
Management for Professionals

For further volumes:
<http://www.springer.com/series/10101>

Hartmut Stadler
Bernhard Fleischmann
Martin Grunow
Herbert Meyr
Christopher Sürle

Advanced Planning in Supply Chains

Illustrating the Concepts Using
an SAP® APO Case Study

Prof. Dr. Hartmut Stadtler
Universität Hamburg
Fak. Wirtschafts- u. Sozialwissenschaften
Inst. für Logistik und Transport
Von-Melle-Park 5
20146 Hamburg
Germany
hartmut.stadtler@uni-hamburg.de

Prof. Dr. Martin Grunow
Technische Universität München
TUM School of Management
Production and Supply Chain Management
Arcisstr. 21
80333 München
Germany
martin.grunow@tum.de

Dr. Christopher Stürle
Ringgartenstr. 20
64625 Bensheim
Germany
cstuerle@web.de

Prof. em. Dr. Bernhard Fleischmann
Universität Augsburg
Wirtschaftswissenschaftliche Fakultät
Universitätsstr. 16
86159 Augsburg
Germany
bernhard.fleischmann@wiwi.uni-augsburg.de

Prof. Dr. Herbert Meyr
Universität Hohenheim
Lehrstuhl für Supply Chain Management (580C)
70593 Stuttgart
Germany
H.Meyr@uni-hohenheim.de

ISBN 978-3-642-24214-4 e-ISBN 978-3-642-24215-1
DOI 10.1007/978-3-642-24215-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011939772

© Springer-Verlag Berlin Heidelberg 2012

This publication contains references to the products of SAP AG. SAP, SAP AII, SAP APO, SAP Business Suite, SAP CRM, SAP EM, SAP ERP, SAP EWM, SAP PLM, SAP SCM, SAP SNC, SAP SRM, SAP TM, and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and other countries.

SAP AG is neither the author nor the publisher of this publication and is not responsible for its content. SAP Group shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP Group products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Listening to a Stradivari played by an expert is a real treat, but if it is played by a beginner. . .

In a figurative sense this also applies to implementing and making use of an advanced planning system (APS) in industry. Much frustration with APS is due to a misconception and incompetent utilization of APS. Often companies have not been aware that APS require different skills compared to transactional systems known as enterprise resource planning. Especially in complex decision situations in which many alternative activities have to be combined in order to construct a feasible, high-quality plan for a supply chain APS can unfold their full potential.

This book aims at providing the necessary knowledge needed to make the best use of APS in industry. It is intended especially for a quantitative master course in “Supply Chain Management”, “Industrial Engineering” or “Management Science”. To lower expectations a little we only provide the necessary background information of the concepts and capabilities underlying today’s APS as well as the look-and-feel of one APS, namely SAP®’s Advanced Planning and Optimization (APO). This should at least enable you to become a valuable member of an experienced APS implementation team.

Preparing the content of this book started more than a decade ago when we made first attempts together with our students to test various modules of APS from different software vendors. We are indebted to these students and hence have prepared a list of names for our acknowledgements (printed at the end of this book). We would also like to express our sincere gratitude to several experts from SAP Deutschland AG, especially Katrin Diesing, Berthold Hege, Mathias Göbelt, and Clemens Kriesel for providing valuable advice regarding the use and description of SAP APO on top of their daily duties. From the very start Heinz Braun promoted our project, e.g. by providing access to the SAP APO software for our students and research assistants as well as initiating contacts with experts from the development team of SAP APO – this was great! Thomas Pöhler, datango AG, Berlin, has to be mentioned, too, for teaching us how to make the best use of the datango e-learning software which was used to create the interactive learning units accompanying our book.

We would also like to thank our co-authors which have helped us writing some chapters. Their names are indicated at the start of these chapters.

Last but not least we would like to mention Christopher Haub for recording the interactive learning units accompanying our book as well as Julian Wulf for his relentless efforts in preparing the .tex files for the Springer publisher.

Now it is up to you to make the best use of our book and the interactive learning units – we wish you great success!

Bernhard Fleischmann
Augsburg

Martin Grunow
München

Herbert Meyr
Hohenheim

Hartmut Stadler
Hamburg

Christopher Suerie
Walldorf

in July 2011

Contents

Preface	V
Introduction	1
<i>Hartmut Stadler, Christopher Haub</i>	
Part I	
<hr/>	
1 The Frutado Case	11
<i>Bernhard Fleischmann</i>	
1.1 The Frutado Company	11
1.2 The Current Planning System	14
1.3 Data Analysis	14
1.4 Purpose of the Frutado Case	17
1.5 Characteristics	18
2 Hierarchical Planning and the Supply Chain Planning Matrix	21
<i>Hartmut Stadler, Bernhard Fleischmann</i>	
2.1 Principles of Hierarchical Planning	21
2.2 Rolling Schedules	26
2.3 The Supply Chain Planning Matrix	28
2.4 Planning Tasks in the Frutado Case	32
3 SAP® APO - Module Matrix and General Principles	35
<i>Christopher Sürle</i>	
3.1 Module Matrix and Related Systems	35
3.2 Data Flows (Technical and Process-Related)	39
3.3 General Terms and Principles	44
3.3.1 Models and Planning Versions	44
3.3.2 Master Data	45
3.3.3 Transactional Data	50
3.3.4 User Interface	53
3.4 The SAP® APO Solution for the Frutado Case	59

Part II

4 Demand Planning (DP)	67
<i>Herbert Meyr</i>	
4.1 Introduction to Demand Planning	68
4.1.1 Measuring the Forecast Quality	69
4.1.2 The Objects to Forecast	70
4.1.3 Basic Forecasting Approaches	72
4.1.4 The Demand Planning Process	74
4.2 Demand Planning Models in the Literature	77
4.2.1 Level Demand	78
4.2.2 Trend and Seasonality	79
4.3 Demand Planning Methods for Time Series Analysis	80
4.3.1 Level Demand	80
4.3.2 Additive Trend	83
4.3.3 Multiplicative Seasonal Demand	85
4.3.4 Methods for other Demand Models	86
4.4 Planning Tasks and Data for the Frutado Company	87
4.4.1 Available Data	87
4.4.2 Planning Tasks and Level of Detail	88
4.5 Modeling the Frutado Planning Tasks	89
4.5.1 Introduction to SAP® APO DP	89
4.5.2 Modeling with SAP® APO	94
4.6 Implementation and Disaggregation of Results	99
4.7 Demand Planning Learning Units	99
4.7.1 Overview	99
4.7.2 Basic Stream	100
4.7.3 In-Depth Stream	102
5 Master Planning - Supply Network Planning	109
<i>Hartmut Stadler</i>	
5.1 Medium-Term Planning Models in the Literature	110
5.2 Solution Procedures for LP and MIP	118
5.3 Planning Tasks and Data for the Frutado company	122
5.3.1 Planning Tasks and Level of Detail	122
5.3.2 Data	123
5.4 Modeling the Frutado Planning Tasks	126
5.4.1 Introductory Remarks	126
5.4.2 Basic Frutado Model	128
5.4.3 Extensions	132
5.5 Implementation and Disaggregation of Results	135
5.6 SNP Learning Units	136
5.6.1 Overview	136
5.6.2 Basic Stream	138
5.6.3 In-Depth Stream	144

6	Production Planning and Detailed Scheduling (PP/DS)	149
	<i>Hartmut Stadler, Christopher Sürie</i>	
6.1	Operating Principles of Production Segments	150
6.1.1	Criteria	150
6.1.2	Job Shops	151
6.1.3	Flow Lines with Setups	152
6.1.4	One of a Kind Production	154
6.1.5	Further Production Segments	156
6.1.6	Conclusions and Additional Remarks	157
6.2	Detailed Scheduling – Solution Algorithms	157
6.2.1	Overview	157
6.2.2	An Example	159
6.2.3	Solution by a Priority Rule	160
6.2.4	Solution by a Genetic Algorithm	162
6.3	Planning Tasks and Data for the Frutado Company	170
6.3.1	Planning Tasks	170
6.3.2	Data	171
6.4	Modeling the Frutado Planning Tasks	173
6.4.1	Basic Frutado Model	173
6.4.2	Extensions	175
6.5	Implementation and Results	178
6.6	PP/DS Learning Units	179
6.6.1	Overview	179
6.6.2	Basic Stream	181
6.6.3	In-Depth Stream	190
7	Global Available-to-Promise (global ATP)	195
	<i>Bernhard Fleischmann, Sebastian Geier</i>	
7.1	ATP: Basics and Literature	195
7.2	Planning Tasks and Data for Frutado	201
7.2.1	Planning Tasks	201
7.2.2	Data	201
7.3	Modeling the Frutado Planning Tasks and Implementation in Global ATP	202
7.3.1	Introduction to SAP® APO Global ATP	202
7.3.2	Customization of Order Promising at the Frutado Com- pany	203
7.3.3	Basic Global ATP Model and Implementation for Frutado	204
7.3.4	Extensions	207
7.3.5	Processing the Results	208
7.4	Global ATP Learning Units	210
8	Deployment	217
	<i>Martin Grunow, Poorya Farahani</i>	
8.1	Introduction to Deployment	217
8.1.1	Deployment Modeling Framework	223

8.1.2	Deployment Model Classes	224
8.2	Planning Tasks and Data for Frutado	233
8.2.1	Planning Tasks and Level of Detail	233
8.2.2	Data	233
8.3	Modeling Deployment for Frutado	234
8.4	Implementation	239
8.4.1	Deployment Planning Initialization	241
8.4.2	Solution Methods in SAP® APO	242
8.5	Deployment Learning Units	245
8.5.1	Overview	245
8.5.2	Basic Stream	245
8.5.3	In-depth Stream	246
9	Transportation Planning/Vehicle Scheduling (TP/VS)	249
	<i>Martin Grunow, Bryndís Stefánsdóttir</i>	
9.1	TP/VS in the Literature	250
9.1.1	Transportation Load Building	250
9.1.2	Formulation of the Basic Vehicle Routing Problem	252
9.1.3	Typical Problem Classes of Vehicle Routing Problems	256
9.2	Solution Approaches for Vehicle Routing Problems	259
9.3	Planning Tasks and Data for the Frutado Company	261
9.3.1	Planning Tasks	261
9.3.2	Data	263
9.4	Modeling the Frutado Planning Tasks	265
9.4.1	TLB for the Frutado Company	265
9.4.2	Frutado’s Vehicle Routing Problem	266
9.4.3	Extensions	272
9.5	Implementation and Integration with Deployment	274
9.6	TP/VS Learning Units	276
9.6.1	Overview	276
9.6.2	Basic Stream	277
9.6.3	In-Depth Stream	282

Part III

10	Final Remarks	289
	<i>Hartmut Stadler</i>	
10.1	Implementation of an APS	289
10.2	Evaluation of APS	291
	Index	295
	About Contributors	301

Part I

The Frutado Case and Foundations of Advanced Planning

The Frutado Case

Bernhard Fleischmann¹

¹ University of Augsburg, Department of Production & Logistics, Universitätsstraße 16, 86135 Augsburg, Germany

1.1 The Frutado Company

The fictitious company Frutado has been designed in a master thesis (Lebreton 2001) at the University of Augsburg. The Frutado case was derived from a real case and has been used many times for teaching and research (see for instance Mauch 2010).

Frutado is a medium-sized manufacturer of beverages with the headquarter situated in Augsburg, Germany. Originally specialized in the production of *fruit juices*, the company has expanded the range of products and the capacities by taking over two competitors in the mid 90s. It now produces fruit juices as well as *ice teas* at three sites, located in Augsburg, in Ludwigshafen, and in Magdeburg. Each site consists of a production plant and an adjacent distribution center (DC).

The production process consists of blending the beverages and filling them in various formats. Only the filling stage, which takes place on automated continuous flow lines, is considered in the Frutado case. The filling lines are potential bottlenecks, whereas the blending step is not critical. In each of the three plants, two parallel filling lines (FL) are available. Each filling line is applicable for a certain group of products only, due to its technical equipment.

The production department is working around the clock from Monday to Friday in three shifts per day and 15 shifts per week. In case of bottlenecks, up to three additional Saturday shifts can be used at extra overtime costs. As each filling line has to process several products, change-overs occur. They take between 10 minutes and 2 hours as *setup time* and cause *setup costs*, between 10 and 500 \$, depending on the sequence of the products.

The product range comprises 13 types of fruit juices and 6 types of ice teas. The fruit juices are produced without preservatives and have a limited storage life. In order to guarantee a sufficient shelf life to the retailers, the duration in Frutado's stock must not exceed three weeks. The ice teas, in contrast, are not perishable. Some products show a seasonal demand with a strong peak in summer, some are designed as special drinks for the cold season and have a peak demand in winter, others have a rather flat demand over the whole year. Most products can be produced at a single plant only. A few products can be produced at two or even three plants. In order to distribute the full range of products from each DC to the customers, the DCs must be fully assorted and cross shipping between the sites is necessary.

Immediately after production, the finished products are either stored into the DC at the same site or shipped to one or two of the other DCs. Cross shipping from DC to DC is undesired, because it entails duplicate handling in the DCs and may cause unnecessary transports. Nevertheless it is used as an emergency action in case of unforeseen shortages.

The customers of Frutado are retailers and restaurants with 60 delivery points altogether, mainly in Southern and Eastern Germany. The delivery points are simply called "customers" in the following. There is a fixed allocation of 20 customers to every DC. The transports from the plants to the DCs and between the DCs are done in trucks of 18 tons. For the deliveries from a DC to its customers trucks of 18 tons and of 9 tons are available, which operate in tours linking several customers. All vehicles are owned by Frutado and the transports are organized by the Logistics department. [Figure 1.1](#) shows the structure of Frutado's supply chain and [Figure 1.2](#) illustrates the geographical locations of the DCs and the customers.

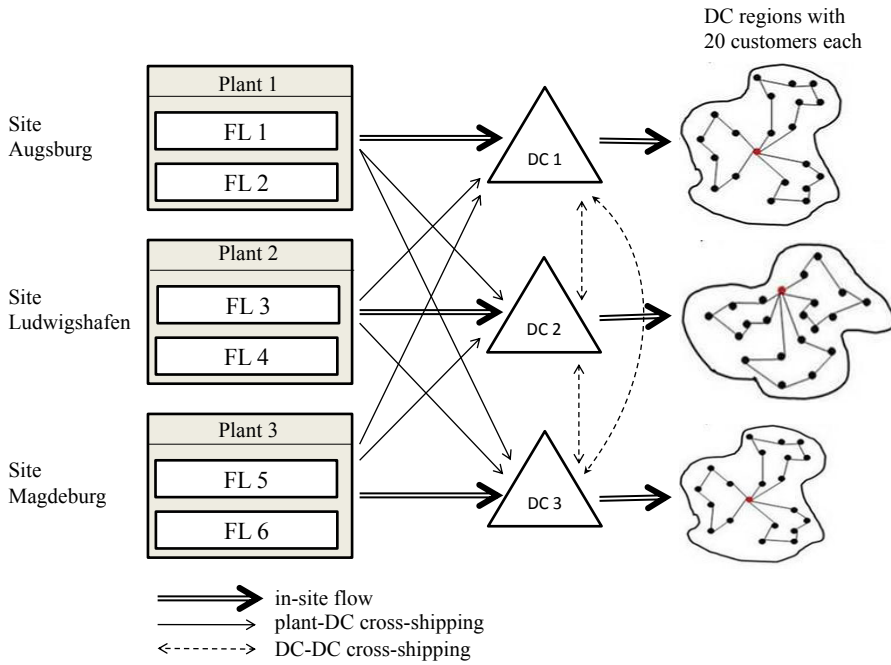


Figure 1.1
Supply chain of frutado with distribution centers (DC), plants, and filling lines (FL)

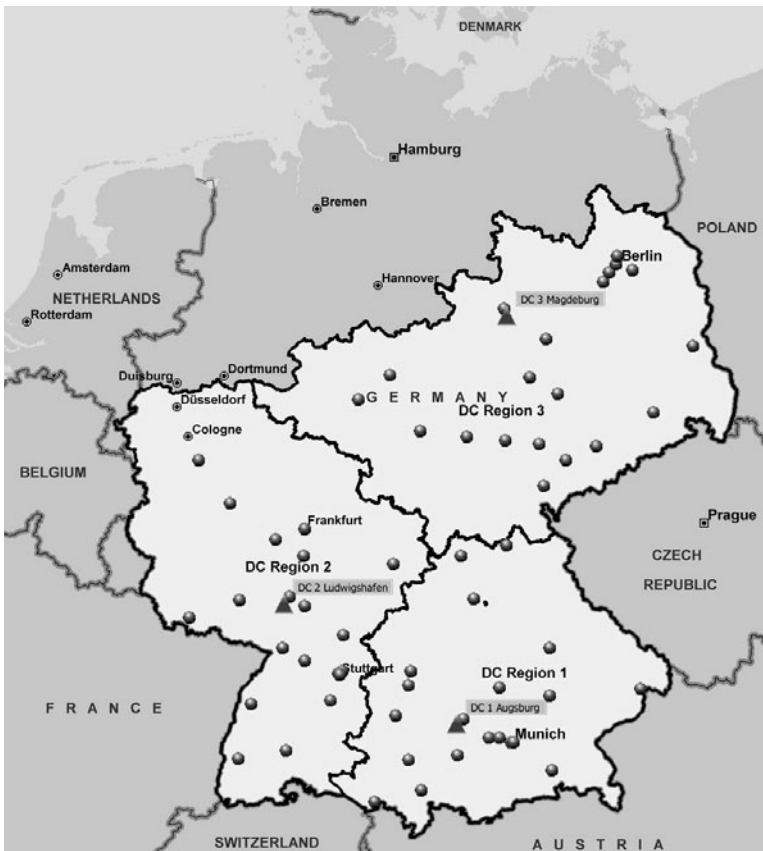


Figure 1.2
Geographical location of the DCs and customers

1.2 The Current Planning System

So far, the planning system of Frutado is based on the concept of Material Requirements Planning (MRP) (see Heizer and Render 2011, Chap. 14) and can be roughly described as follows: For every DC region, a sales manager is responsible for estimating the future sales and establishing a sales plan for three months. These sales plans are the input for the self-developed MRP software, which allocates the sales volumes to the plants and aggregates them month by month into production orders (POs). For those products which can be produced in several plants, the Frutado management has fixed priorities in the following way: Each plant should satisfy the demand of its own DC for the local products. If a product can be produced in exactly two plants, the third DC should be supplied by that plant which minimizes the production and transport costs. In each plant, the production planner transforms the POs into schedules for the two local filling lines. Usually he splits the monthly POs into weekly cycles and tries to keep a sequence with low setup times. Unfortunately, the overload of some lines, frequent stock-outs, and the pressure of the Logistics department, who wants to fulfill the current customer orders, enforce him to improvise more often than not. In particular, reallocations of the production between the plants often have to be agreed by phone. As a counter-measure, the production planner tends to produce in advance whenever there are free capacities.

As a result, the operations are far from running in an efficient way and cause high stocks, unnecessary high costs of production and extra shifts and a low service level for the customers. Finally, Frutado's management has decided to reorganize the planning system fundamentally. First, they established a new Supply Chain Department, responsible for the coordination of all operations from the procurement up to the fulfillment of the customer orders. Second, this department is in charge of selecting and implementing an advanced planning system.

