which is similar to Figure 3.6—it can be seen that the calculation of T_0 via (3.59) occurs across the L base customer segments of the hierarchy. The objective now consists of finding an expression which permits calculating T_0 in an iterative manner using data at the intermediate level of the K aggregate customer segments in Figure 3.11. Each aggregate customer segment k consists of n_k base customer segments, with $\sum_{k=1}^{K} n_k = L$.

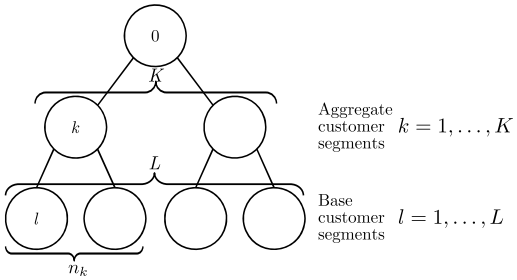


Figure 3.11. – Calculation of the Theil index at the level of aggregate customer segments

Lemma 2. *The Theil index T_0 in a customer hierarchy can be computed in an iterative, decentral manner using*

$$T_0 = \sum_{k=1}^{K} \frac{p_k \cdot d_k}{p_0 \cdot d_0} \ln \frac{p_k}{p_0} + \sum_{k=1}^{K} \frac{p_k \cdot d_k}{p_0 \cdot d_0} T_k. \tag{3.67}$$

T_k corresponds to the level of customer heterogeneity in the subtree below each aggregate node k and is given by

$$T_k = \sum_{l \in \mathcal{D}_k} \frac{p_l \cdot d_l}{p_k \cdot d_k} \ln \frac{p_l}{p_k} \tag{3.68}$$

Proof. In the econometric literature, the "self-similar" (Conceição, 2001) nature of the Theil index is used frequently. The subsequent argumentation only exploits the established analogies between income inequality and customer heterogeneity. An alternate form of the proof which rather departs directly from the basic expression (3.59) for the calculation of T_0 over all leaf nodes in the customer hierarchy will be provided in Appendix A.1.

For the short form of the proof, recall the original definition of $T^I = GE^I(1)$ from (3.52) in the context of inequality measurement and the notation introduced with the help of Figure 3.6. In Equation (3.56), a representation of T^I based on aggregate data at the level of the L income classes was introduced. However, as stated before, this form ignored the within-component of income inequality which might exist within each income class. To account for this component, only an additional term must be added. Following Conceição and Galbraith (2000, p. 63), the full level of income inequality at the level of the L customer segments is given by

$$T^I = \sum_{l=1}^{L} \frac{n_l}{n} \cdot \frac{\mu_l}{\mu} T_l^I + \sum_{l=1}^{L} \frac{n_l}{n} \cdot \frac{\mu_l}{\mu} \ln \frac{\mu_l}{\mu}. \tag{3.69}$$

Here, T_l^I is the level of income inequality within each income class l. Essentially, the first sum in (3.69) corresponds to the within-component of inequality while the second term describes the between-component. This representation also holds at higher hierarchical levels, i.e.

$$T^I = \sum_{k=1}^{K} \frac{n_k}{n} \cdot \frac{\mu_k}{\mu} T_k^I + \sum_{k=1}^{K} \frac{n_k}{n} \cdot \frac{\mu_k}{\mu} \ln \frac{\mu_k}{\mu}. \tag{3.70}$$

Observing the analogies between income inequality and customer heterogeneity and making the substitutions $n_k \to d_k$, $n \to d_0$, $\mu_k \to p_k$, $\mu \to p_0$, $T^I \to T_0$ and $T_k^I \to T_k$ in (3.70) leads to (3.67). □

Equation (3.68) can also be derived directly based on the basic expression (3.59) for the calculation of T_0 over all leaf nodes. This alternative derivation is given in Appendix A.1. It has the advantage of directly illustrating that the additively decomposability property of T follows directly from this basic form of the heterogeneity measure. Note that (3.67)

can be expressed as

$$T_0 = T_0^B + T_0^W, \qquad \text{with } T_0^B = \sum_{k=1}^{K} \frac{p_k \cdot d_k}{p_0 \cdot d_0} \ln \frac{p_k}{p_0} \quad \text{and } T_0^W = \sum_{k=1}^{K} \frac{p_k \cdot d_k}{p_0 \cdot d_0} T_k \tag{3.71}$$

$$= T_0^B + \sum_{k=1}^{K} w_k \cdot T_k, \qquad \text{with } w_k = \frac{p_k \cdot d_k}{p_0 \cdot d_0}. \tag{3.72}$$