1.3 Data Analysis

As a first step, the Supply Chain Department analyzed the operational data of the last year. [Table 1.1](#) shows the partition of the total yearly sales volume of 95 Million liters over the 19 products and the three DC regions. The DC 3 (Magdeburg) has to supply the largest demand (47%), followed by DC 1 (Augsburg, 32%) and DC 2 (Ludwigshafen, 21%). The product mix, however, does not show significant regional differences.

To analyze the seasonality of the demand, even two years of past demand data have been used. [Figure 1.3](#) (adapted from Christ 2003) shows the demand curves for three groups of products with summer season, winter season, and steady demand, respectively.

[Table 1.2](#) presents the result of an ABC classification of the products and the allocation of the products to the production sites and lines. The three A products together represent 66% of the total sales volume, whereas the

Product	Type	DC1	DC2	DC3	Total
1	Juice	808	579	1095	2482
2	Juice	670	477	935	2083
3	Juice	8067	5607	11204	24877
4	Ice Tea	4495	2965	5945	13404
5	Juice	8014	5425	10751	24190
6	Ice Tea	1575	1050	2046	4670
7	Juice	1014	606	1279	2898
8	Ice Tea	202	117	216	534
9	Ice Tea	146	77	169	392
10	Juice	173	134	273	580
11	Juice	174	117	254	545
12	Juice	1260	916	1737	3913
13	Juice	1331	860	1728	3920
14	Juice	187	104	248	540
15	Juice	127	95	146	368
16	Juice	135	64	135	335
17	Ice Tea	1238	861	1731	3830
18	Juice	538	389	745	1672
19	Ice Tea	1105	722	1612	3439
Total		31258	21166	42250	94673

Table 1.1
Sales volume last year
in [1000 liters]

ten C products share only 10% of it. A similar partition can be observed regarding the customers: For each DC region, the first four “A customers” are responsible for 60-75% of the demand of all customers. Regarding the days of the week, however, no significant differences can be found.

Product	Type	Cumulated Percentage of Sales	ABC Class	Production Possibilities					
				Plant 1 Augsburg		Plant 2 Ludwigshafen		Plant 3 Magdeburg	
				FL 1	FL 2	FL 3	FL 4	FL 5	FL 6
3	Juice	26.3%	A	0.22				0.1	0.11
5	Juice	51.8%	A	0.23				0.2	0.17
4	Ice Tea	66.0%	A	0.26				0.12	0.15
6	Ice Tea	70.9%	B		0.53		0.54		
13	Juice	75.1%	B		0.57		0.58		
12	Juice	79.2%	B		0.51				
17	Ice Tea	83.2%	B					0.08	
19	Ice Tea	86.9%	B			0.45			
7	Juice	89.9%	B		0.53		0.44		
1	Juice	92.6%	C		0.4	0.5			0.4
2	Juice	94.8%	C		0.46	0.38			0.29
18	Juice	96.5%	C					0.15	
10	Juice	97.1%	C		0.53				
11	Juice	97.7%	C		0.41				
14	Juice	98.3%	C		0.5				
8	Ice Tea	98.8%	C		0.6				
9	Ice Tea	99.3%	C		0.55				
15	Juice	99.6%	C		0.49				
16	Juice	100.0%	C		0.5				

Table 1.2
ABC classification of
the products and
production coefficients
[hours per 1000 liters]

The right-hand part of [Table 1.2](#) shows the production coefficients for the possible allocations of the products to the filling lines. The inverse of these coefficients is the throughput of the line, which ranges from 1,700 to 12,500 liters per hour. The fastest lines FL 5 and FL 6 are mainly used for the A products.

The priority rules for allocating the sales volumes to the plants, as explained in Section 1.2, concern the three A products, produced in Plants 1 and 3, and the B Products 6, 7, and 13, produced in Plants 1 and 2. For the supply of DC 3 with the latter products, the cost difference between Plants 1 and 2 is low. As FL 2 in Plant 1 has a very high load, DC 3 is supplied from Plant 2. For the A products, the most recent filling line FL 6 has by far the lowest production cost, because it requires less workers. Therefore, it

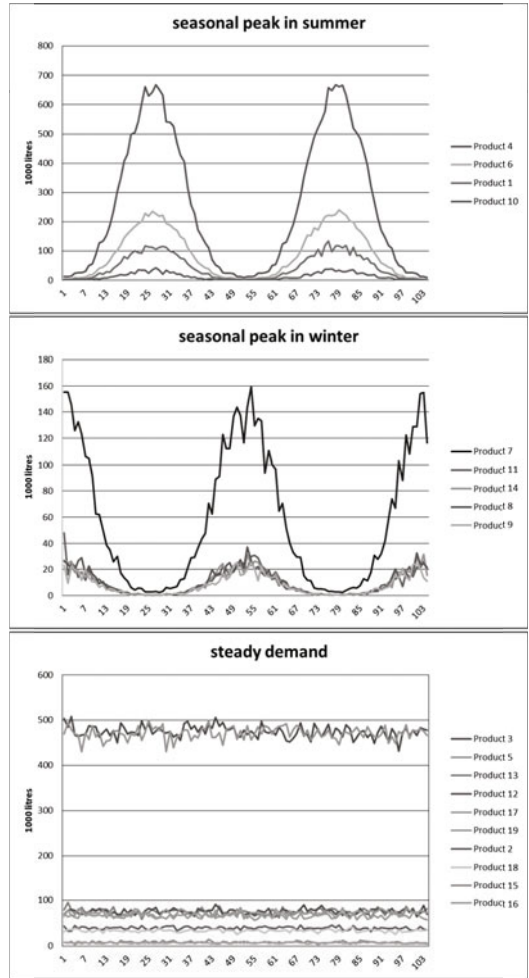


Figure 1.3
Demand curves for products with summer season, winter season, and steady demand

is prioritized for both the supply of DC 2 and DC 3. These priorities are in boldface in [Table 1.2](#).

Based on these priorities, it is possible to calculate the average utilization by multiplying the sales volumes per DC (see [Table 1.1](#)) with the production coefficients of the prioritized lines. The resulting production hours of the lines per year are shown in [Table 1.3](#). These figures do not yet contain the setup times, which can be estimated as the total setup time in the optimal sequence through all products of a line. For the current weekly production cycles, this is shown as “setup time per week” in [Table 1.3](#). The total load per year is composed of the production and the setup times. Related to the total available working time per year of about 6120 hours, the average utilization of every line is shown in [Table 1.3](#). It reveals a high overload of FL 2 and FL 6 and a rather high utilization of FL 1 and FL 4. These averages do not take the seasonality of the demand into account. In peak seasons, the utilizations will even be higher.

Thus, the production management is often forced to deviate from the nor-

Product	FL 1	FL 2	FL 3	FL 4	FL 5	FL 6
3	1775					1849
5	1843					2750
4	1169					1336
6		835		1671		
13		759		1501		
12		1996				
17					306	
19			1548			
7		537		829		
1		323	289			438
2		308	181			271
18					251	
10		307				
11		224				
14		270				
8		321				
9		216				
15		180				
16		167				
total occupation (hours p.a.)	4786	6442	2018	4002	557	6645
setup time per week	1	6,5	1	1	1	4
setup time p.a.	52	338	52	52	52	208
occupation p.a.	4838	6780	2070	4054	609	6853
average utilization	79.1%	110.8%	33.8%	66.2%	10.0%	112.0%

Table 1.3
Occupation of the lines in hours per year and average utilization

mal product allocation. Discharging FL 6 by shifting production to FL 5 can be done easily by local decisions in Plant 3, but bottlenecks on the other filling lines require short-term agreements between the plants and the Logistics Department.

Questions and Exercises

1. The sequence-dependent setup times on FL 6 are given in [Table 1.4](#) in form of a from/to matrix. Verify the setup time per week on FL 6, as shown in [Table 1.3](#) (4 hours) by searching for a cyclic sequence with lowest total setup time. Note that the precise result is 3.833 hours.

from/to product	1	2	3	4	5
1		0.167	2	2	2
2	0.167		2	2	2
3	1	1		0.333	0.333
4	1	1	0.333		0.5
5	1	1	0.333	0.333	

Table 1.4
Setup times in [hours] on FL 6

2. Verify the total yearly occupation of FL 6, as shown in [Table 1.3](#), using the data of [Tables 1.1](#) and [1.2](#) and the allocation rules explained above.

1.4 Purpose of the Frutado Case

The Frutado case is aligned with the operations along the supply chain. It has been designed in view of the following three purposes:

- Teaching Supply Chain Planning:
The Frutado case demonstrates the various planning tasks within a

company and reveals their interdependencies. It provides examples for practicing planning methods and analyzing the results.

- **Research in Supply Chain Planning:**
The Frutado case permits to compare planning concepts which differ in the way how the planning tasks are combined in a planning system and how they are linked by information flows.
- **Test of Advanced Planning Systems:**
The Frutado case helps to understand the implementation and the use of an APS in a company. It can also be used to compare different APSs.

In order to achieve these objectives, it was necessary to include all essential operations of a company with detailed data into an integral case. This requirement had two important implications:

First, it was necessary to restrict the example company to a specific industry sector, because the planning tasks differ essentially in diverse industries (see Meyr and Stadler 2008, for a classification). It is one of the weaknesses of the classical MRP concept, that it neglects these differences. For the Frutado case, the consumer goods sector was selected, more precisely the food and beverages industry. This sector has been less investigated than the mechanical engineering sector which is often assumed tacitly, if planning tasks are described. The characteristics of the consumer goods industry that are considered in the Frutado case are explained in the following section.

Second, a compromise had to be found between the objectives of designing a realistic example company and modeling all operations in detail. In order to keep the model transparent for teaching and research purposes, simplifications were unavoidable. The number of products and of customers is much smaller than in reality, and the production process is simplified.

Thus, the resulting company, Frutado, is not a detailed model of a real company or of a typical consumer goods company. But it is an integral model of typical supply chain operations and planning tasks in a consumer goods company.

1.5 Characteristics

The supply chain of a consumer goods manufacturer, in particular in the case of foods and beverages, shows the following characteristics, which appear in the Frutado case.

Products

The company produces standard products on stock. The production is driven by forecasts of the future demands, which may exhibit seasonal variations. The products have a rather simple structure, they are composed of a few

raw materials according to recipes. However, by varying the ingredients and the packing formats, the result may be a large number of end product varieties and stock keeping units (SKUs). This is a so-called *divergent product structure*. Often the products are perishable with a shelf life of some weeks only.

In the Frutado case, the number of products has been fixed to only 19, for the sake of simplicity, as explained in the previous section. This number is untypically low for a consumer goods company.

Production Process

The production process consists of only 1-2 stages, often characterized as “Make and Pack”. It takes place on highly automated mass production lines generating a continuous flow of output. There may be several production lines working in parallel which can be dedicated to a certain product group or general-purpose lines. As high automation is only economic in case of high utilization, the capacity of the lines is usually tightly limited.

As a rule, the number of products is much larger than the number of parallel lines. Therefore a production line has to process a number of products one after the other. Product changeovers require setup of the lines, for instance for cleaning the line or tuning it to another packing format. This causes unproductive *setup times* and *setup costs*, which may depend on the sequence of the products before and after the changeover.

Therefore, the production typically takes place in a sequence of batches or *lots* of the different products. The interval between two consecutive lots of a certain product is needed for producing the other products. Therefore, the replenishment lead time for the finished product contains part of this interval, in spite of the very short production time per unit of the product.

Distribution

Mostly, a consumer goods producer has several production sites with different ranges of products. Then, the finished products are brought together in a first distribution step, to one or a few *distribution centers* (DCs) which keep stocks of all products. Typically, the DCs are situated at or near a factory.

The destinations of the distribution are the retail outlets. In a traditional distribution system in Germany, the manufacturer delivers up to the outlets. However, recently, the big retail chains have installed their own central warehouses or stockless cross-docking points, from where they organize the deliveries to the outlets by themselves. Then, these retail centers are the final destinations for the manufacturer’s distribution. This concept has reduced the number of delivery points for the manufacturer considerably. In the Frutado case, only 60 “customers” are considered which may represent retail centers or outlets. Again, this number is untypically low, even in case of retail centers, for the sake of simplicity.

The shipments from the factories to the DCs take place on the basis of demand forecasts. Often, they are synchronized with the production in order to avoid duplicate stock keeping. The deliveries from the DCs to the retail delivery points are on order, with very short lead times of 24-48 hours, enforced by the retailers. The allocation of the delivery points to the DCs is fixed for organizational reasons over a medium-term horizon.

Procurement

Given the relatively small number of raw materials, compared to the finished products, the procurement function is easier than the distribution. Usually, the number of suppliers is rather low, and the lead times are short and reliable. There may be complicating factors, however, in case of natural raw materials for food and beverages, which can show a fluctuating quality and seasonal availability, due to harvest periods and, consequently, fluctuating prices. But this is not considered in the Frutado case. There, the procurement function is disregarded, and it is assumed that the required raw materials are always available.

Questions and Exercises

1. Which processes of the Frutado company are made to stock and which are made to order?

Bibliography

- Christ, S. (2003) *Implementierung der Supply Chain einer Modellfirma in SAP APO 4.0 - Demand Planning und Supply Network Planning*, Studienarbeit, Technical University of Darmstadt, Germany
- Heizer, J.; Render, B. (2011) *Principles of Operations Management*, Pearson, Prentice Hall, 8th ed.
- Lebreton, B. (2001) *Aufbau einer Modellfirma und ihre Implementierung mit J.D. Edwards Active Supply Chain*, Diplomarbeit, University of Augsburg, Germany
- Mauch, K. (2010) *Hierarchische Planung von Produktion und Distribution in der Konsumgüterindustrie*, Verlag Dr. Kovač, Hamburg
- Meyr, H.; Stadtler, H. (2008) *Types of supply chains*, in: H. Stadtler; C. Kilger (Eds.) *Supply Chain Management and Advanced Planning*, Springer, Berlin et al., 4th ed., 65–80

Hierarchical Planning and the Supply Chain Planning Matrix

Hartmut Stadtler¹, Bernhard Fleischmann²

¹ University of Hamburg, Institute for Logistics and Transport, Von-Melle-Park 5, 20146 Hamburg, Germany

² University of Augsburg, Department of Production & Logistics, Universitätsstraße 16, 86135 Augsburg, Germany

In Section 2.1 we will introduce the principles of hierarchical planning – the architecture of today’s APS. Rolling schedules follow in Section 2.2. In Section 2.3 an overview of the various planning tasks occurring in a supply chain is presented. These tasks can be structured and visualized in the Supply Chain Planning Matrix. Finally (Section 2.4), the concept of the hierarchical planning system designed for the Frutado company will be outlined.

2.1 Principles of Hierarchical Planning

“Divide and conquer” is an old, well-known strategy for solving complex decision problems. However, this saying does not provide a clue *how* to divide a given decision problem. Hierarchical planning (HP) – albeit it does not give a general answer – at least provides some principles and examples on how to divide a decision problem for industrial practice. The aim is to achieve a feasible solution with a good quality. Hax and Meal (1975) were the first providing recommendations on how to divide an operational production planning problem into hierarchical planning levels, and on how to construct solvable decision models for each level. Especially, these individual models are linked in a specific manner in order to provide a solution to the overall decision problem. The resulting HP system was applied to a tire company. The concept of HP can be described by its five principles:

- Decomposition and hierarchical structure
- Coordination
- Aggregation
- Model building, anticipation, and disaggregation
- Model solving

Subsequently these building blocks will be explained in detail.

Decomposition and Hierarchical Structure

A monolithic (i.e. total) model of all the planning tasks arising in a supply chain usually will neither be solvable nor accepted by the various managers in charge of specific tasks. Also, a monolithic model will require large amounts of up-to-date data, and revising data will result in frequent re-planning. Last but not least, bottom line managers will be reluctant to input their local knowledge (like experiences regarding their workforce) into an abstract model at the top of an organization's hierarchy.

This reasoning provides a first clue how to divide the total model of the SC, namely to generate a (sub-) model for each *decision unit* in the hierarchy – which then becomes a *planning unit*. The model at the top (planning) level will incorporate decisions with a large impact on the profitability and competitiveness of the supply chain as a whole. In order to evaluate the impact of these decisions the planning interval at the top level is the longest. The next lower level may control the decisions at the plants or the distribution system. A third level may look at specific operations at each plant (or DC) and be attributed to respective group leaders or process managers. The simplest view of an HP system is a system with only two hierarchical levels and only one decision unit at the bottom level (often called base level, see [Fig. 2.1 – Schneeweiss 2003](#)).

Once decisions become final they are implemented, i.e. applied to the object system which is the physical supply chain with its material and financial flows.

Coordination

To link decisions of the various models pertaining to the decision units within the hierarchy *coordination* is necessary. This may be achieved by *instructions* imposed on the lower level decision unit(s). These may be primal instructions (e.g. by allocating products to be produced in specific plants) or it may be done by dual instruction (i.e. imposing a price for utilizing specific resources). Often “directives” are used as a synonym for “instructions”. These are always applied top down. If a subordinate decision unit is not content with the instructions (e.g. because it leads to an infeasible plan) it may respond by a counter proposal *before* the plan becomes effective. This is a *feedback*

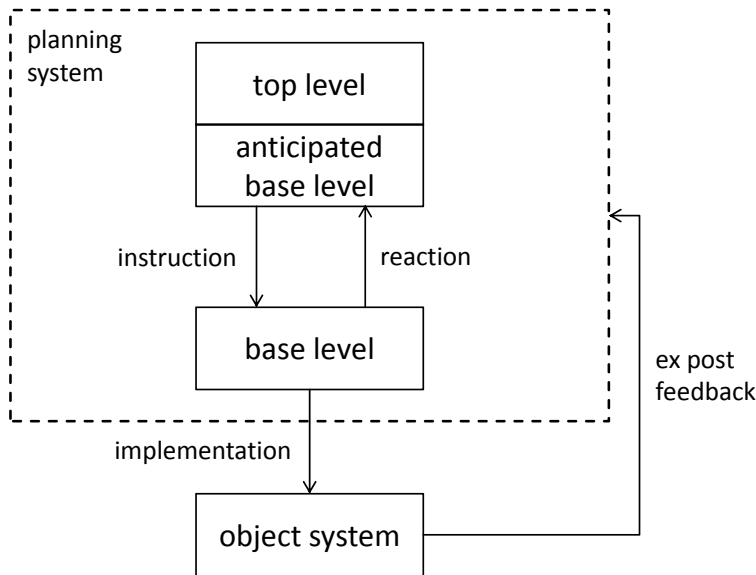


Figure 2.1
Hierarchical planning system

or *reaction* which may improve the overall plan but will require additional efforts and time for a further iteration.

An ex-post feedback is returned from the object system to the HP system reporting what has been achieved (production amounts, capacity utilizations etc.) during the re-planning interval (usually after one period has passed).

Now we are able to define an HP *system*: It comprises a set of decision units where each decision unit is assigned to a specific planning level, and for each decision unit – except for the one at the top – there is a single superior decision unit at an upper planning level which controls or limits its potential decisions by imposing instructions.

Aggregation

A means to reduce complexity of higher level planning models (especially at the top) is *aggregation*. Aggregation can be performed on two dimensions – time and entities. Instead of looking at the system in (nearly) continuous time, like it is often done in scheduling, we may look at quantities relating to certain intervals of time – named periods. As an example we have demand forecasts for specific products per period or available capacities of resources per period.

An aggregation of entities implies that a set of entities with similar characteristics forms an aggregate (entity). As an example we can group all trucks of one's own fleet with the same load capacity and the same engine into a specific group of trucks. A popular way is to aggregate products into product groups. Here, products with similar production characteristics (e.g. production coefficients) may form a product group. An appealing feature of aggregating end products to product groups is the (potential) reduction of the forecast error: If the coefficient of correlation of forecast errors of

products forming a product group (where demand forecasts are calculated as the sum of demand forecasts of the individual products) is less than 1, then the error of demand forecasts of the product group is smaller than the (weighted mean of) forecast errors of the individual products.

Unfortunately, aggregation always incurs some loss of information. Care must be taken that aggregation will not result in infeasible plans (at lower levels of the planning hierarchy). As an example we discuss the aggregation of product variants into a product (group) and the aggregation of respective demand forecasts into *effective demands*. Consider “Ice Tea - Product 4” (see Table 1.1 in Chap. 1). So far the Frutado company only serves the German market. Since the product sells well, management has decided to produce and sell the product also for Denmark and Poland. These bottles require labels in the respective language. Hence, *inventories* have to be distinguished although the production process is the same.

In order to aggregate demand forecasts for these product variants (see Table 2.1) into demand forecasts for the product (group) *net demands* have to be calculated first for each product variant. In a second step (see Table 2.2) these net demands are summed over all product variants period by period resulting in effective demands of the product (group) “Ice Tea - Product 4”. Note that simply adding the columns in Table 2.1 would result in a demand of ‘zero’ in period 1. However, without producing “Ice Tea - Product 4” in period 1 lost sales for product variants “Ice Tea - Product 4 - DK” and “Ice Tea - Product 4 - PL” of five units each would occur.

Table 2.1
Initial stocks and effective demands for product variants of “Ice Tea - Product 4”

Product variant	Initial stock	Gross demands for period t=			
		1	2	3	4
DK	0	5	10	5	10
PL	10	15	10	10	10
D	30	10	15	10	10

Table 2.2
Calculation of effective demands for product “Ice Tea - Product 4”

Product variant	Net demands for period t=			
	1	2	3	4
DK	5	10	5	10
PL	5	10	10	10
D	0	0	5	10
Effective demands of product (group)	10	20	20	30

Model Building, Anticipation, and Disaggregation

For each planning unit the decision situation can be documented in a model. In an HP system these are usually formalized, mathematical models. The

model will incorporate variables representing the decisions the planning unit has to take. Also, the constraints limiting the decision space will be specified at an appropriate level of detail (see aggregation). In order to improve the quality of plans the subordinate planning unit (s) – its capabilities and potential reactions to our instructions – should be considered in the model too. This – so called *anticipation function* – can range from non-existence to perfect anticipation. The latter requires the incorporation of the complete model of the sub-ordinate planning unit(s) in the respective detail. Obviously, this will result again in a monolithic model and thus will not reduce planning efforts associated with HP. Hence, a compromise is needed such that the most important features of the sub-ordinate planning unit(s) are taken into account by the superior planning unit.

As an example we will discuss three different anticipation functions for a period's capacity of a machine (group). Recall that available productive capacity is reduced by setup times. If these are sequence dependent this loss of capacity is difficult to estimate at a planning unit which does not know the actual sequence of products. Firstly, a simple anticipation of the capacity loss is to monitor the loss of capacity in previous periods and to calculate the mean loss which then will be subtracted from gross capacity.

As a second alternative we make use of the fact that the loss of capacity depends on the subset of products to be actually produced in a given period. Hence, anticipation can be performed by estimating the “usual” setup times for this combination of products (see Rohde 2004 for an anticipation function for setup times based on neural networks).

Thirdly, a perfect anticipation would be to do the sequencing also at the superior planning unit. In essence an anticipation function can improve the quality of the overall plan and avoid instructions leading to infeasible plans at the lower planning levels.

If instructions from the superior planning unit are based on aggregate items (like product groups) these have to be *disaggregated* at sub-ordinate planning unit(s). Usually, disaggregation incurs some degree of freedom which should be used intelligently. Let us refer to the example calculating effective demands (see [Table 2.1-2.2](#)). Now, let us assume an instruction which requires to produce 20 units of product “Ice Tea - Product 4” in period 1. How much should we produce of each variant?

If each product requires a major setup on a filling line (and only a minor setup in between product variants) then it is wise to split production amounts among variants such that they have equal *runout times*. For the example above this would mean production of 0 units of variant D and 10 units of variants DK and PL each. The resulting runout time would then be in the middle of period 2.

Model Solving

When creating a model for a planning unit one should already consider the solution approach. Potential solution approaches may span from manual

planning (perhaps with the help of a spreadsheet) to simple heuristic rules, intelligent meta-heuristics, or even optimization models. Here we have to find a compromise between solution quality, efforts, and available (computational) times. Note that in each module of SAP APO the user may choose among various solution approaches. In the Frutado case we will concentrate on the most advanced solution approaches available.

2.2 Rolling Schedules

In the preceding section we argued that aggregation of time results in a planning interval divided into several periods, and that data and variables relate to periods. Assume there are T periods in the planning interval (see Fig. 2.2), usually of equal length. To reduce the model size also a telescopic time frame is possible where periods become longer the more they reach into the future.

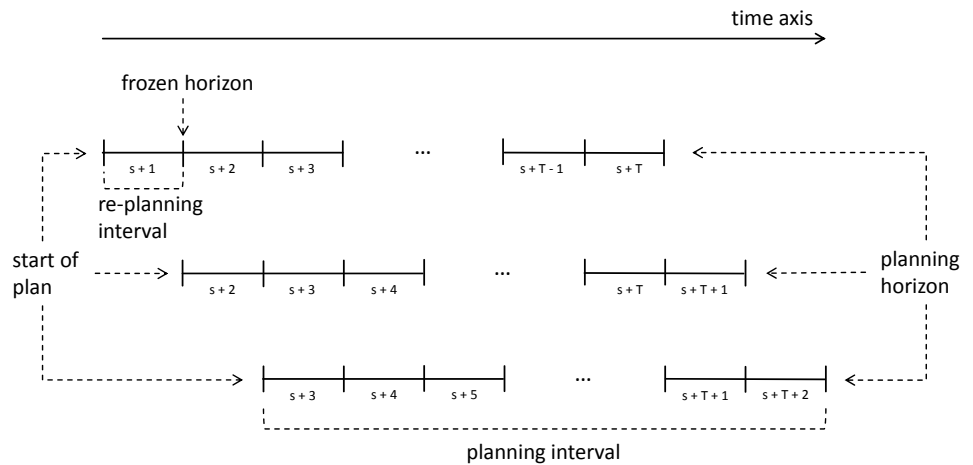


Figure 2.2
Rolling schedules with a planning interval of T periods

But how long should we look into the future? It is obvious that uncertainty of data grows the more we look into the future. On the other hand some decisions have to be prepared beforehand (e.g. negotiations with a new supplier, product introduction, building up of seasonal stocks). Actually, there will be no general answer – a compromise must be obtained for each supply chain and decision level.

Usually, these multi-period plans will not be implemented in total (i.e. re-planning only after T periods). Instead, *rolling schedules* are used: Once the re-planning interval has passed – usually the first period of the planning interval – a new planning period $s+T+1$ will be added, the data of the other periods will be updated taking into account latest information about the future (e.g. demand forecasts), and finally a new schedule will be generated (see Fig. 2.2). If a plan has been generated and confirmed, the first period(s) is (are) declared the *frozen horizon*. There, no changes are allowed in order to secure the implementation of decisions within the frozen horizon – in order to make its preparation effective.

The drawback is that a company will not be able to accept rush customer orders to be produced within the frozen horizon and hence might lose money. Since re-planning is rather easy with APS compared to pure manual planning, the introduction of an APS often results in smaller frozen horizons.

The advantage of rolling schedules is that this is a way to cope with uncertainty even in a deterministic modeling approach.

A next and important issue is to design rolling schedules for the different levels of an HP system. Here, the time frames of rolling schedules may not be designed independently because we have to make sure that the information to be exchanged between adjacent levels is synchronized.

If the length of periods is the same for two adjacent hierarchical planning levels we are done (see Fig. 2.3). The same holds if there is no aggregation of time at the lower planning level (i.e. a continuous time scale like in scheduling).

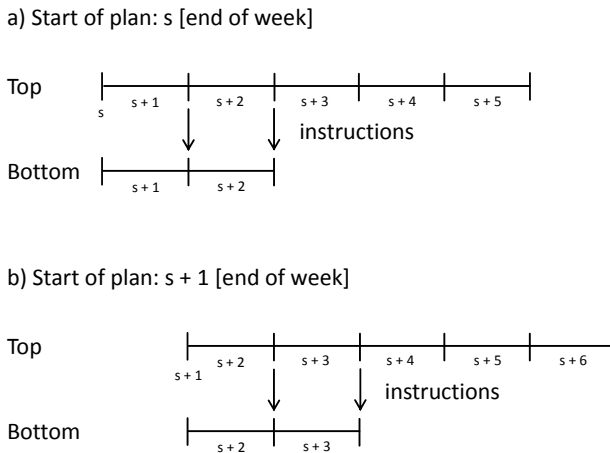


Figure 2.3
Equal length of periods for all hierarchical planning levels, fixed planning intervals

Note that instructions may either be given at the end of each period or only at the lower level’s planning horizon.

Rolling schedules become “tricky” if the re-planning interval (periods) of the subordinate planning level becomes smaller. Actually, this situation is typical because the lower the planning level the greater the level of detail should be (i.e. the smaller the periods). A rolling schedule then may require variable lengths of lower level planning intervals in order to always have instructions at the planning horizon (see Fig. 2.4).

It should have become clear that there is no single way to design an HP system (even if using pre-specified software modules like SAP APO). The overall aim is to derive feasible, good quality solutions for complex decision problems arising in the SC with reasonable efforts. Solutions should be accepted by all the people in the decision hierarchy. One very important feature of HP is that decision makers at all levels will keep some degree of freedom in making decisions but still be in concert with the other decision units (due to coordination).

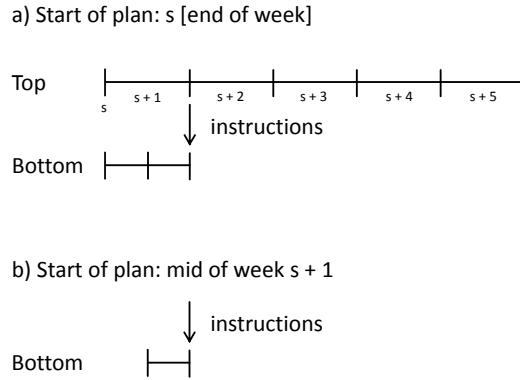


Figure 2.4
 Shorter re-planning intervals on lower hierarchical planning levels, variable planning intervals

HP systems are not limited to APS. There may also be individual solutions as a closer look into the recent literature shows.

Zäpfel and Mitter (2010) develop a tactical-operational HP system with four levels (material requirements planning, personnel planning, process planning, resource scheduling). This system is implemented for a logistics service provider with a rolling schedule for the two upper levels. A special focus is on the flexibility of the workforce and equipment.

A hierarchical production planning system for general supply chain planning with uncertainty can be found in Gebhard (2009). Here, aggregate planning (top level) and lot sizing (bottom level) are both used in combination with robust planning and rolling schedules in order to generate a plan in light of uncertain information.

Volling (2008) implements a two-level tactical-operational HP system for a company in the automotive sector. The system combines a make-to-order environment with a multi-variant batch production. Order taking is addressed at the top level, and master production scheduling at the bottom level. Both levels use rolling schedules.

An HP system to plan production for a company in the railway industry is developed by Timm (2008). The HP system is divided into four hierarchical levels which represent process design and machine capacity, personnel planning, inventory management, and lot-sizing. There are two rolling schedules, one for the two upper and the other for the two bottom levels.

Next, a structured overview of the various planning tasks arising in a supply chain is presented.

2.3 The Supply Chain Planning Matrix

The HP concept allocates the planning tasks to several planning levels. Combining this classification with the decomposition of a company’s supply chain into the sections procurement, production, distribution, and sales, results in the supply chain planning matrix (SCP matrix, see Fleischmann et al. 2008). It provides a useful structure for classifying the various planning tasks along the supply chain.

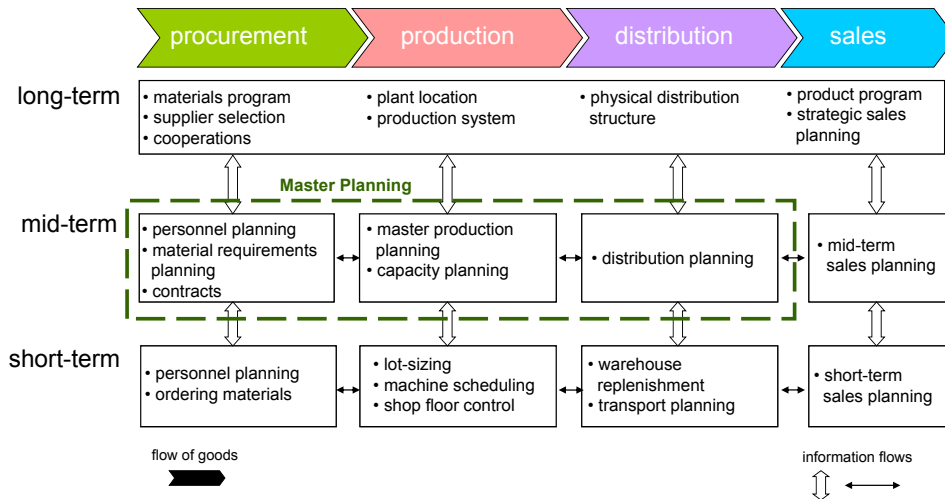


Figure 2.5
Supply chain planning tasks (see Fleischmann et al. 2008)

Figure 2.5 shows the SCP matrix with the supply chain sections as horizontal dimension and the planning levels as vertical dimension. The three levels help to differentiate the characteristics of the planning tasks on the top level, the mid-level, and the bottom level. More detailed classifications with four or more levels are feasible. The vertical information flows correspond to the instructions and feedback information in the HP system. The horizontal information flows serve to coordinate the neighboring supply chain processes. Upstream they contain primarily demand data, customer orders and internal requests, and downstream available capacities, current stock and lead times.

The planning tasks in the different levels do not only exhibit a hierarchical relationship, but also differ significantly in the following characteristics:

- **Planning interval:** The planning interval comprises several years on the top level and decreases to about six to twelve months on the mid-level and to typically a few weeks on the bottom level. Thus, the three levels correspond to the classical distinction of long-term, mid-term, and short-term planning.
- **Type of decisions:** Top level planning supports decisions on the structure of the supply chain, such as locations of plants and warehouses. These strategic decisions are non-recurrent in their content and are often prepared in temporary *strategic planning* projects. Planning on the lower levels concerns regular decisions on the supply chain operations, e.g. production quantities of the next month or the next day, and is called *operational* planning. Moreover, so-called *tactical* decisions occur on the mid-level, which set a mid-term framework for the short-term operations, e.g. contracts with suppliers and service providers.
- **Impact:** Top-level planning impacts the whole company. Mid-level planning concerns a larger part of it, like a whole factory, whereas short-term planning considers only a limited area, like a production department or the operations in a warehouse.

- Level of detail: Top-level planning is highly aggregated and works with rather rough models of the supply chain. The level of detail increases from top to bottom, and on the bottom level, the planning results must be so precise that they can be used immediately as instructions for the execution.
- Responsibility: Decisions on the top level are taken by the company's top management, often supported by a strategic planning department. Mid-term planning and decisions are usually the responsibility of a central supply chain or logistics department, whereas short-term planning and decisions are tasks of local planners. This way, the hierarchical structure of SCP is often reflected in the organizational hierarchy of the company.

The content of the single planning tasks depends on the type of business and on the design of the HP system. The specifications in [Figure 2.5](#) are typical examples.

Top Level

The top-level planning tasks are shown in a common box to emphasize the comprehensive character of strategic planning which always concerns the whole company. The key decision is the product program that the company intends to offer in the future and the estimated sales volumes. This has an impact on the required production system and materials program. The selection of production locations and of suppliers is interdependent, and the distribution system must be appropriate to link the production locations with the intended sales markets.

Mid-Level

Mid-term planning of procurement, production, and distribution establishes quantities for the corresponding supply chain processes and allocates them to “time buckets”, usually months or weeks within the planning horizon. It considers the relevant capacities, in particular in the production system, and decides on capacity adjustments, as far as possible in the mid-term horizon, e.g. adjustments of the working time and of the head count. Mid-term planning is essential for considering seasonal demand fluctuations. It determines the appropriate measures, e.g. adjustment of capacities or build-up of seasonal inventories by pre-production. Note that planning on this level does not consider single operations, like processing a production order or the trip of a vehicle, and hence no sequencing of operations.

Planning in process quantities and time buckets allows an easy combination of the planning tasks in the single sections to a common *Master Planning*. It extends the notion “Master Production Schedule” which originates from MRP. Master Planning is able to coordinate all processes along the supply

chain in a simultaneous planning step. Using Linear Programming as planning technology (see Chap. 5) minimizes the total cost of the supply chain processes so as to satisfy a given sales plan. As an important building block in an HP system, it coordinates the various short-term planning tasks, which cannot be accomplished simultaneously.

Bottom Level

Short-term distribution planning specifies the daily shipments and allocates them to vehicles or it releases the corresponding orders to a logistics service provider. This concerns the transports from the factories to the warehouses, i.e. the *warehouse replenishment*, and the delivery of customer orders. In addition, transports occur as part of the procurement, but there, the transport activities are often controlled by the supplier, as it is true of the Frutado case.

Short-term production and procurement planning comprises the determination of lot-sizes, scheduling the resulting production orders on the machines, and preparing the necessary resources of personnel and materials.

Mid-Term and Short-Term Sales Planning

Operational sales planning plays a particular and crucial role in SCP, because it must provide the current demand data for all other planning tasks. Its main function is forecasting future sales. As the various planning tasks differ in the length of the planning interval and in the level of detail, the forecasting module must be flexible regarding these attributes. In particular, it must provide possibilities to aggregate and disaggregate the forecasts according to the needs of the using planning module, as to the time buckets, product groups, and market regions.

Information on future demands can also be available in form of known customer orders. This is more often the case on the short-term planning level and depends on the position of the decoupling point which will be explained in the following.