The above two expressions contain an important insight for hierarchical data: In (3.71), T_0^B corresponds to the heterogeneity *between* the K aggregate customer segments whereas T_0^W is the level of heterogeneity *within* all K aggregate customer segments. Comparing Equation (3.72) with the definition of additive decomposability from (3.40) illustrates that the Theil index in multi-stage customer hierarchies indeed fulfills this property. Moreover, the following lemma holds:

Lemma 3. *The Theil index in the form of* (3.67) *is strictly additively decomposable, i.e. the within-heterogeneity T_0^W is a true weighted average of the individual heterogeneity contributions T_k within each aggregate segment.*

Proof. It suffices to show that the sum of the weights w_k in (3.72) equals unity. First, apply the definitions of p_0 and d_0 as introduced in Expression (3.38). This leads to

$$\sum_{k=1}^{K} w_k = \sum_{k=1}^{K} \frac{p_k \cdot d_k}{p_0 \cdot d_0} = \sum_{k=1}^{K} \frac{p_k \cdot d_k}{\frac{\sum_{l=1}^{L} p_l \cdot d_l}{d_0} \cdot d_0} = \frac{\sum_{k=1}^{K} p_k \cdot d_k}{\sum_{l=1}^{L} p_l \cdot d_l}$$

$$= \frac{\sum_{k=1}^{K} p_k \cdot d_k}{\sum_{k=1}^{K} \left(\sum_{l \in \mathcal{D}_k} p_l \cdot d_l \right)} \tag{3.73}$$

$$= \frac{\sum_{k=1}^{K} p_k \cdot d_k}{\sum_{k=1}^{K} p_k \cdot d_k} \tag{3.74}$$

$$= 1 \tag{3.75}$$

Step (3.73) follows by realizing with the help of Figure 3.11 that the sum over all leaf nodes $\sum_{l=1}^{L}$ can also be written as two nested sums $\sum_{k=1}^{K} \sum_{l \in \mathcal{D}_k}$. Realizing that $\sum_{l \in \mathcal{D}_k} p_l \cdot d_l = p_k \cdot d_k$ then leads to (3.74). □

While all measures of the GE class can be decomposed in a form which corresponds to (3.72), the sum of the weights usually does not equal unity, i.e. the strict criterion of additive decomposability is not fulfilled. T and MLD are the only heterogeneity measures with this property (see Bourguignon, 1979, p. 916). An implication of this property is that for T and MLD the within- and between-group heterogeneity components are fully independent of each other, allowing for a full decomposition of overall customer heterogeneity. For all other members of the GE class, the remaining difference between the sum of the weights and one, i.e. $1 - \sum_{k=1}^{K} w_k$, is proportional to the between-group heterogeneity term. This means that the decomposition coefficients w_k of the within-group

inequality are no longer independent of the between-group inequality (see Shorrocks, 1980, p. 624).

While the above argumentation has focused on the decentral and iterative calculation of T_0 at the root node of the customer hierarchy, similar arguments can be made for any other node i in the customer hierarchy. For later reference, the Theil index measuring the level of customer heterogeneity in the subtree below a particular node i can be computed in an iterative, decentral manner using

$$T_i = \begin{cases} 0, & \text{if } i \in \mathcal{L}, \\ \sum\limits_{k \in \mathcal{D}_i} \frac{p_k \cdot d_k}{p_i \cdot d_i} \ln \frac{p_k}{p_i} + \sum\limits_{k \in \mathcal{D}_i} \frac{p_k \cdot d_k}{p_i \cdot d_i} T_k, & \text{else.} \end{cases} \tag{3.76}$$