Decoupling Point

The decoupling point in a supply chain separates the upstream processes which are forecast-driven from the downstream processes which react to the known customer orders. At the decoupling point, a safety stock is held to account for forecast errors. Common decoupling points are:

- Before the first production stage (Make-to-Order, MTO)
- Before the final assembly (Assemble-to-Order, ATO)
- At the finished product stock (Make-to-Stock, MTS)

Shifting the decoupling point downstream reduces the lead time for the customer orders, but increases the value and the holding costs of the inventory. In the Frutado case, the decoupling point is at the DCs, hence MTS. Note that even a part of the distribution, the replenishment of the DCs, is forecast-driven, and the only order-driven processes are the deliveries to the customers and emergency cross-shipments between the DCs.

Unfortunately, the SCP matrix does not show the position of the decoupling point. Depending on it, the character of the planning tasks, particularly on the short-term level, can be quite different, either forecast-driven with a focus on inventory management, or order-driven with a focus on reliable fulfillment of the customer orders. The latter is called *demand fulfillment*. In case of ATO and MTS, the main bottleneck for demand fulfillment is the stock of components or finished products, respectively, that is “*Available-to-Promise*” (ATP). The term ATP is often used to denote all functions of the demand fulfillment (see Chap. 7). The box in the bottom-right corner of the SCP matrix, the short-term sales planning, can correspondingly be filled out with “ATP” or “Demand Fulfillment”.

2.4 Planning Tasks in the Frutado Case

The analysis of Frutado’s present planning system, as explained in Section 1.2, and the consideration of the planning concepts outlined in this chapter, result in the following requirements which should be satisfied by a new system for mid- and short-term operational planning.

Forecasting

Forecasting should be done in a more systematic and objective way, over a horizon of at least half a year, in order to consider future seasonal peaks, and in a weekly granularity. For short-term production scheduling and distribution, even daily forecasts are required.

Moreover, *safety stocks*, which are necessary to protect against the uncertainty of demands, have to be determined, based on an analysis of past forecast errors, so that they guarantee a desired service level. However, this is not part of the Frutado case. There it is assumed, that safety stocks have already been determined and are given data for the other planning tasks.

Master Planning

The most serious deficiency of the former planning system is the missing overall coordination of the operations in the different functions and locations. The medium-term Master Planning task at Frutado consists in deciding on the allocation of production to the plants and filling lines and on the distribution to the DCs, taking into account the limited capacities of the filling lines and the seasonal demand curves. This way, bottlenecks can be anticipated and the optimal mix of counter-actions – building up stocks, overtime or

cross-shipping – can be determined. This central planning function stabilizes the environment for the following short-term planning tasks.

Production Scheduling

This task has to be done locally in every plant. It concerns decisions on the lot sizes and sequences of the products on the filling lines, considering the setup times and costs, the results of the Master Planning, and the short-term demand forecast.

Distribution Planning

For the planned production quantities of the various products, it has to be decided in which DC to ship the finished products, considering the short-term demand forecast. In order to meet known customer orders, it may also be necessary to use cross-shipping between the DCs. Finally, the deliveries of the customer orders have to be scheduled by assigning them to the vehicles and determining the vehicle routes.

Order Promising

At every order arrival, it has to be checked, whether the order can be satisfied at the desired due date, and a confirmation with a promised date has to be generated.

Finally, the management of Frutado has decided to test the SAP Advanced Planning and Optimization (APO) which satisfies the above planning requirements. Its modules, which are explained in the following chapter, cover exactly the planning tasks of the mid-term and short-term levels of the SCP matrix.

Questions and Exercises

1. Why do we have a planning interval of several time periods if only the first period's decisions are implemented?
2. What is your proposal regarding the aggregation of production coefficients for "Ice Tea - Product 4", if for the DK- and the PL-variant the production coefficients are 10% higher than for Germany (D)?
3. Discuss the pros and cons of aggregation in HP.
4. Which are the means to cope with demand uncertainty in the Frutado case?

Bibliography

- Fleischmann, B.; Meyr, H.; Wagner, M. (2008) *Advanced planning*, in: H. Stadtler; C. Kilger (Eds.) *Supply Chain Management and Advanced Planning*, Springer, Berlin et al., 4th ed., 81–109
- Gebhard, M. (2009) *Hierarchische Produktionsplanung bei Unsicherheit*, Gabler, Wiesbaden
- Hax, A. C.; Meal, H. C. (1975) *Hierarchical integration of production planning and scheduling*, *Logistics: TIMS Studies in Management Sciences*, vol. 1, North-Holland, Amsterdam, 53–69
- Rohde, J. (2004) *Hierarchical supply chain planning using artificial neural networks to anticipate base-level outcomes*, *OR Spectrum*, vol. 26, no. 4, 471–492
- Schneeweiss, C. (2003) *Distributed Decision Making*, Springer, Berlin, 2nd ed.
- Timm, T. (2008) *Ein Verfahren zur hierarchischen Struktur-, Dimensions- und Materialbedarfsplanung von Fertigungssystemen*, Universität Paderborn, Paderborn
- Volling, T. (2008) *Auftragsbezogene Planung bei variantenreicher Serienproduktion*, Technische Universität Braunschweig, Braunschweig
- Zäpfel, G.; Mitter, J. (2010) *Hierarchische Planung für industrielle Logistikdienstleister*, *Zeitschrift für Betriebswirtschaft*, vol. 80, no. 12, 1277–1304

SAP[®] APO - Module Matrix and General Principles

Christopher Sürrie¹

¹ SAP Deutschland AG & Co. KG, Hasso-Plattner-Ring 7, 69190 Walldorf, Germany

This chapter provides an introduction into SAP[®] Advanced Planning & Optimization (SAP APO). First, SAP[®] APO is put in the context of the SAP software portfolio and its modules are characterized. Second, the data flows within the software are described, both from a technical as well as a process-related point of view. Thereafter, the concept of models and versions is described as well as the most important master data elements, that will be used throughout this book. Different types of transactional data are described as well, and an overview of the user interface is given. For more details on SAP[®] APO the reader is referred to a comprehensive documentation of SAP[®] APO which can be found at <http://help.sap.com> (SAP 2011). This website provides documentation on any SAP software in various languages. SAP[®] APO can be found as part of the SAP Business Suite → SAP Supply Chain Management.

3.1 Module Matrix and Related Systems

An overview of applications related to SAP[®] APO that are part of the *SAP[®] Business Suite* are shown in [Figure 3.1](#). SAP Advanced Planning & Optimization (SAP[®] APO) is part of SAP[®] Supply Chain Management (SAP SCM) which is an integral part of the SAP[®] Business Suite that covers a wide range of business software including SAP[®] Customer Relationship Management (SAP CRM), SAP[®] Supplier Relationship Management (SAP SRM) and SAP[®] Product Lifecycle Management (SAP PLM) to name only a few.

SAP® APO primarily covers planning tasks compared to other components of SAP® SCM which support the execution of processes directly like SAP® Extended Warehouse Management (SAP EWM) with a wide range of functionality concerning warehousing processes, SAP® Supplier Network Collaboration (SAP SNC), which supports collaboration processes concerning the exchange of goods and monitoring of inventories, and SAP® Transportation Management (SAP TM) which covers all kinds of processes in transportation logistics (from order capturing to freight billing).

SAP® APO is based on the SCM infrastructure including SAP® Event Manager (SAP EM) which offers track and trace functionality and SAP® Auto ID Infrastructure (SAP AII) to integrate RFID technology to SAP systems to monitor the flow of data and goods within a supply chain.

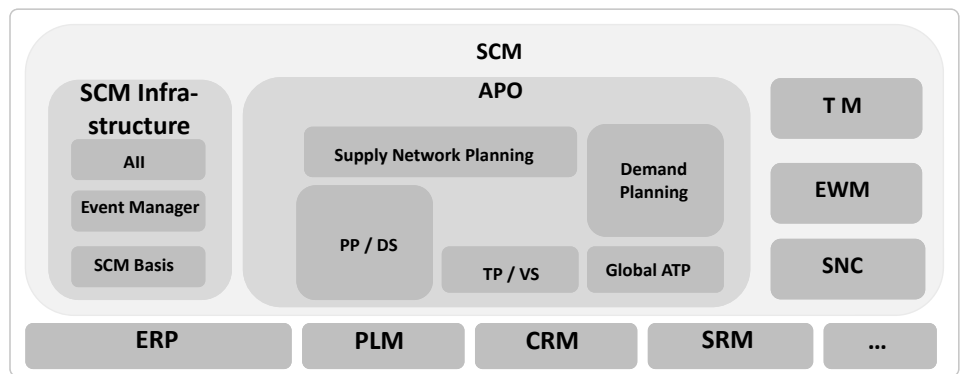


Figure 3.1
SAP® Business Suite

From a technical perspective, SAP® APO itself consists of the following modules:

- Demand Planning (DP)
- Supply Network Planning (SNP)
- Production Planning / Detailed Scheduling (PP/DS)
- Transportation Planning / Vehicle Scheduling (TP/VS)
- Global Available-To-Promise (Global ATP)

The objective of the *Demand Planning* module is to calculate and determine future demand, which will help planners further up in the supply chain to produce, stock, and deploy the right quantity of products / components at the various locations / entities of the supply chain. To achieve this objective the DP module includes a set of *statistical* and *causal forecasting methods*. This means that depending on the individual time series to be forecasted an analysis can be made to assess the underlying demand structure in the past (constant, linear trend, seasonal patterns, ...), and based on that find a mathematical expression to extrapolate this time series into the future. The user is supported in this activity by e.g. automatic assignment of a best-fit-model

or outlier detection procedures. Forecast accuracy is monitored, and the user can be alerted to check or revise the forecasting model, if forecast accuracy falls below a defined threshold.

Furthermore, product lifecycle recognition (e.g. for products with frequent model changes like consumer electronics or mobile phones) as well as promotion planning to take into account effects of past and planned promotions (e.g. television advertising or retailer promotions) are part of DP core functionality. An aggregation / disaggregation logic allows to create forecasts on different hierarchical levels (e.g. aggregation of individual forecasts of sales representatives for their individual customer accounts). Finally, collaborative planning features can be used to allow internal and external stakeholders to be part of a joint forecasting effort.

Master Planning is covered by the *Supply Network Planning* (SNP) module. Its objective is to calculate quantities to be produced and delivered to the various locations that are part of the supply chain to match (final) customer demand and maintain desired service levels. The SNP module provides an overview of the complete supply chain and integrates purchasing, production, distribution, and transportation decisions. By taking a supply chain wide view, the SNP module supports inventory planning within the supply chain considering constraints and penalties (e.g. maximal inventory levels or storage costs) and therefore plans the product flow from the source (e.g. raw materials) up to the final customers (e.g. finished products). Similar to DP, SNP offers medium-term planning on different levels of detail (aggregated planning).

The SNP module offers basically three different planning algorithms: heuristic, rule-based, and optimization-based. While the heuristic algorithm is merely an unconstrained demand covering algorithm that picks the first valid element (e.g. inventory position, production in same location, transportation from an upstream location) to cover a demand, the rule-based algorithm (CTM - capable to match) is much more sophisticated. In comparison to the heuristic algorithm, CTM can consider capacities (e.g. production capacity) and performs backtracking to find a feasible solution to cover a demand. Finally, the optimization-based algorithm is the most sophisticated algorithm offered by SNP. It does not only provide a feasible solution like CTM, but seeks the best feasible solution based on a cost objective including numerous cost elements like production cost, transportation cost, inventory cost, ... The optimization engine is based on an exact algorithm, that (if time permits) will return the optimal solution of the supply chain planning problem. In addition, the SNP module offers advanced safety stock methods considering multilevel supply chain networks and demand variability.

While from a process perspective deployment is a separate process step, technically all deployment functionality is part of the SNP module. The objective of *Deployment* is to match supply and demand in a short-term horizon. If all forecasts materialize into sales orders, and all production orders are executed as planned, deployment would not be necessary. However, in

the real world, sales orders deviate from forecasts, and production output deviates from planned production. Therefore demand and supply do not match and need to be synchronized in the short-term horizon. Deployment helps to decide which demands have to be met (in case of short supply) and what to do with excess supply (in case of short demand). Deployment can be done with a rule-based heuristic algorithm taking into account priorities and quota arrangements or using an optimization algorithm similar to SNP.

The *Production Planning / Detailed Scheduling* module consists of a production planning (PP) part and a detailed scheduling (DS) part. PP delivers a short-term plan that matches overall supply to demand by creating e.g. planned production orders or purchase requirements to cover all demand elements. In this step lot sizes are determined without taking resource capacities into account. This is done across multiple levels and also provides order pegging, which is a concept to establish a link between requirement elements and receipt elements that will be explained later in more detail. Algorithms (called heuristics in APO terms) are delivered as standard to support these processes but also own developed algorithms can be added easily.

The DS part determines a (near) optimal production sequence for execution to meet delivery commitments based on actual constraints on the shop floor. This includes the best-usage of available resource capacities as well as optimal sequencing based on e.g. setup considerations. This can be achieved by a wide range of options. The simplest way of course may be manual (user-based) planning in a flexible graphical detailed scheduling planning board. Additionally, several heuristics are offered to cope with scheduling problems, while the most sophisticated option is a powerful planning engine based on a genetic algorithm.

PP/DS offers what-if analyses and deterministic simulation (e.g. resource breakdown, different demand situations ...) and provides the user with a configurable exception alert monitor. Furthermore, PP/DS enables special planning processes tailored to industry-needs: Planning with characteristics is offered for mill industries, production campaign planning is offered for chemical / pharmaceutical industries, planning with shelf-life is offered for food industries, sequencing and model-mix planning on production lines is offered for automotive industries ...

The objective of the *Transportation Planning / Vehicle Scheduling* module is to plan and optimize shipments for orders (sales orders, purchase orders, returns, and stock transport orders) and deliveries. Due to its tight integration to SAP® ERP and Global ATP it is a commonly used transportation planning solution for shippers. TP/VS allows to plan shipments that are inbound (e.g. based on purchase orders), that are intra-company (e.g. based on stock transport orders), and those that are outbound (e.g. based on sales orders). TP/VS includes two optimization engines. The first one covers vehicle scheduling and routing functionality. This means that orders or deliveries are consolidated on vehicle resources based on their attributes and proximity of

locations, such that tours and a delivery schedule result that fulfills a variety of constraints (e.g. vehicle capacity, loading and unloading capacity, opening hours ...). While this first planning engine selects the means-of-transport (i.e. vehicle type like train, truck or van), the second optimization engine can be used (provided the user of TP/VS does not possess an own fleet) to select the *transportation service provider* (TSP) to execute the shipment. If several TSPs offer the chosen kind of truck, the TSP selection optimizer will find the right one based on real (!) costs (provided freight costs have been set up in ERP) and at the same time considering contractual obligations like business shares or allocations. The planning result of TP/VS are so-called planned shipments that can be transferred directly to ERP shipments for execution. Furthermore, TP/VS offers (web- and EDI-based) collaboration features allowing the TSP to accept or reject assigned planned shipments.

The objective of *Global Available-to-Promise* (Global ATP) is to provide online information about the most recent state of a plan and to allow order promising that will execute according to the customers' expectation. A typical use case is to quote / promise a delivery date to a customer (order) based on the current inventory positions and production plan of the supply chain.

The Global ATP module supports rule-based ATP which means that defined rules are checked to fill a specific customer demand, e.g. if a product is not available in a certain location, the upstream location is checked or an alternative product is proposed. Furthermore, multi-level ATP can be done, which means that the availability of components is checked to derive the promised date for the customer. The Global ATP module is integrated with PP/DS to check against the actual production plan and consider capacity constraints. Furthermore, an integration with DP is available to consider product allocations.

Considering that in typical customer scenarios many (Global)ATP-checks need to be performed and a reasonable runtime is to be expected, Global ATP relies on the liveCache (3.2).

Questions and Exercises

1. Describe the five modules of SAP® APO (DP, SNP, PP/DS, TP/VS, and Global ATP). In which time horizon are they primarily used and which planning tasks do they support?

3.2 Data Flows (Technical and Process-Related)

With respect to data flows between the SAP® APO modules as well as data flows to and from external systems one has to distinguish between different views, a more technical system architecture related view and a business process driven view. Even within the individual modules there exist data flows from a system architectural point of view.

Figure 3.2 shows how SAP® APO (being part of SAP® SCM) is embedded in a system landscape. The heart of SAP® APO is the application server depicted in the center of Figure 3.2. All data is kept in a database layer (shown at the bottom) while the presentation layer containing the user frontend is using SAPGUI (SAP Graphical User Interface) technology to be represented e.g. in the SAP Enterprise Portal or via the Netweaver Business Client (NWBC).

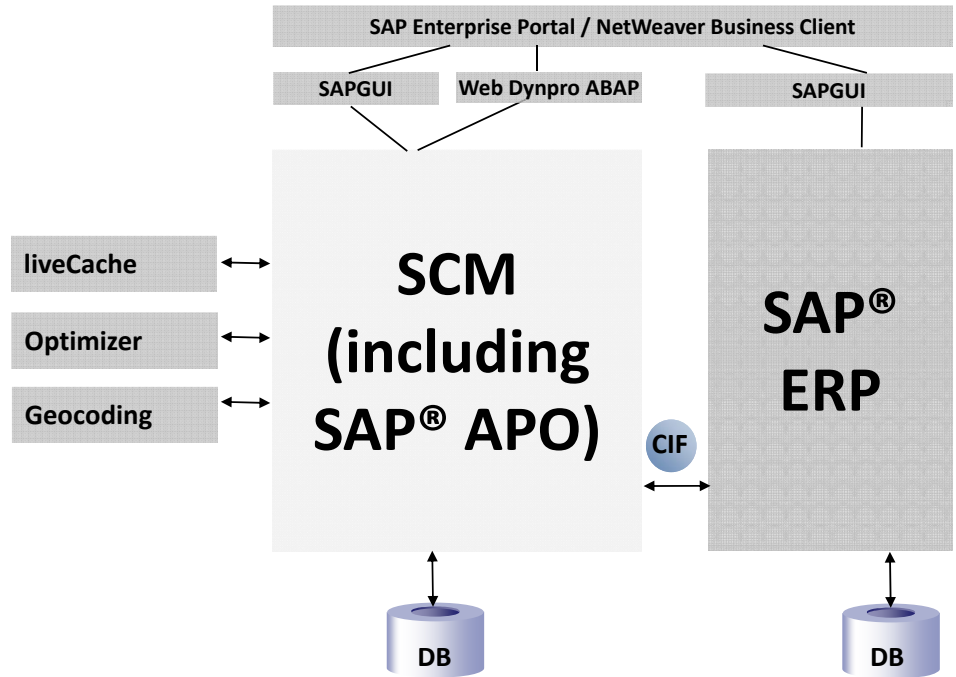


Figure 3.2
SAP® SCM server
architecture

To deal with huge data loads and performance critical planning activities, SAP® APO transfers performance-critical application logic to the liveCache instance, where the data is kept persistent. This allows fast access to all application data which can be cached completely in main-memory and is organized in relational but mainly object oriented manner. The liveCache is a high performance, memory resident data processor based on SAP DB (Database) technology. This enables SAP® APO to perform the planning tasks where the data is located ensuring high performance.

The optimization engines of SAP® APO also reside on a separate server. They operate in an asynchronous mode which means that they are usually started for a defined runtime which they use to collect relevant data from the liveCache and / or database via the application server, to generate a suitable (optimal) solution to the planning problem, and afterwards to return the result to the application server, while the user can continue with his work.

A geocoding server and software may also be part of the system landscape, especially if transportation planning is in scope of the supply chain operation. Geocoding is used to locate addresses and determine the exact road distance

between addresses, which is obviously a critical master data element, if vehicle routing is part of the planning problem.

From a business process perspective point of view, it is important to be aware of the fact, that SAP® APO is a planning system and all transactional order data (from sales orders representing customer demand to purchase orders representing raw material orders to suppliers) are kept in the ERP system as leading system. Between SAP® ERP and SAP® APO there exists a standard interface called CIF (*Core Interface*).

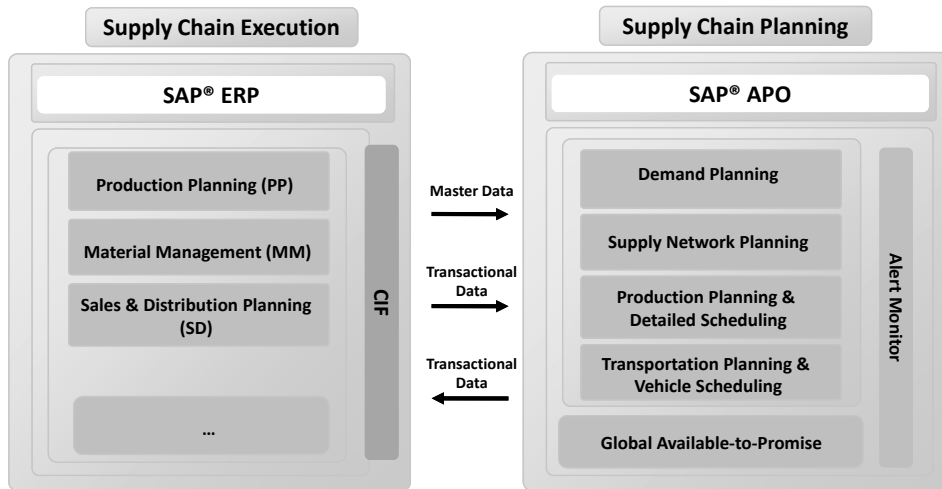


Figure 3.3
Core interface (CIF)

Figure 3.3 explains the CIF in more detail. CIF is responsible for both, the exchange of master data as well as transactional data. The setup of the interface is relatively easy. In SAP® ERP a so-called integration model is defined. The integration model is basically a selection of data elements that are relevant for replication to SAP® APO and holds the information of the target system (e.g. if multiple SAP® APO systems are connected to the same SAP® ERP system). Data elements may be master data objects or transactional data elements. Both are selected by their object type (e.g. material) and additional selection criteria (e.g. materials that are used in a specific plant).

For *master data*, ERP is the leading system. This means that all relevant changes (e.g. address changes, bill of material changes) need to be made in the ERP system and will be replicated to SAP® APO. The data transfer for some master data objects can be configured to be instantly, while for most master data objects a periodical transfer is standard. Master data objects that are replicated from SAP® ERP to SAP® APO include

- Locations,
- Products,
- Resources, and
- BOM and Routings.

For *transactional data* the transfer in both directions is continuously in real time. Transactional data elements that are transferred from SAP® ERP to SAP® APO include

- Sales orders,
- Production orders,
- Confirmations, and
- Stock information.

Transactional data elements that are transferred in the opposite direction, i.e. from SAP® APO to SAP® ERP include

- Planned orders
- Stock transfer requisitions (or orders), and
- Shipments.

Having explained the technical / architectural view on the integration of SAP® APO to its ecosystem the process view which is even more important from the business perspective shall be focused at. As there are numerous business processes that are supported by SAP® APO using different business objects or documents, one TP/VS process for outbound transportation is picked as an example (see Fig. 3.4).

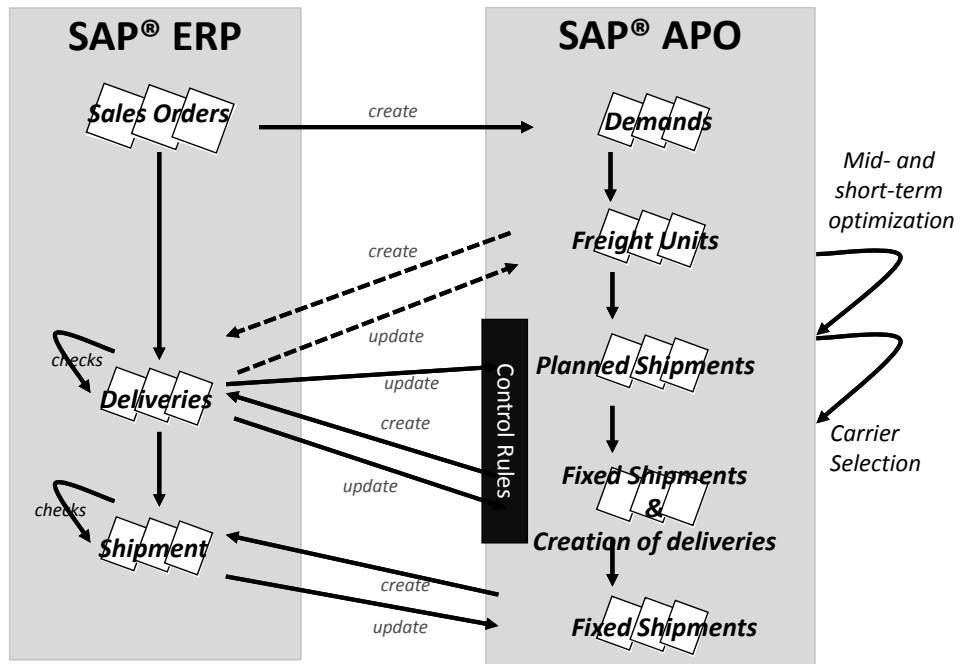


Figure 3.4
Transportation
planning – outbound
process

The relevant documents or business objects on ERP side are sales orders, deliveries, and shipments. Sales orders represent a customer demand for one or multiple materials in a certain quantity at a certain location (ship-to party). Not all items of a sales order may be shipped together to the customer. Therefore, several deliveries can be created from the original sales order. One or more deliveries (originating from one or more sales orders) can be consolidated into one shipment, which is the document representing e.g. the individual tour of a truck. The same delivery may be assigned to multiple shipments, e.g. in case of a multi-modal transportation chain. The shipment is the execution document for the transport that keeps the routing information (stages of a shipment) and the carrier, if the shipment is not executed by an internal fleet.

How does SAP® APO integrate into this execution process and document flow? Sales orders are transferred via CIF from SAP® ERP to SAP® APO and represent transportation demands in TP/VS. Very often it is hard and also not even desired to transport a complete sales order (e.g. because different items of the sales orders have different required delivery dates or the full quantity of the sales order does not fit on any available vehicle resource). Therefore, transportation demands are split (using split rules) into freight units. Freight units represent transportable units and are the logistical unit that is not split into smaller units throughout the complete journey from their source (e.g. plant) to their final destination (e.g. customer).

Freight units are the basis element for any planning activities within TP/VS. Planning activities (may they be manual, heuristic-based or using the optimization engine) assign freight units to available vehicle resources (which can be trucks, railcars, ships, airplanes or any other means-of-transport), which results in planned shipments. These planned shipments can be reviewed by the planner. Before these planned shipments can be published to ERP, TP/VS must trigger the delivery creation in ERP based on the TP/VS planning result. This step is necessary, because the ERP system requires deliveries to assign them to a shipment, as sales orders cannot be assigned to a shipment directly. This means that deliveries are created based on the freight units that are assigned to a planned shipment. These deliveries then update the freight unit, such that they do no longer reference to the sales order but to the delivery. Now the planned shipments can be fixed, and ERP shipments can be created based on the fixed planned shipments.

Generally, transportation-related changes made to an ERP document are updated in the corresponding TP/VS document. This means that subsequent changes on the ERP documents (delivery or shipment) are replicated to SAP® APO such that the current situation is always visible. The update from ERP documents to SAP® APO is controlled by so-called control rules. These can be used to define actions in case a change of an ERP document leads to a non-desired situation (e.g. an increase of the delivery quantity results in a resource overload for the shipment).

In the real world, master data and transactional data have their origin

in an ERP system. A sales order impacts demand planning by consuming forecasts. SNP and PP/DS use sales orders as requirement elements for planning and scheduling of planned (production) orders. Deployment uses sales orders for making the right assignments of supply to demand, and TP/VS plans shipments to deliver to customers according to their sales orders on-time. Consequently, integration between different systems (e.g. ERP and SAP® APO) is a vital part of any supply chain planning project. However, the focus of the Frutado case is on planning concepts and its modeling in a supply chain planning software. Therefore, integration is not in scope of this book and rather an exemplary stand-alone SAP® APO solution is described in the subsequent chapters of this book to reduce complexity.

Questions and Exercises

1. Describe the server architecture of a SAP® APO installation. What is the purpose of the individual components?
2. Which data is the basis for planning? Where is the origin of the data and how is the data kept consistent?
3. Describe the flow of information between an ERP system and SAP® APO for an outbound transportation planning process.

3.3 General Terms and Principles

3.3.1 Models and Planning Versions

Master data and transactional data are structured within SAP® APO into models and versions. The concept of *models* and *versions* allows the user to simulate master data or transactional data changes and assess their implications on the supply chain.

The general idea is to define model dependent master data that means if master data modifications have to be analyzed different models should be created while transactional data is defined (planning) version dependent. Each model can have several planning versions assigned to it that means that for each master data set (supply chain configuration), several transactional data sets (e.g. different demand situations) can be analyzed.

As SAP® APO is the planning system and ERP is the leading system containing the “real” master data and current transactional data, there can be only one model and planning version being defined that is integrated to ERP and filled via the CIF. This model and planning version both are called “000”. Planning version “000” of model “000” can therefore be used as a master and copied to other models or versions to allow for simulations. While master data created via CIF is automatically assigned to model 000, master data that is created in SAP® APO needs to be explicitly assigned to a model before usage.

In addition to planning versions there exists also the concept of transactional simulation versions. This concept is being used e.g. when the optimization engine in PP/DS is run. While the optimizer runs on its (separate) server, the application system cannot be stopped. Therefore, the optimizer is run in a transactional simulation (local copy of the relevant data), and once the planning result has been determined and accepted by the decision maker, this transactional simulation is merged with the planning version again. That way, it is possible to have users continue to work with the system (same model and planning version) in parallel and e.g. confirm executed operations in ERP which is the information that is passed to SAP® APO via CIF. During the merge operation rules are applied to assess which change to the data is more relevant (e.g. confirmation of an operation in ERP overrules a rescheduling of the operation made by the PP/DS optimizer).

3.3.2 Master Data

For all planning tasks master data quality is crucial. Production planning will be inaccurate if standard times for operations or resource availability and resource efficiency do not match reality. The same holds true for distances and duration in transportation planning. Therefore, *master data* plays an important role in supply chain planning, and this chapter will explain which master data objects are available to model the supply chain in SAP® APO. However, only the most important objects can be covered here that have relevance to the Frutado case:

- Location
- Resource
- Product
- Hierarchy
- Transportation Lane
- Quota Arrangement
- Production Process Model (PPM) / Production Data Structure (PDS)
- Product Storage Definition
- Setup Matrix
- Schedule
- Characteristic Value Combination (CVC)

Locations in SAP® APO are merely addresses. They originate from different master data objects in ERP (e.g. plants, vendors or customers). Although each location has a location type in SAP® APO, this has (almost)

no relevance for its further usage. Locations are important to model the (geographical) flow of goods, but can also be used to model a (logical) flow of goods (e.g. plant and distribution center at the same geographical location). Especially for TP/VS, the location is an important object. It keeps assignments like the handling capacity or opening hours for inbound and / or outbound shipments and of course provides geographical information important for vehicle routing decisions. To reduce master data maintenance efforts a special location type called transportation zone is available. *Transportation zones* represent (geographical or logical) groups of locations (e.g. ZIP code areas). Individual locations can be assigned to a transportation zone, which allows in a subsequent step to define vehicle resource availability on transportation zone level instead of location level.

As planning often deals with the assignment of scarce resources to operations, *resources* are a central master data object. SAP® APO knows different resource types and resource categories. The most important resource categories are production, transportation, and handling. The most relevant resource types in this case study are single, single-mixed, multi, multi-mixed, vehicle, and transportation resources. Single and multi refers to the number of operations that a resource can execute in parallel. Single resources can only be occupied by one operation at a time and are used to model machines with setup decisions, if machines can have only one status at a time, and this status is defined by the current operation being executed or planned at this resource. In contrast to this, multi resources can be used by several operations at once. Multi resources can be used to model a pool capacity (e.g. personnel). While single and multi resources are used for continuous scheduling in PP/DS, single-mixed and multi-mixed resources can be also used by bucket-based planning applications like SNP. Transportation resources are used to restrict transportation capacity in bucket-based planning like SNP, whereas vehicle resources represent individual vehicles that are routed and scheduled on a continuous time scale.

Resources have capacities in different dimensions. A single resource has capacity “1” and a time availability (e.g. workdays from 9am to 5pm), while multi resources have capacity “n” and also a time availability. Additionally, resources can have storage properties (e.g. tanks), that means they have a max fill level (e.g. 5,000 l). For vehicle resources up to 8 storage dimensions can be defined at the same time (e.g. mass, volume, and footprint). That way, different properties of products can be respected during planning: Light products will hit the volume limit, while heavy products will hit the weight limit, and bulky products may hit the footprint limit.

Resource networks can be defined to model a pre-defined flow of material. For example if a softdrink manufacturer has three production resources and three bottling lines, but not all production resources can feed the product into all bottling lines. Resource networks can be used to define the allowed flows of products.

The *product* originates from the material in SAP® ERP (if transferred via CIF). In product master data physical and logical properties (e.g. dimensions like mass or volume and attributes like transportation group) of the product are stored as well as a multitude of planning relevant parameters. A product contains “global” data and “location-dependent” data. Cost parameters are a typical example for location-dependent planning parameters. Products can have different storage costs at different locations in the supply chain to push the product to a desired storage location. Other planning parameters are the procurement type that tells the system, whether the product can be produced in this location, whether it has to be externally procured or whether both is possible. Furthermore, lot-sizing procedure parameters, shelf-life parameters, and information on safety stock methods are kept in the product master data.

Hierarchies are used in SAP® APO to

- allow for aggregation / disaggregation procedures,
- reduce master data maintenance effort, or
- model special scenarios in transportation planning.

A hierarchy can be defined for products to assign them to product families or product groups. The transportation zone hierarchy has already been mentioned in the paragraph on locations as a means of grouping locations to transportation zones to reduce master data maintenance effort. A rather particular hierarchy is the hub hierarchy which needs to be defined for special transportation planning scenarios. In a hub hierarchy, locations (or transportation zones) are assigned to a (hub) location. This means that a location can be procured either directly or via the hub (depending on the setup of transportation lanes). Characteristic to a hub is that the means-of-transport in a transportation chain can only be changed at a hub. For example if an outbound delivery needs to be planned from the plant in Germany via the port in Hamburg to a customer in the US, then both ports (Hamburg and the port in the US) have to be defined as hubs, and the plant and customer location have to be assigned to their respective hubs in the hub hierarchy, because the means-of-transport changes on the different stages of the transportation chain (e.g. rail from plant to Hamburg, ship from Hamburg to US, truck from US port to customer).

Transportation lanes are one element to represent the geography in SAP® APO. Transportation lanes store the information about the distance as well as the duration between locations. Distances and durations are means-of-transport specific, because a rail transport will usually take more time between two locations compared to a road transport. Transportation lanes keep the information which means-of-transport is available on this explicit relationship. In case the transportation service provider (TSP) selection is relevant, choices have to be made regarding which TSP offers which means-

of-transport. Furthermore, planning relevant information like business shares of TSPs are stored in the transportation lane.

As indicated above transportation lanes are often not maintained on location level, but rather on transportation zone level. In this case the transportation lane represents more a template, because the distance and duration will be different for any pair of locations of the respective transportation zones. In this case, geocoding will provide the real distance and duration for each pair of locations to allow for accurate routing decisions. This information is usually buffered in SAP® APO for performance reasons to avoid a call to the geocoding server for each pair of locations in a planning run. From the transportation lane “template” means-of-transport provided by each TSP are obtained.

The *quota arrangement* is a master data object that is used if no dynamic split but rather a static split of required quantity from a product is to be sourced (incoming quota arrangement) or distributed (outgoing quota arrangement) from or to different locations. It indicates the percentage of a required quantity of a product that is to be procured from a source of supply or delivered to a location and therefore specifies which part of a product quantity split is to be assigned to a transportation lane.

Bill of material information (BOM) as well as routing information for production purposes is stored in one object in SAP® APO. Currently two master data objects are offered for this purpose:

- *Production Data Structure* (PDS)
- *Production Process Model* (PPM)

PPMs being the older object type can be created and edited in SAP® APO, while PDSs cannot be maintained in SAP® APO, but have to be created via CIF from ERP. This follows more stringently the concept of establishing ERP as leading system for master data and allowing only the planning relevant enhancements in SAP® APO that are not needed in ERP. However, as explained previously, for the purpose of this book an exemplary standalone SAP® APO solution has been created. Therefore, PPMs will be used in the Frutado case but not PDS.

Besides this both objects store almost the same content. [Figure 3.5](#) displays a sample PPM. In the graphical overview one can identify the most relevant information that can of course be accessed and edited in detail by either clicking on one of the objects in the graphical overview or by using a tree-like navigation structure. The displayed PPM includes the following elements:

- Bill of material
 - Final product (I_100, top left)
 - Two components (I_W00, I_C00, bottom left)

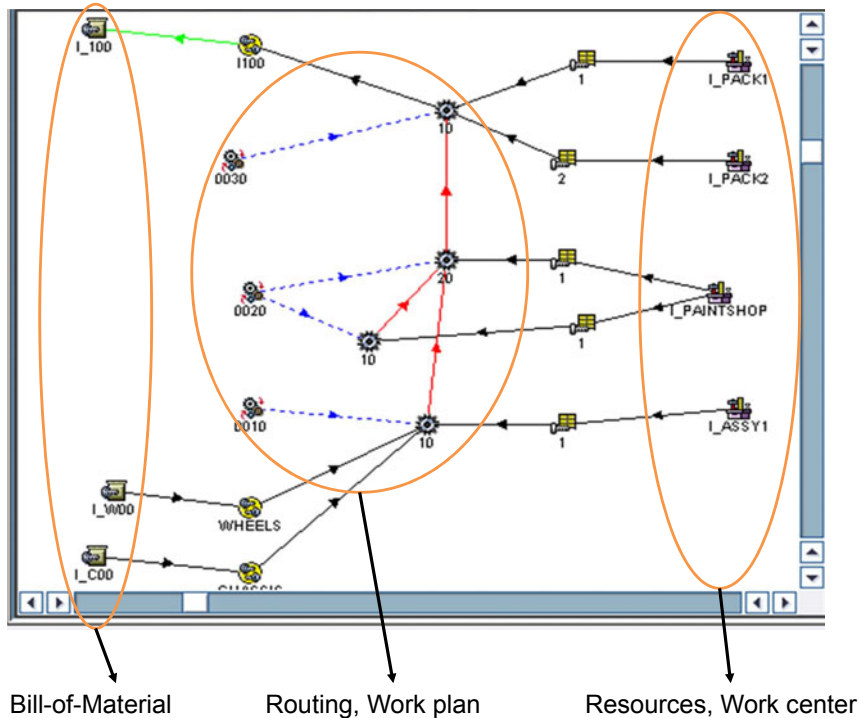


Figure 3.5
Production process model (PPM)
© Copyright 2011. SAP
AG. All rights reserved

- Routing
 - Operation 0010 Activity 10 (bottom center)
 - Operation 0020 Activities 10 and 20 (middle center)
 - Operation 0030 Activity 10 (top center)
 - Precedence links between the activities (red arrows)
- Resource information
 - Operation 0010 requires I_ASSY1 (bottom right)
 - Operation 0020 requires I_PAINTSHOP (middle right)
 - Operation 0030 requires either I_PACK1 or I_PACK2 (top right)

PPMs are very powerful objects and influence the planning result to a great extent. Their accuracy and level of detail is crucial for any production planning project using SAP® APO. Although PPMs can be created on a very detailed level, this should only be used up to the required level to avoid complexity for both the user of the planning system as well as the system itself considering that solving planning problems like this is a hard task also from an algorithmic perspective.