In the above expression, T_k can be determined in the same manner at the next lower level. For a central planner, this iterative calculation is not necessary. She can calculate the level of heterogeneity in the subtree below node i directly over all leaf nodes of this subtree. Use the symbol $\mathcal{L}_i$ to refer to all direct and indirect leaf nodes below node i, i.e. the group of all leaf nodes which are direct or indirect descendants of intermediate node i. Then a centralized calculation of T can be performed using

$$T_i = \sum_{l \in \mathcal{L}_i} \frac{p_l \cdot d_l}{p_i \cdot d_i} \ln \frac{p_l}{p_i} \quad \forall i \notin \mathcal{L}. \tag{3.77}$$

Note that if i is a leaf node, $\mathcal{L}_i = \{i\}$. With $d_i > 0$ and $p_i > 0$, the term at the right hand side of Equation (3.77) corresponds to

$$T_i = \frac{p_i \cdot d_i}{p_i \cdot d_i} \ln \frac{p_i}{p_i} = 0 \quad \forall i \in \mathcal{L}. \tag{3.78}$$

This reflects the prior assumption that all leaf nodes are homogeneous where all demand units contribute identical unit profits.
The possibility to calculate Theil's index in a decentralized, i.e. distributed manner makes this measure an excellent candidate to aggregate data also within other hierarchical multi-level structures. The main applications of Theil's index can naturally be found in the domain of economic inequality measurement. Conceição et al. (2001) studied wage inequality at several levels of the Standard Industrial Classification (SIC) and illustrated the information gain from calculating the inequality measure using more disaggregated data, i.e. when considering more of the within-subgroup inequality.

Recently, the Theil index has also been applied successfully to determine aggregate measures in software engineering: Serebrenik and van den Brand (2010) studied the number of lines of source code in large software packages. They used Theil's index to represent this metric at different levels of aggregation and also tracked its evolution over time by comparing the development stages of software systems. Furthermore, Theil's decomposition property was employed to test different explanations for the observed overall level of heterogeneity in terms of lines of code, e.g. differences in the package type, implementa-

tion language or software maintainer. In a related study, Vasilescu et al. (2011) discussed and tested the applicability of a broader range of heterogeneity measures, including the Theil index, to aggregate software metrics.

3.5.4. Comparison of Heterogeneity Measures

The purpose of this final section is to highlight the superiority of Theil's index to measure customer heterogeneity in comparison to the other presented measures.

As a first step, Table 3.4 gives an overview of the heterogeneity measures which have been presented in the previous sections. The following comparison will cover the relative range (RR), the standard deviation (σ), the coefficient of variation (CV), the Gini coefficient (G) as well as the mean logarithmic deviation (MLD) and Theil's index (T) as representatives for the general class of GE measures. Table 3.4 indicates which of the key properties from Section 3.5.1 apply to each measure. Regarding the sensitivity to the transfer effect, it will only be stated whether the weak principle of transfers—or Pigou-Dalton principle—is fulfilled.[43] Furthermore, the table indicates if a measure is additively decomposable and whether it allows for a full decomposition of aggregate heterogeneity, i.e. if the weights sum to unity (strict additive decomposability).

Property	**Heterogeneity measures**					
	RR	σ	CV	G	MLD	T
Symmetry	x	x	x	x	x	x
Population size independence	x	x	x	x	x	x
Additive scale invariance		x				
Multiplicative scale invariance			x	x	x	x
Pigou-Dalton principle		x	x	x	x	x
Additive decomposability			(x)[a]		x	x
Strict additive decomposability					x	x

[a] Fulfilled for the associated $GE(c)$ measure with $c = 2$ which corresponds to $1/2$ of the squared coefficient of variation.

Table 3.4. – Properties of key heterogeneity measures for customer hierarchies

As this summary suggests, both the MLD and T appear to be favorable measures, particularly for multi-level customer hierarchies where a full decomposition of heterogeneity is needed. While all measures of the class GE^c are additively decomposable, only MLD and T fulfill the stricter requirement that the within-component of heterogeneity at an aggregate level is a true weighted average of the individual heterogeneity contributions of the lower level (see Bourguignon, 1979, p. 918). In fact, T is the better choice among

[43] I.e. whether a small transfer of profitability from the high end of the profitability distribution to the lower end reduces overall heterogeneity.

these two measures as it is insensitive with respect to the position of transfer effects. In that sense it is neutral as to what should be done to reduce heterogeneity in a customer hierarchy, i.e. certain transfers in the profitability distribution are not 'better' suited to reduce heterogeneity than others.