Product storage definitions are relevant for tank scheduling scenarios. They enhance the PPM/PDS by storage relevant information. For any input or output product of a PPM/PDS it can be defined whether the product is filled into or drained from a tank and which tank(s). Furthermore, it defines

whether the PPM/PDS consumes the product from the tank at the beginning or end of the filling / draining activity or whether there is a continuous consumption.

Another relevant master data object for production planning are *setup groups* and *setup matrices*. Setup groups are basically attributes of production operations. For example there are two products, a yellow and a blue one. Then there will exist two PPMs/PDSs in the system with the yellow and blue products as output component respectively. The painting operations of the respective PPM/PDS will have assigned the setup group yellow or blue. To model setup times and / or setup costs for PP/DS, a setup matrix is needed defining the corresponding values.

Schedules are a master data object relevant for TP/VS. A schedule represents a regular transportation option between different locations at defined times. Schedules are defined by an itinerary, which is a sequence of locations to be visited and a departure calendar that defines when the scheduled resource departs from the first location of the itinerary. Schedules can be used to represent sailing calendars of ships or truck shuttle operations between distribution centers / warehouses of an enterprise.

Forecasting in DP is based on so-called *Characteristic Value Combinations* (CVC). It is the combination of characteristic values which serve as a basis for demand planning. CVCs need to be created explicitly, because usually not all possible combinations of characteristic values are required. Think about a scenario with five products and five sales organizations. Not each product is sold by each sales organization, so only a subset of 25 possible CVCs need to be created in SAP® APO.

3.3.3 Transactional Data

Technically, *transactional data* is organized in SAP® APO as

- key figure (time series),
- order, or
- quantity (ATP time series).

The reason for this is the usage of the information in the respective planning modules. DP and SNP are planning modules covering a longer planning horizon than PP/DS or TP/VS. They work based on buckets. This means that the time horizon is separated into several non-overlapping buckets (periods). Bucket profiles can be defined with different bucket sizes (day, week, month, year) or any telescopic combination thereof. Key figures for different purposes (sales orders, forecasts, planned production ...) are defined and store the quantity for each bucket. DP and SNP usually work on the granularity of location-products. This means that there will be one key figure time series for each of the purposes for each product at each location.

Planning is then based on these key figure time series, which means that for each bucket e.g. a production quantity is planned in SNP.

Figure 3.6 shows the data structure for DP. Planning object structures build the basis. In the planning object structure characteristics are defined (e.g. location and product). However, not all possible combinations exist (i.e. not all products are available at each locations), the allowed combinations for planning are stored as characteristic value combinations (CVCs). A planning object structure can be used in several planning areas. In the planning area the key figures (e.g. forecast, production, projected stock) and their properties (e.g. unit of measure, time bucket profile) are defined, and for each key figure and CVC a time series is stored. For each planning area one or more planning books which represent the user (data) selection with one or more data views are defined. In a planning book the interaction with users takes place (see Fig. 3.9). Planning books select a subset of key figures from the planning area and a subset of characteristics from the planning object structure. The different data views help the user to structure the information that is available in the planning book to fit their needs (e.g. one data view to display the regional values of the forecast (disaggregation) and one data view to compare forecast with planned production to allow checks for feasibility). Planning books do not actually store the data, but these are kept in planning versions (see 3.3.1). Thus by changing the data selection in the planning book, the user is able to analyze different scenarios (planning versions) using the same screen layout (planning book/data view).

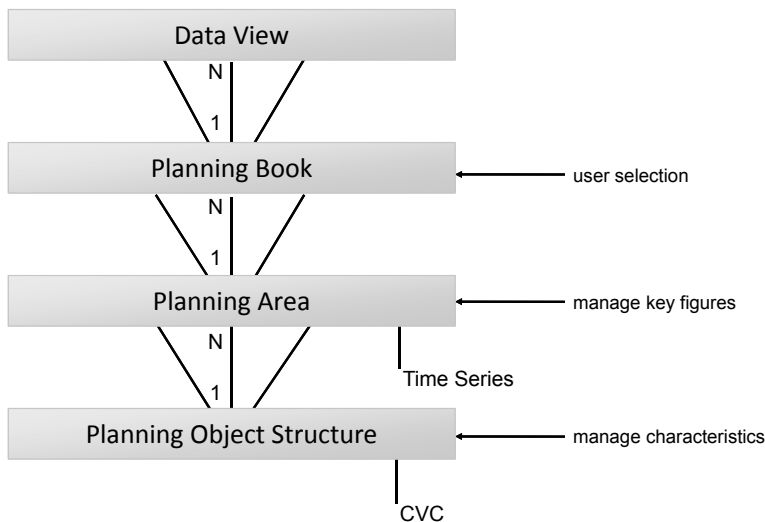


Figure 3.6
Data structure for demand planning

On the other hand, PP/DS and TP/VS are order-based. This means that there is a specific data record for each (planned) production order or customer sales order. Orders have a category which allows to group or select them category-specific in different planning circumstances. Examples for order categories are:

- Purchase Orders (Receipt)

- Purchase Requisitions (Receipt)
- Production Orders (Receipt)
- Planned (Production) Orders (Receipt)
- Sales Orders (Requirement)
- Scheduling Agreements (Requirement)
- Dependent Demand (Requirement)
- (Unrestricted) Stock (Stock)
- Safety Stock (Stock)
- Forecast (Forecast)

There are four order category types (in parentheses, above) that determine the possible usage of the individual orders. The different granularity of transactional data information allows for different planning details with reference to the module used. In PP/DS there may be four different sales orders for next March, while in SNP (assuming a monthly bucket profile) only the total quantity (sum) is displayed in the respective key figure time series.

Planning based on key figure time series compared to order based planning constitutes some fundamental differences not only with respect to planning granularity, but also with respect to the presentation of data to the user and the algorithms being used for planning.

For order-based planning *pegging* is an important concept being used to establish links between the receipt (e.g. production order) and the requirement elements (e.g. sales order). [Figure 3.7](#) shows an example, in which all receipt elements (boxes, production orders) are assigned to requirement elements (triangles, sales orders).

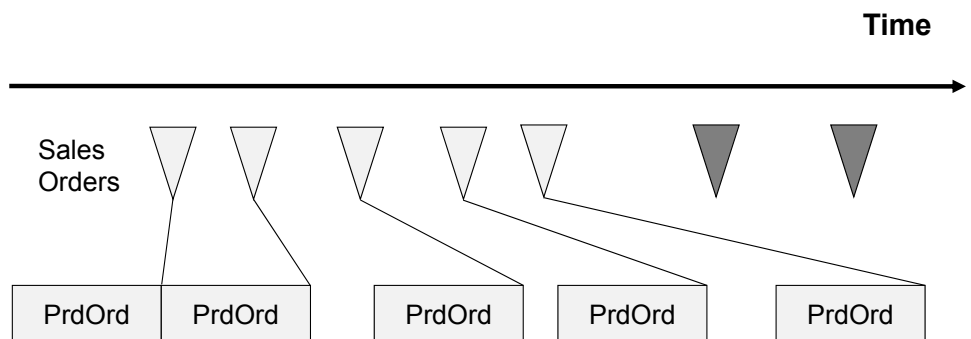


Figure 3.7
Pegging

Pegging can be dynamic or fixed depending on the planning situation. Dynamic pegging has the advantage that receipt elements are re-assigned to other requirements elements automatically, e.g. if the corresponding demand

is changed (quantity or due date). Fixed pegging on the other hand is easier to track in complex planning situations. If sales orders have different priorities these are inherited to the (planned) production orders during their creation. Later in planning when the sequence of the orders changes, dynamic pegging may result in high priority production orders being dynamically pegged to low priority demands and vice versa. However, both concepts may be combined using heuristics to fix and unfix pegging in the production planning run. Furthermore, for dynamic pegging, two different pegging strategies are offered and can be selected per (location) product. The “FIFO” strategy operates based on the “first in, first out”-principle, whereas the “use latest receipt” strategy pegs a requirement element to the latest (in time) available receipt element in case of excess supply (exception: stock is always pegged).

Finally, the Global ATP module relies on ATP-Time Series. In an ATP time series, the data elements (receipt, requirement or stock) are stored in an aggregated form. The aggregation is on a daily basis. The storage of data in key figure (time series), order-based, and ATP time series leads to some redundancy, but is done for performance reasons.

3.3.4 User Interface

The SAP GUI is needed to log on to and to use SAP systems like SAP® APO. [Figure 3.8](#) shows how SAP® APO is presented to the user after log on. A role-based authorization concept allows a user to work only with those transactions and data that are released for him. All available transactions are presented to the user in a tree-like navigation structure. The user can drill down and open a transaction either via double-clicking on the respective line or by entering the transaction code in the field on top of the menu. Users can build their individual navigation structure by creating a favorites folder. SAP® APO is customized using a similar tree-like navigation structure which is usually not available to the standard user.

Planning Books that are used in DP and SNP have a look and feel similar to the SNP planning book shown in [Figure 3.9](#). On the left side of the screen the selected objects (characteristic value combinations) are shown (top). In the screenshot the selection is based on products and locations. Selections can be made individually or predefined per user. A list of predefined selections of the planner is also shown on the left (middle). On the lower left side of the screen the user can select the planning book and data view he wants to work with.

Depending on the *data view* and selection made on the left side of the screen, the screen layout on the right side is determined. Key ingredient of the right side is the table of key figure time series that are filled with the selected data (interactive planning grid). Key figure time series can be read-only (e.g. customer orders) or read/write (e.g. planned production) which allows for direct (manual) planning activities. Also automatic planning activities like heuristic planning, CTM or the SNP optimizer can be started from this screen for the current selection.

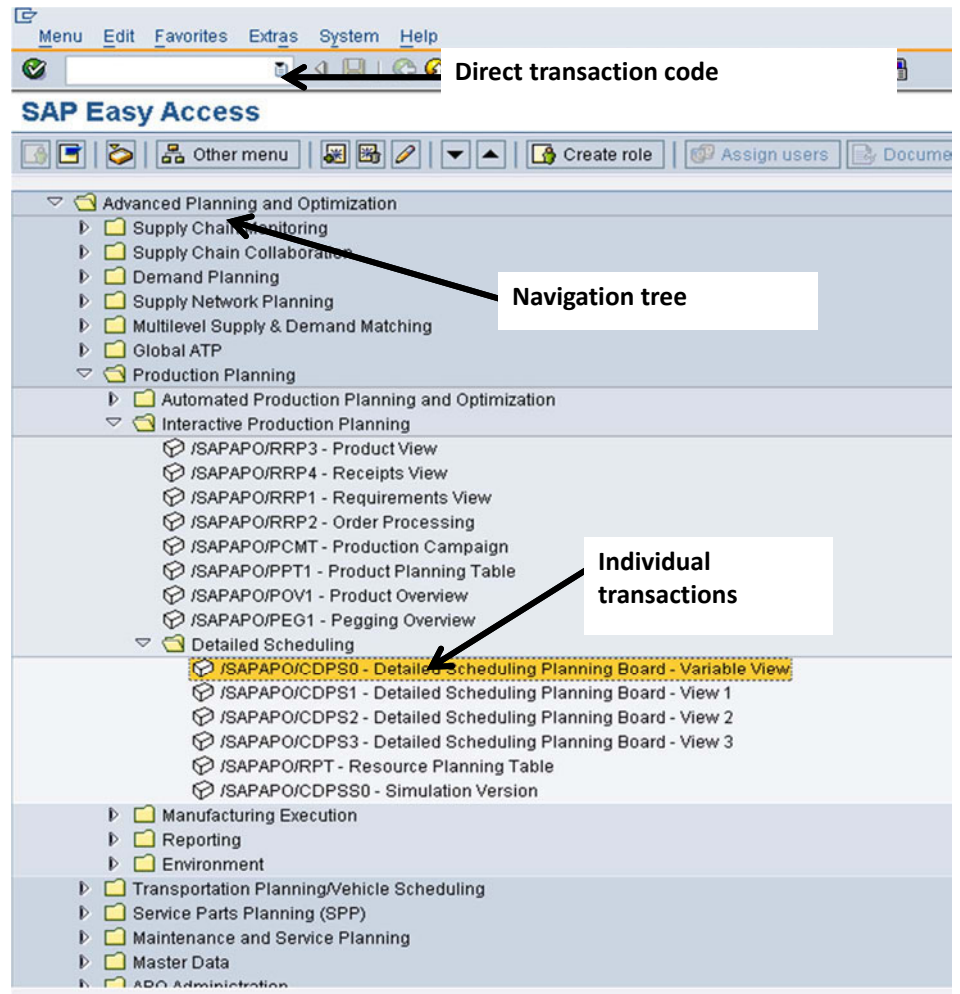


Figure 3.8
SAP® APO navigation
© Copyright 2011. SAP
AG. All rights reserved

PP/DS offers a similar table-based planning view (see Fig. 3.10) as well as a graphical detailed scheduling planning board (see Fig. 3.11). In Figure 3.10 the product planning table is shown. The selection on the left side allows navigation based on the data (e.g. products, resources) as well as the configuration of the right side of the screen. In the screenshot the right side is divided horizontally into three “charts”. On top the product overview is displayed which shows the most relevant data for each product (e.g. inventory position). The table layout can be changed individually by hiding or displaying columns and changing their sequence. The second “chart” is the planning table which shows the capacity load per resource, and for each resource which products account for the capacity load. Here, interactive planning can take place directly in the table. On the lower right part of the screen, alerts relevant to the context can be displayed.

The detailed scheduling planning board in Figure 3.11 works according to a similar concept. While the left side shows the resource selection, the right side can be individually configured with different “charts”. The screenshot

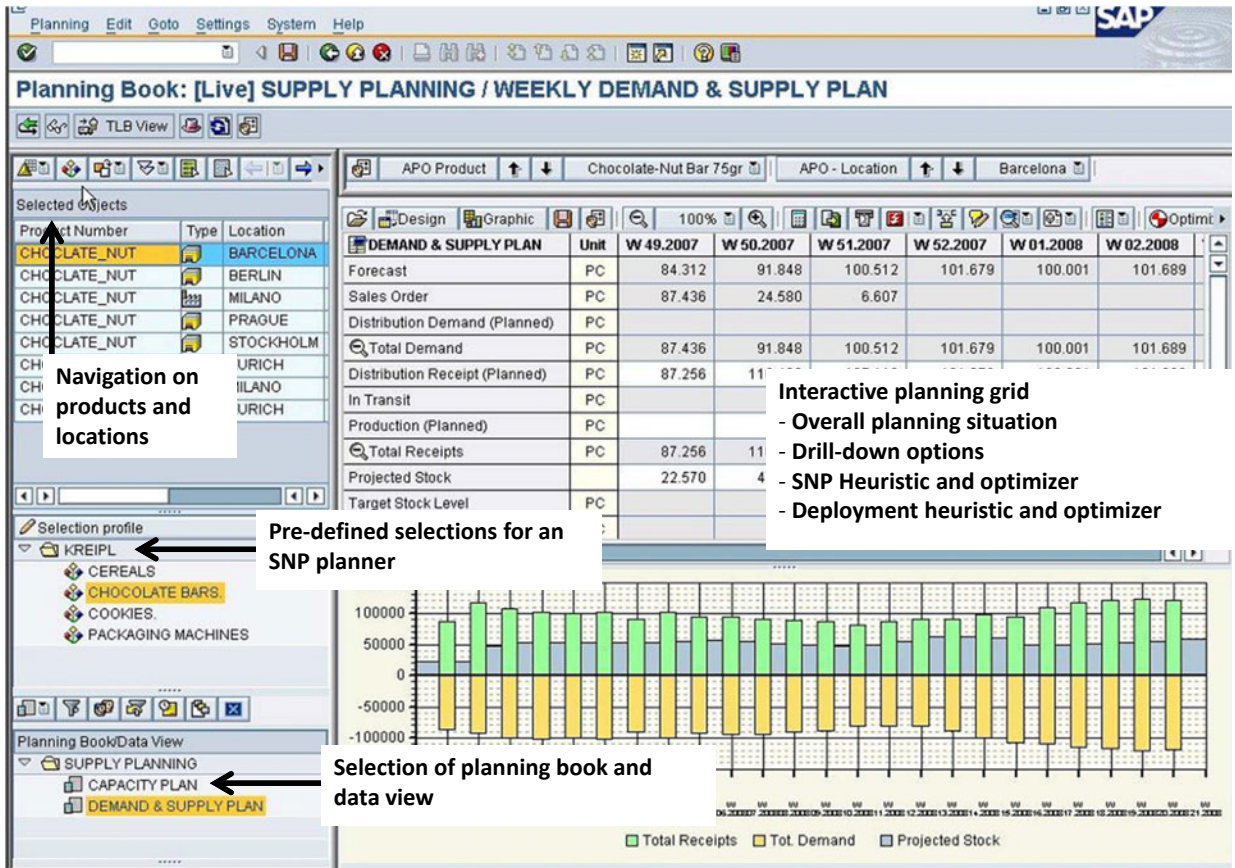


Figure 3.9

Planning book (here: SNP)

© Copyright 2011. SAP AG.
All rights reserved

shows a Gantt chart like display of resources and scheduled operations on the top of the screen. Manual planning activities including drag & drop are supported as well as heuristics (e.g. find next slot, insert operation). The second chart shows the product flow. Each line represents a product, and each box represents an order for this product. The links between the boxes represent the pegging links explained in Section 3.3.3. The example shows pegging links between planned production orders as well as to the demand (triangle). The third chart shows the resource utilization. This can be the utilization depending on the number of operations per resource (single or multi resources) or with reference to a tank capacity like shown in the example here. In this case the utilization represents the fill level of the tank at any given point in time. Finally, the last chart shows the inventory situation for the selected products.