Before illustrating this argument with an example, it is helpful to consider the numerical values which result from applying different heterogeneity measures to a particular customer hierarchy. Consider the three-level hierarchy given in Figure 3.12. At the lowest level, the leaf nodes of the hierarchy represent four customer segments $l = 1, \ldots, 4$ with total demands of $d_l = 10$, 30, 40 and 20 units and associated unit profits of $p_l = 2$, 10, 20 and 100, respectively. Higher level customer segments are given by the nodes a, b and 0.

The aggregate values for d_a, d_b, p_a and p_b result from applying Equations (3.16) and (3.17), respectively. Corresponding values for the root node 0 can be obtained either in a decentral manner by a further application of (3.16) and (3.17) or by using a central calculation directly over all leaf nodes as given in (3.38). Both approaches lead to $d_0 = 100$ and $p_0 = 31.2$ for the total demand and the demand-weighted arithmetic mean of the unit profits in the customer hierarchy.

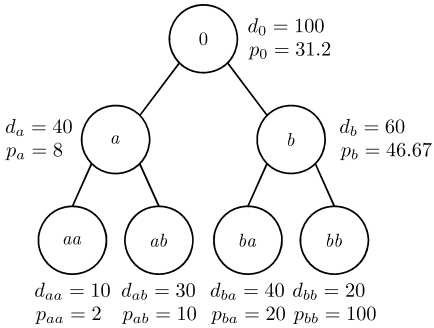


Figure 3.12. – Example hierarchy for heterogeneity measures

The values of different heterogeneity measures may now be calculated at the aggregate nodes a and b and at the root node. Furthermore, also the level of heterogeneity which exists between the aggregate nodes a and b may be determined. This has been done for the measures from Table 3.4 which are additively decomposable, i.e. CV,[44] MLD and T will be considered. The results are given in Table 3.5 below.[45]

The above example hierarchy may also be used to illustrate the differences between CV, MLD and T regarding their sensitivity to the transfer effect.[46] Assume that in the hierarchy given in Figure 3.12 a small progressive (heterogeneity-reducing) transfer of

[44] Technically, only the closely related measure $GE(2)$ is additively decomposable, but not in the strict sense.

[45] Note that the value of T is significantly below the theoretical maximum of $\ln d_0 = \ln 100 = 4.6$.

[46] This example has been adapted from Shorrocks (1980, p. 623).

	Heterogeneity measures		
	CV	MLD	T
Node a	0.4330	0.1792	0.1226
Node b	0.8081	0.3108	0.3023
Heterogeneity between a and b	0.6071	0.3028	0.2217
Root node 0	1.1181	0.5610	0.5056

Table 3.5. – Different measures for level of customer heterogeneity in example hierarchy

unit profitability occurs between nodes *aa* and *ab*. A single unit of demand in customer group *ab* now fetches a unit profit of 9.999 instead of 10 whereas a single unit of demand in customer group *aa* now fetches a unit profit of 2.001 instead of 2. All other demand units in these two segments remain unaffected. Note that the size of this transfer is small compared to the absolute value of the unit profits. Also, this transfer does not change the order of the heterogeneity distribution. However, the average profitability in node *aa* has increased slightly whereas the average profitability in *ab* has decreased.

By contrast, this transfer leaves both the total demand d_0 as well as the overall total profit in the hierarchy $\sum_{l=1}^{L} p_l \cdot d_l$ unchanged. Therefore, also the demand-weighted arithmetic mean unit profit p_0 in the hierarchy, calculated via (3.38), remains unchanged at $p_0 = 31.2$ by such a transfer.

Clearly, any heterogeneity measure which fulfills the Pigou-Dalton principle should register a reduction of overall heterogeneity as a result of this small transfer of profitability. For each of the three considered measures CV, MLD and T, the change in the heterogeneity measure is reported as $\Delta M1$ in the second data row of Table 3.6.

Now undo the first transfer and initiate a second transfer, this time between two single units of demand in the two most profitable customer segments, i.e. a single unit in node *ba* now fetches 20.001 whereas the profit associated with a single unit of demand in node *bb* has been reduced to 99.999. For each of the measures in scope, the corresponding change $\Delta M2$ in the level of heterogeneity is reported in the third data row of Table 3.6. The last row of this table gives the ratio between both changes $\frac{\Delta M1}{\Delta M2}$ and thus expresses the impact of the second transfer in terms of the first.

For the Theil index, both transfers reduce heterogeneity by an equal amount since $\frac{\Delta M1}{\Delta M2} = 1$. This illustrates that the magnitude of the transfer effect is *independent of the position* in the profitability distribution where such transfers occur. Now consider the second GE measure, the mean logarithmic deviation MLD. Since $\frac{\Delta M1}{\Delta M2} = 10$, the first transfer, taking place in the lower part of the distribution, has a significantly larger impact than the second transfer.

The reverse is true for the coefficient of variation. Here, the second transfer has a significantly larger effect. Therefore, choosing the coefficient of variation as a heterogeneity measure implicitly places more weight on transfers between demand units at the higher end of the distribution of profitability. Both a high sensitivity in the lower (MLD) or in

the upper tail (CV) of the profitability distribution are undesirable in multi-stage customer hierarchies as long as there are no specific reasons to prefer either over a measure which is neutral like T.[47]

	Heterogeneity measures		
	CV	MLD	T
Customer heterogeneity	1.12	0.561	0.506
First transfer effect $\Delta M1$	$-7.35 \cdot 10^{-8}$	$-4.00 \cdot 10^{-6}$	$-5.16 \cdot 10^{-7}$
Second transfer effect $\Delta M2$	$-7.35 \cdot 10^{-7}$	$-4.00 \cdot 10^{-7}$	$-5.16 \cdot 10^{-7}$
Relative magnitude $\frac{\Delta M1}{\Delta M2}$	0.1	10	1

Table 3.6. – Transfer effect in example hierarchy per heterogeneity measure

To summarize, Theil's index T has been shown to be an attractive measure to quantify the level of heterogeneity in customer hierarchies. Not only complies the measure with a number of standard properties such as population size independence and multiplicative scale invariance, it is also unique in fulfilling three essential properties:

- T is strictly additively decomposable, allowing for a full decomposition of the within-heterogeneity component at an aggregate level.
- All customer segments are weighted based on their contribution to total profits.
- T has a constant transfer effect sensitivity, i.e. it does not imply that there is a higher utility from heterogeneity-reducing transfers which either occur among the least or the most profitable leaf nodes.

The first property has been shown with the help of Lemmas 2 and 3, the second property follows from (3.59) and the third property has been illustrated above with a numerical example. Furthermore, the use of T is also aligned with the use of the demand-weighted arithmetic mean to aggregate the unit profit figures in the hierarchy. Hence, it is important to note that T may not be used in connection with other aggregation operators for p such as the geometric or harmonic mean.

Generally, this convenient heterogeneity measure may also be used for a number of other important management tasks. T is particularly useful for intra-company and inter-company comparisons and benchmarking efforts. For example, the application of T to quantify customer heterogeneity allows assessing different classification criteria for customer segmentation or testing whether a particular customer base is heterogeneous 'enough' to justify a profit-based allocation approach.

This concludes the introduction to customer hierarchies. The presentation in this chapter has led to a formal model of customer hierarchies. Furthermore, the links to the

[47] In the context of income inequality, Shorrocks (1980, footnote 8) has dubbed the above behavior of the coefficient of variation as "perverse" since it suggests that balancing transfers between the very rich are the best way to reduce overall inequality.

corresponding hierarchical sales organization have been illustrated. The discussion of incentive systems has shown that there exist multiple means to mitigate or even suppress principal-agent problems in hierarchical organizations, in particular with respect to forecast misrepresentations. This permits to ignore such problems in customer hierarchies in the following chapters. Lastly, the overview of heterogeneity measures has identified that Theil's index is an ideal candidate to formally quantify to what extents customer segments differ in terms of profitability in customer hierarchies. With this background on multi-stage customer hierarchies, the presentation in the following two chapters can focus on solution approaches to the DMC problem.