Figure 3.12 shows the interactive vehicle scheduling screen. The left side of the screen either presents a list of orders (selected for planning), a list of resources (available for planning) or a list of shipments (being the result of planning). The right side of the screen is divided into several tabs. The tabular planning tab (depicted in the screenshot) shows three tables (which can have an individually configured layout). The first table shows all activities

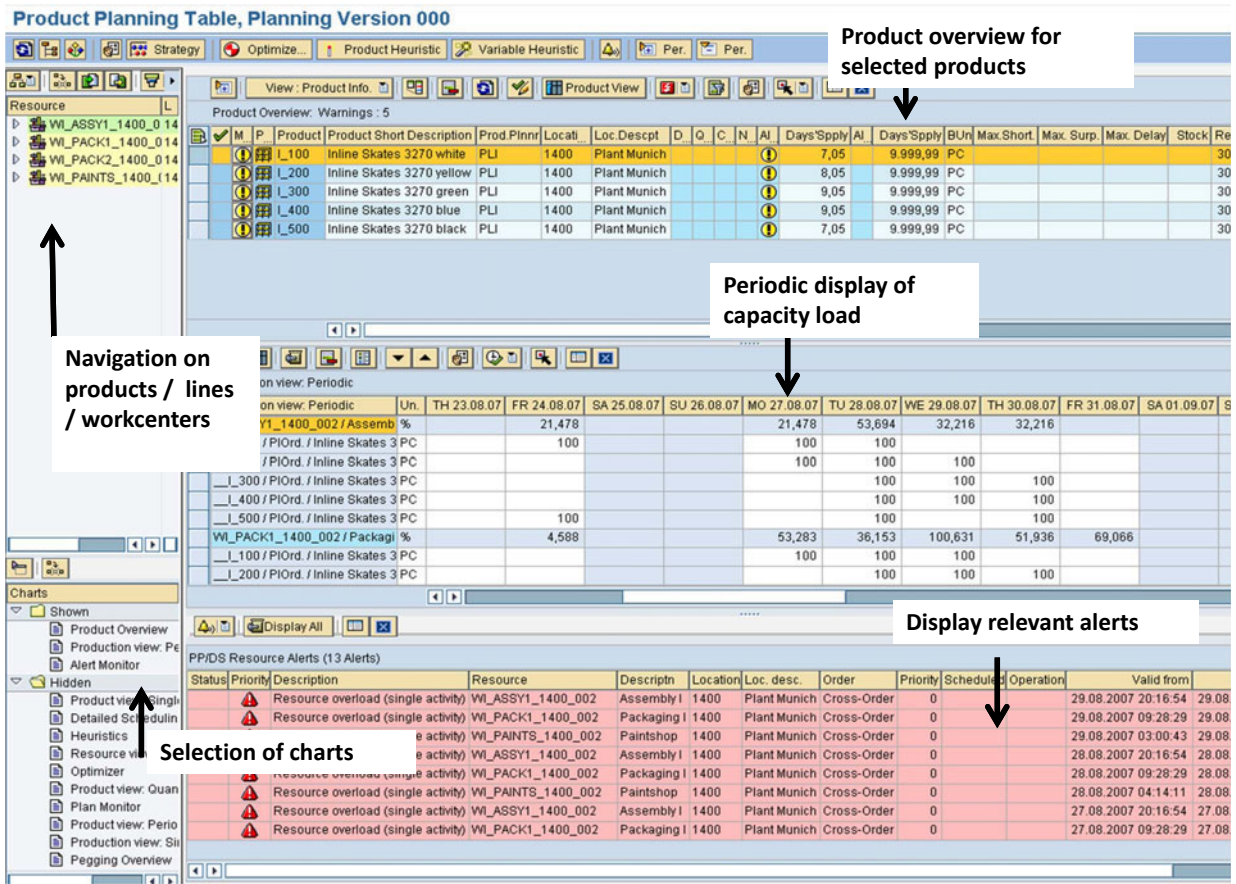


Figure 3.10
 Product planning table (PP/DS)
 © Copyright 2011. SAP AG.
 All rights reserved

of the selected resource or shipment. These can be handling activities (loading / unloading) or transport activities (routing of the shipment). The second table shows the freight units that are loaded on the resource / shipment and the third table shows the freight units that have not been assigned to any shipment. Drag & drop of freight units to resources and shipments is supported. Other tabs available are

- multilevel planning (different view of the shipment, tree-like),
- the overview planning board (a graphical scheduling screen similar to PP/DS),
- object planning board (similar to overview planning board, but displaying only the selected objects),
- geographical overview (map display of a shipment), as well as
- details (details of the selected object: resource, shipment, order).

On the lower left side of the screen alerts can be displayed.

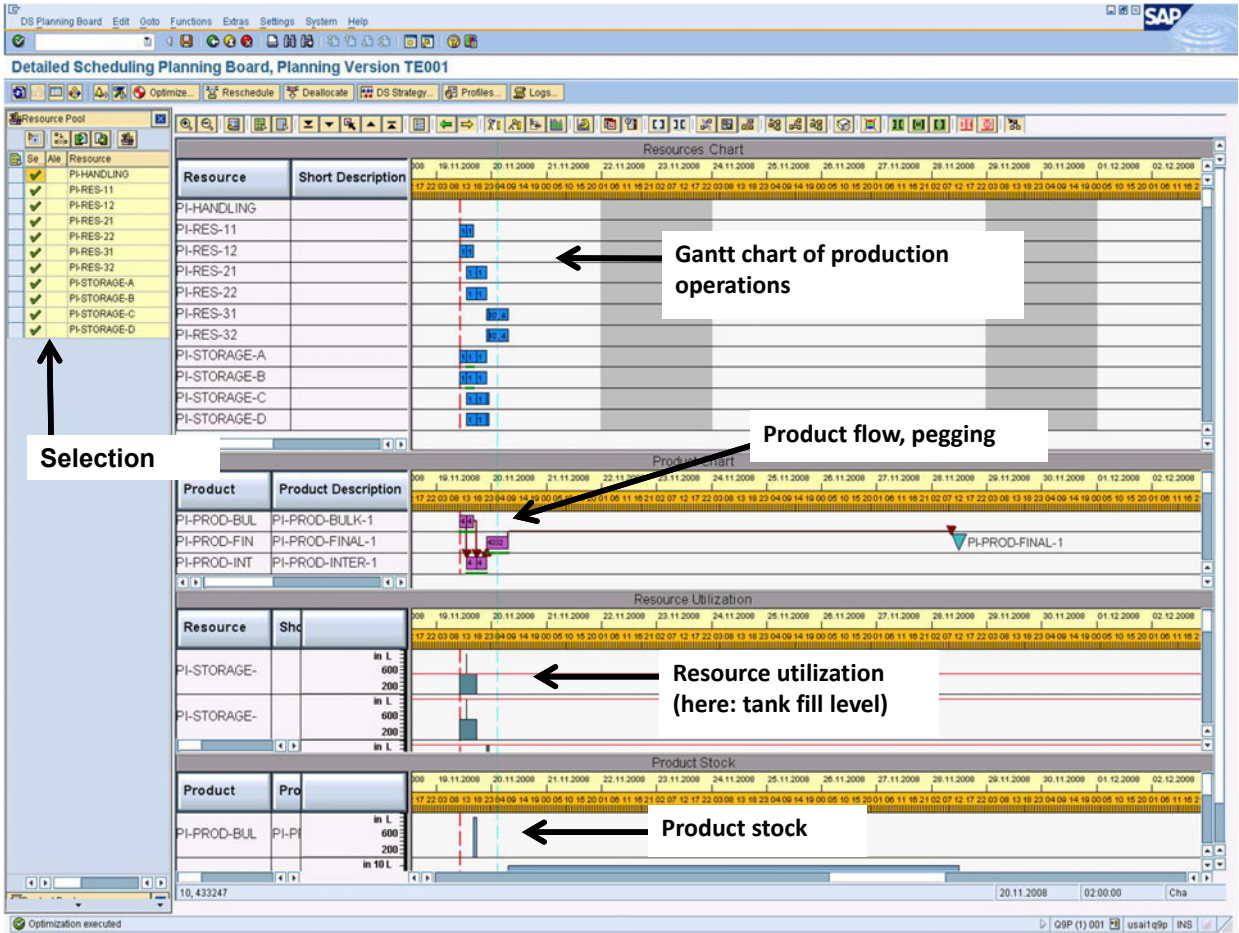


Figure 3.11
Detailed scheduling planning board (PP/DS)
© Copyright 2011. SAP AG.
All rights reserved

Finally, Figure 3.13 shows the alert monitor. The alert monitor is a powerful exception message system integrated in all SAP® APO planning modules. Alerts are displayed according to the alert profile which configures what types of alerts are displayed. It is possible to drill down from the alert monitor directly to the corresponding application to resolve the issue generating the alert.

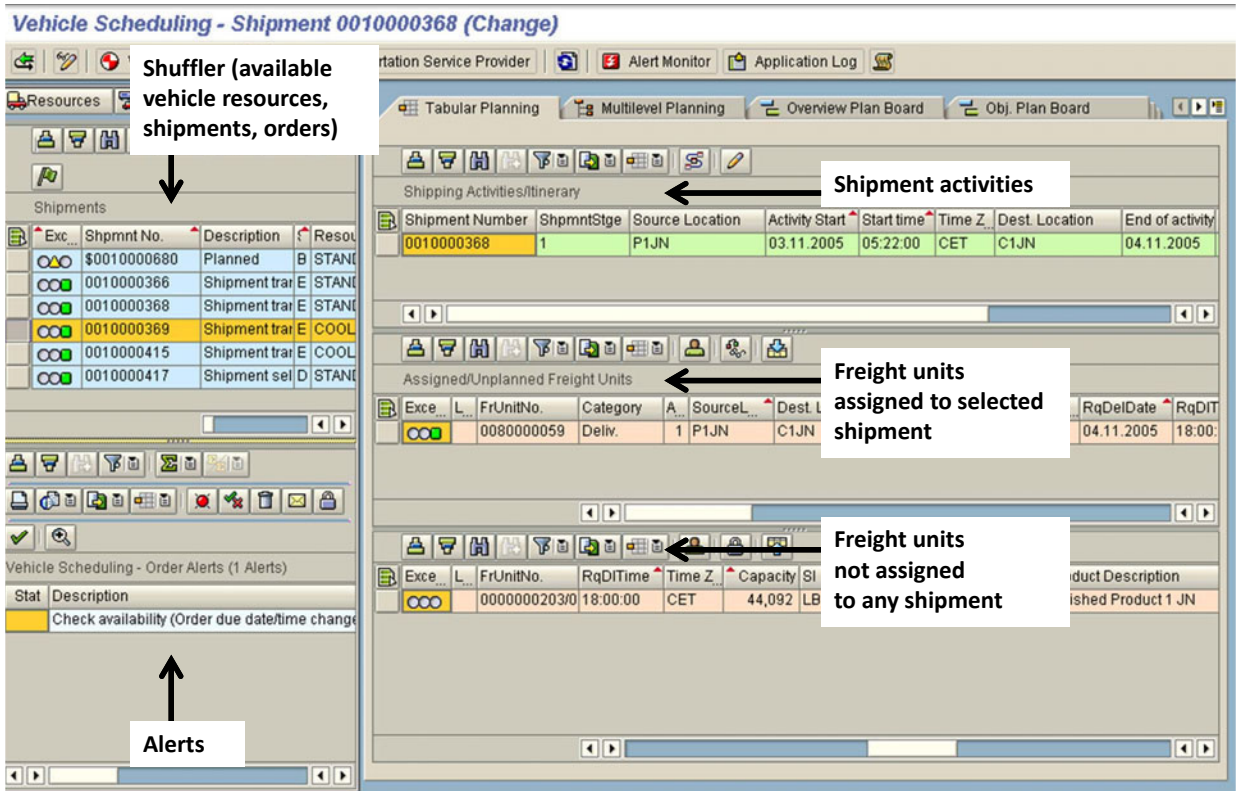


Figure 3.12
Interactive vehicle scheduling (TP/VS)
© Copyright 2011. SAP AG.
All rights reserved

Questions and Exercises

1. What is the purpose of having models and planning versions in SAP® APO? How is the relationship to a transactional system (ERP) being managed?
2. Name five master data objects that are relevant for supply chain planning. What is their role in SAP® APO?
3. What are production process models (PPMs)?
4. Explain the concept of pegging. Why is pegging important for planning?

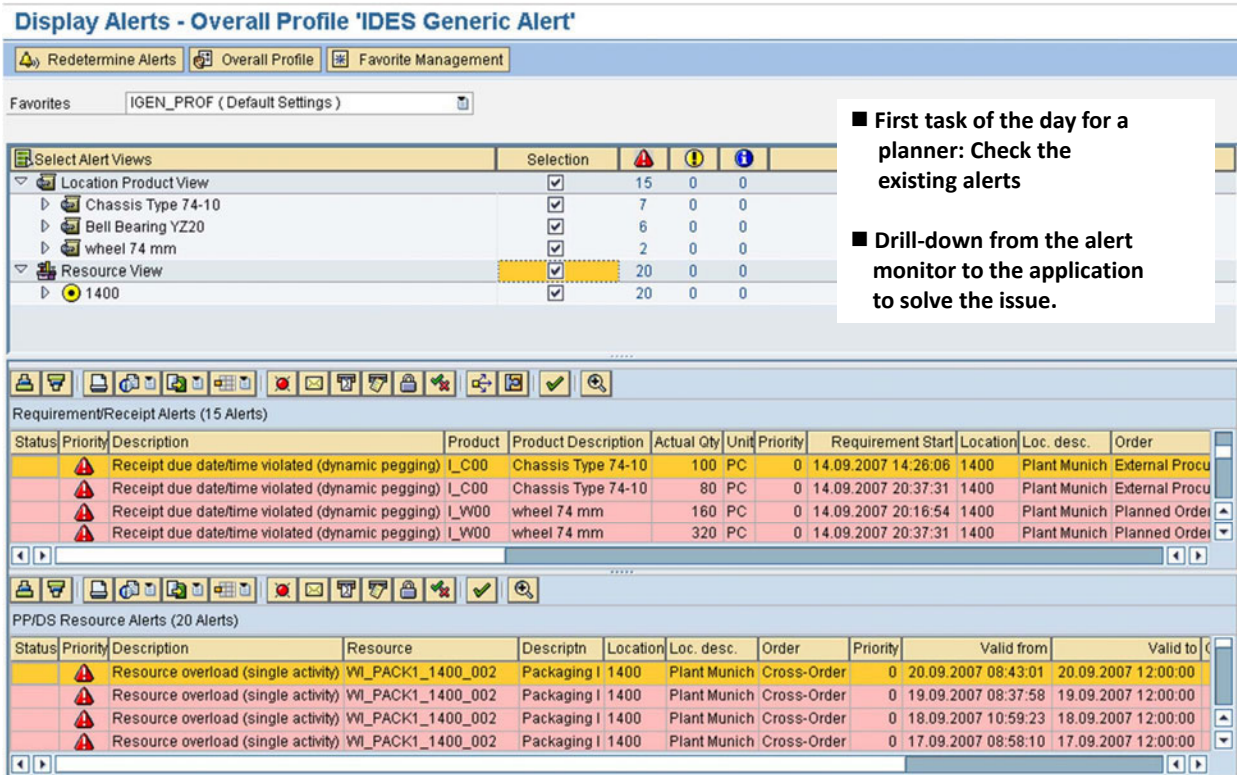


Figure 3.13

Alert monitor

© Copyright 2011. SAP AG.

All rights reserved

3.4 The SAP® APO Solution for the Frutado Case

The following planning tasks are covered by the case study.

- Demand Planning (short-term and medium-term) → DP
- Master Planning (cross-location) → SNP
- Short-term Production Planning and Scheduling → PP/DS
- Distribution Planning / Deployment → SNP
- Transportation Planning → TP/VS
- Available-to-Promise / Capable-to-Promise → Global ATP

Figure 3.14 gives an overview of the involved SAP® APO modules and planning tasks as well as their interactions / data flows. Note, that the flows to and from ERP are shown as they could occur in practice. However, to simplify the scenario and focus on the planning tasks, the Frutado case has been implemented as a SAP® APO standalone solution without ERP integration. Therefore, these flows are not presented as part of the learning units.

In the DP module medium-term and short-term demand planning is carried out. It generates medium-term and short-term forecasts of the sales per

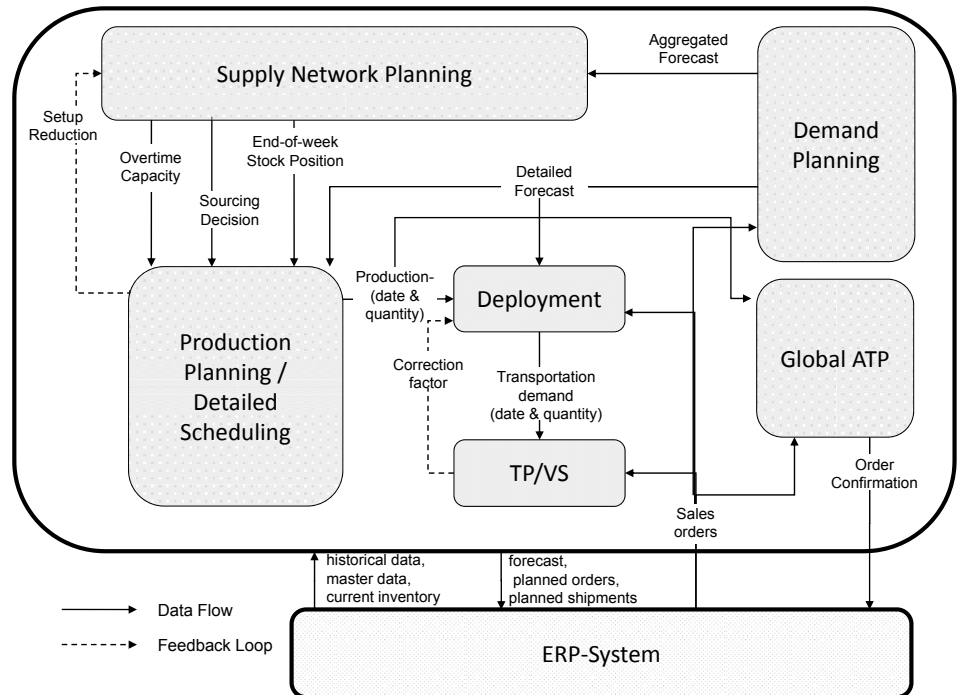


Figure 3.14
 Planning tasks and
 data flows in the
 Frutado case

product and per DC, based on past sales data. Medium-term demand planning covers a planning horizon of one year on a weekly basis which is the input for cross-location Master Planning by the SNP module. Short-term demand planning covers a planning horizon of four weeks on a daily basis, which is input for short-term production planning in PP/DS as well as distribution planning using the deployment functionality of the SNP module. Both forecasts rely on historical data covering the past two years.

The SNP module accomplishes the task of Master Planning considering production and transportation data capacities. It decides on the production quantities and their allocation to the plants and lines, on the overtime shifts to be used and on the transport quantities between the distribution centers. The objective is to minimize the costs of production, overtime, transportation, and inventory. The planning horizon is 26 weeks and divided into weekly buckets. The level of detail is single products, three plants, and three DCs. In total, the SNP module comprises 15 transport relations (9 from the three plants to the three DCs and six between DCs), 86 location-product combinations (29 plant-product combinations and three times 19 DC-product combinations), and 32 line-product combinations. As SNP considers only weekly production quantities, neglecting lot sizes and sequences, the setup times can only be estimated as an average, based on feed-back information about the actual setup times from the production planning module. Shelf-life restrictions are respected in this planning step to make sure, that no product is produced that does not have sufficient demand within its maturity time. Optimization using linear programming (LP) is used to solve the

planning problem in the base case. As a result of SNP, the allocation of the production lines, the planned overtime shifts, and the weekend target-stocks are transferred to the short-term production planning module.

In the PP/DS module a short-term production plan is created for each of the three plants of the Frutado company for a four week planning horizon with exact continuous timing. The goal is to minimize a weighted objective function including order delays, setup costs, and setup time as well as production (mode) costs. The production plan is created for each plant individually based on the short-term demand planning result and directives in the form of planned inventory positions from SNP.

Based on the short-term production plan and the short-term demand situation a deployment plan is generated using the deployment functionality of the SNP module. Note, that deployment is a separate process step, but the functionality that is used to execute this step is part of the SNP module. Deployment creates a plan of all distribution/transportation activities which need to be executed in a two week horizon on a daily basis considering the priorities of customer sales orders vs. distribution center forecasts. It may revise the medium-term distribution plan provided by SNP in case of shortage or surplus inventory. Such situations may have been caused by deviations of the current sales orders from the demand forecast or by the breakdown of a production line. The objective is to maximize the service level, i.e. to satisfy the current sales orders as well as possible. If, at one day, a complete order fulfillment is not possible, the important decision still remains which customers to serve and which not. This can be based on a priority classification of the customers.

The resulting transportation plan from Deployment specifies how much to transport from where to where and when, but does not yet consider the use of vehicles and their routing. Furthermore, the transport quantities of the Deployment plan, which have been determined for the different products independently, need to be combined by the Transport Load Builder (TLB), considering pallet and vehicle capacities. This palletization step (done by TLB) is part of SNP functionality similar to Deployment. Finally, the use of individual vehicle resources is scheduled so as to minimize the transportation cost and minimize due date violations (lateness) for the given transport requirements. This is the task of the TP/VS module. Here, a detailed transportation plan for each DC is derived. In particular, vehicle tours are scheduled for delivering the sales orders from the DC. The planning horizon for this task is one week with exact continuous timing.

Finally, the Global ATP module accomplishes the task of order promising. The ordinary way to do this is to confirm an arriving sales order, whenever the stock at the responsible DC is sufficient, and to postpone it for a fixed amount of time (standard lead time) otherwise. However, the Global ATP concept is more sophisticated in two aspects: First, it divides available stocks in different classes of “ATP quantities” according to customer priorities, in order to avoid upsetting important customers. Second, it checks all possibilities

of satisfying the order, including the stock at other DCs and forthcoming production.

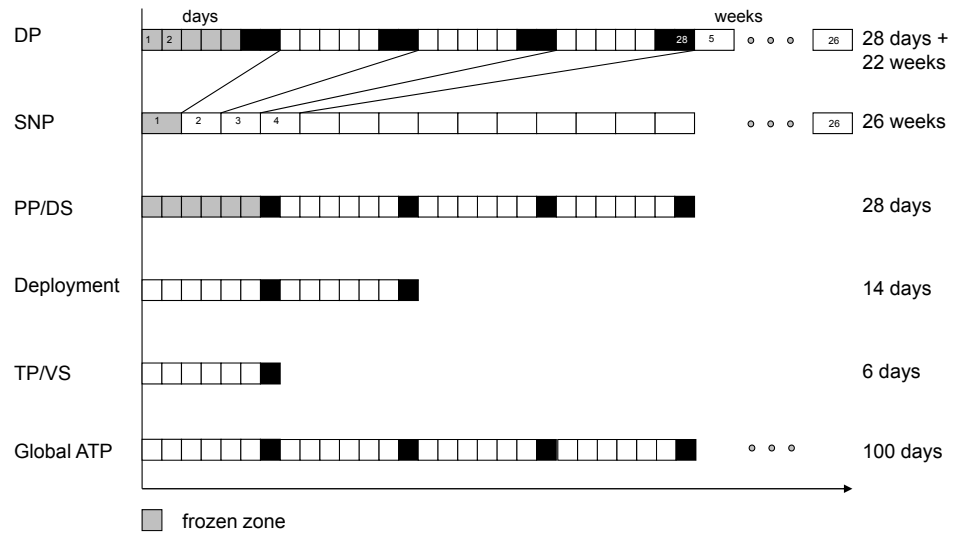


Figure 3.15 Rolling schedules

In Section 2.2 it has been explained that planning is not a one-time task, but happens regularly. The frequency of planning and its planning horizon depend on the planning tasks. Therefore, the different planning modules used by the Frutado company consider different planning intervals with different granularity. As plans are not implemented for the complete planning horizon, but are updated according to a re-planning interval of one week, the entire planning system shall work with a weekly rolling horizon. Thereby, the first week is a frozen zone for the DP, SNP and PP/DS modules. In this frozen zone the planning results of the prior planning run are fixed and have to be realized. At the beginning of week t , the DP and SNP modules generate new plans for the weeks $t + 1$ to $t + 27$, and the PP/DS module for the weeks $t + 1$ to $t + 4$. In contrast, Deployment, TP/VS, and Global ATP have no frozen zone and can run more frequently because they have to react on the current order arrivals. Figure 3.15 illustrates this procedure. In this rolling horizon concept, the ERP system plays an essential role: First, it contains the interface for the arrival of the sales orders. It administrates the sales orders and provides them to the SAP® APO modules. This is of importance not only for the short-term modules, but also for the DP, because the ERP system adds the recent orders to the sales history, which is the input for DP. Second, it administrates the current inventory positions, which are necessary data for all SAP® APO modules.

The complemental learning units which are referenced in subsequent chapters (see also Introduction) are based on a complete implementation of the Frutado case in SAP® APO modules with a common data base and a demand history of two years. For each module, typical user interactions have been recorded with the software “datango”. This software creates a detailed sequence of screen-shots corresponding to the user inputs and the

system responses. As mentioned before, the implementation does not contain the ERP system. Therefore, the rolling horizon frame cannot be captured. Instead, a single planning cycle through all modules can be executed, starting with initial stocks of zero everywhere. Due to a time displacement between the DP module, which creates forecasts for weeks 2 to 26 and the SNP and PP/DS modules, starting with week 1, the forecast demands can be produced in time. The remaining modules start in week 2 and can use the fixed results of week 1. [Figure 3.16](#) illustrates this procedure.

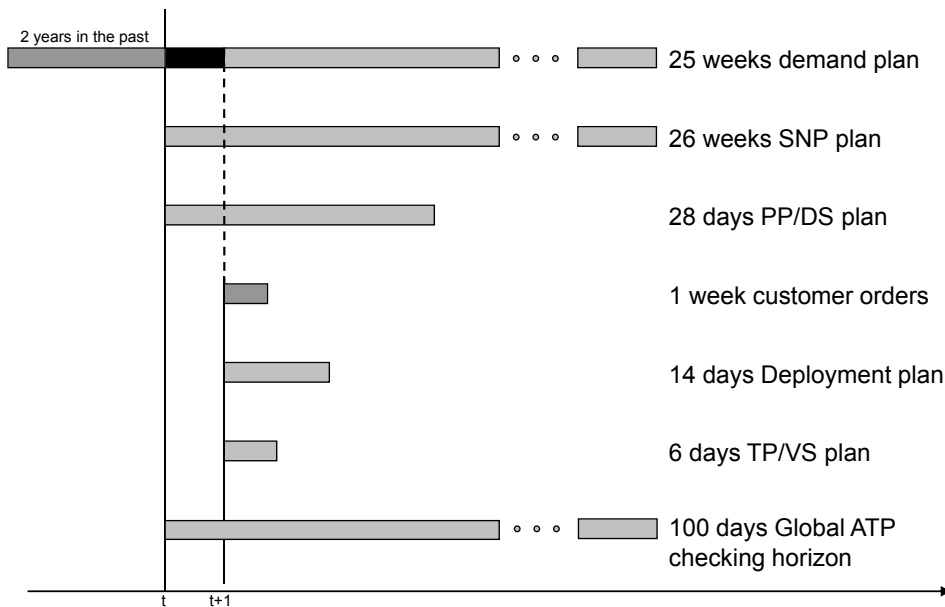


Figure 3.16
Frutado
implementation of
rolling schedules

Questions and Exercises

1. Which modules of SAP® APO are used in the Frutado case and what is the purpose of each module?
2. Explain the concept of rolling schedules and its application to the different planning tasks of the Frutado company.

Bibliography

SAP (2011) *Homepage*, URL <http://help.sap.com>, date: July 29, 2011

Part II

Planning the Frutado Case with
SAP[®] APO Modules

Demand Planning (DP)

Herbert Meyr¹

¹ University of Hohenheim, Institute for Supply Chain Management (580C),
70593 Stuttgart, Germany

Οἶδα οὐκ εἰδώς — “*I know that I know nothing*” is a popular saying that is attributed to the ancient Greek philosopher Socrates and his student Plato. This could also be seen as an early device for Demand Planning (DP).

The aim of planning is to prepare decisions for future execution. Thus planning is always forward-looking. Since we do not have complete information on what will happen in the future, forecasts have to substitute the missing part. Unfortunately, we can never be sure about forecasts. We can make mistakes and we will make mistakes. We have to be aware of the fact that there is uncertainty and that there will be a forecast error. And we have to take care of this error in our planning activities.

Nevertheless, most modules of advanced planning systems (APS) assume that there is a well-known, “deterministic” world. They leave it to the DP module to deal with uncertainty. As a consequence, the most prominent task of DP is to *predict* the future as accurately as possible. However, also the remaining uncertainty has to be taken care of. Usually, in APS this is done by calculating safety stocks as an additional buffer to hedge against the remaining uncertainty. Thus the second task, that is left to DP, is to recommend the other modules *safety stock levels* (see Wagner 2005, Chap. 7.1 or Kilger and Wagner 2008, Chap. 7.6.6). These levels should be just as high to ensure a desired level of service to the customers. And they are crucially dependent on the quality of the forecasts. The most important object that is usually forecast by DP is customer demand. Forecasting means predicting the demand, but not actively influencing it. Nevertheless, some APS also offer a means to actively “manage” demand, e.g. by planning the timing and extent of “promotions” (i.e. temporary price discounts) or advertisements. Therefore, the third task of DP is to check the effects of such activities by

means of *simulation and what-if-analyses* (see Wagner 2005, Chap. 7.1). Since the Frutado case mainly focuses on the forecasting task of DP, the latter two tasks will not further be considered in the remainder of this chapter.

Instead, after a general introduction to DP in Chap. 4.1 – addressing topics like measuring the forecast quality, determining the objects to forecast, quantifying a forecast, or sequencing the necessary steps to establish a forecast – basic forecasting models and forecasting methods, that can be found in the scientific literature, will be discussed. The aim is to demonstrate the general principles and to explain the models and methods that are used in the Frutado case. As a consequence, in Chap. 4.2 we concentrate on rather simple models, assuming a quite constant (“level”) demand, a trend, or some seasonality of demand. Chapter 4.3 then shows forecasting methods that solve these models using a so-called “times series analysis” approach.

Then we change from theory to practice. In Chap. 4.4, the planning tasks and data are introduced that are necessary to put DP at the Frutado company into practice. Chapter 4.5 shows how the Frutado planning tasks have actually been modeled in the case study. Chapter 4.6 briefly describes which other APO planning modules use the output of DP as an input for their planning. Finally, Chap. 4.7 gives an overview of the learning units that demonstrate Frutado’s DP implementation in the SAP APO software.

4.1 Introduction to Demand Planning

The basis for forecasts are always observations about the object to be predicted that have already been made in the past. Often, a so called “time series” of observations is known. Let, for example, t denote the current period (e.g. day or week) and x_t denote the realization of this object (e.g. amount of customer demand) during t . Then, at the end of period t , a time series (x_t) of observations $x_t, x_{t-1}, x_{t-2}, \dots$ would be known for the past periods $t, t-1, t-2, \dots$. Typically, this information about the past is used to predict the future. For instance, the so-called “time series analysis” tries to detect some regular pattern in the time series and to extrapolate this regular pattern into the future.

Obviously data analysis is very important to detect those patterns. But keep in mind that the data of the past can be incomplete (e.g. due to introduction of new products) and contain mistakes (e.g. due to manual order booking processes) or singular events that will not be repeated in the future and thus should not be extrapolated (e.g. the above mentioned promotions). Therefore, tools are necessary which do not only help to analyze the data of the past, but also to correct them or to supplement them for missing information. Usually a graphical representation of the time series is helpful to support these tasks. However, due to the huge amount of objects to be forecast also an automated support should be possible.

4.1.1 Measuring the Forecast Quality

The preceding chapters have shown that forecasts are inputs to many other planning modules. Thus, it is very important to know whether a forecast is of low or high quality. In other words: you need to be able to measure the quality of a forecast. The so-called “forecast error” builds the basis to do this.

The time series (x_t) has already been introduced above to express observations that have been made in the past. Let the superscript $\hat{}$ now denote forecasts that are made for the future. Then, for example, $\hat{x}_{t,t+1}$ represents a forecast that has been made at the end of t for the next period $t+1$, $\hat{x}_{t,t+2}$ represents a forecast for the period after, and so on. In general, $\hat{x}_{t,t+s}$ denotes a forecast that has been made at (the end of) period t for the future period $t+s$ ($s \geq 1$). Using this, Equation (4.1) defines the *forecast error* e_t^s that has come up when the realization x_t of period t has already been forecast s periods ahead by the estimate $\hat{x}_{t-s,t}$:

$$e_t^s := x_t - \hat{x}_{t-s,t}. \quad (4.1)$$

As everybody knows from listening to weather forecasts, the larger the lag s the higher the forecast error e_t^s usually is. In order to simplify the notation, we implicitly assume $s = 1$ and write e_t instead of e_t^1 when the length of the lag s is not crucial to understand a general principle. Analogously we use \hat{x}_t instead of the longer notation $\hat{x}_{t-1,t}$ if we want to express that a “one-period-ahead forecast” (i.e. $s = 1$) has been or shall be made for period t or that the lag s does not play an important role in this context.

The expected value of the forecast error of a “good” forecasting method should be 0 in order to be unbiased, i.e. to prevent systematic under- or overestimation. Furthermore, the standard deviation and variance of the forecast error should be as low as possible.

Many different measures are known from literature and supported by APS to judge the quality of forecasts and forecast methods. Somehow, they all use the forecast error as a basis. Some important ones are shown in the following. These measure the forecast quality at the end of period t ex post, by averaging over the last K periods. They are defined by the common Equation (4.2), but apply different functions $f(k)$ to judge the forecast quality:

$$\text{quality measure} := \frac{1}{K} \sum_{k=t-K+1}^t f(k). \quad (4.2)$$

In order to compute the quality measure “*mean error*” ME, $f(k)$ has to be set to $f(k) := e_k$. This way, the expected value of the forecast error is estimated – which should be close to 0 as postulated above.

On the contrary, the following quality measures represent the forecast variability. The “*mean absolute deviation*” MAD is given by $f(k) := |e_k|$. Here positive and negative deviations of the forecast cannot balance out anymore. Thus every mis-estimation is considered according to its absolute

deviation from the observation. The “*mean squared error*” MSE is defined by $f(k) := (e_k)^2$. As compared to MAD, here small errors (< 1) are mitigated, and large errors (> 1) are amplified. As we will see later on, minimizing the MSE is a popular objective when designing forecast models and methods. However, a disadvantage of the MSE is that it is hard to interpret in practice because its value does no longer express the same dimension as the forecast error. Like the MAD, the “*root mean squared error*” RMSE $:= \sqrt{MSE}$ can overcome this problem again.

Finally, “*mean absolute percentage error*” MAPE is defined by $f(k) := 100 \cdot \frac{|e_k|}{x_k}$. Since it measures the percentage deviation from an average observation, it eliminates any dimension. This is advantageous if the forecast quality has to be compared for very different types of forecasting objects like the value (e.g. expressed in \$) and the quantity (e.g. expressed in tons) of customer orders. However, MAPE is not defined if periods with observations $x_t = 0$ can occur – as might be the case for products showing sporadic demand. In this case, more refined quality measures are necessary. For this, and a deeper discussion of the advantages and disadvantages of forecast measures, the interested reader is referred to Kilger and Wagner (2008, Chap. 7.5). Forecasting measures that are used by SAP APO are explained by Hoppe (2007, Chap. 4.3).

The forecast quality has to be continuously controlled in order to ensure that the forecasting method used is still appropriate. A further dimensionless – and thus generally applicable – measure that can (besides others) be used for this purpose is the “*error tracking signal*” SIG. It is defined by $SIG := \frac{ME}{MAD}$. It can easily be shown that $|SIG| \leq 1$ holds. A popular way of controlling the forecast quality is to define a certain threshold $f < 1$ and to initiate an alert if $|SIG| > f$. APS usually support the automated control of this or similar tracking signals and the automated generation of alerts, which are sent to human planners to trigger manual intervention. For recommendations on how to set f and how to update SIG in a more clever way see Trigg (1964), Silver et al. (1998, Chaps. 4.6.3 and 4.6.4) and Tempelmeier (2008, p. 36).

4.1.2 The Objects to Forecast

Until now we have used the quite general term “forecasting objects” when referring to what is being predicted in DP and just indicated that this is usually the customer demand. This has to be specified in the following.

Please, recall that the forecasts made in a DP module usually serve as input for other planning tasks and planning modules like SNP or PP/DS. These take into account further constraints of the supply chain like limited supply of material or capacitated resources. Since the input to these planning modules are forecasts, of course, their output can only be forecasts, too. This is the reason why sometimes the former ones are called “unconstrained forecasts” and the latter ones are called “constrained forecasts”. However, in order not to confuse the reader and in order to emphasize the planning character of

the remaining planning tasks, we will abstain from this denomination in the following. Instead, we use the term “forecast” to address the output of the DP module and the term “plan” to address the output of the others. Nevertheless, the decisions to be made in the other planning modules drive the necessary input for planning and thus determine the objects to be forecast.

These objects usually have to be characterized by several attributes (or “dimensions”) like the type of product that has to be produced (e.g. Ice Tea 1, Ice Tea 2, Juice 1, Juice 2), the location where it has to be produced or delivered from (e.g. DC 1, DC 2), and the point in time the customer wants to get it (e.g. Day 1, Day 2). Of course, also further attributes are possible, for instance distinguishing different distribution channels or customers’ priorities. What is to be forecast are combinations of these attributes. Accordingly, the other planning modules request certain attribute combinations from the DP module. To give an example: the Basic Frutado Model, used for the Supply Network Planning of Chapter 5.4.2, needs the gross demand d_{jlt} of product j at location l in period t (which is measured in some quantity units QU) as input data of the Linear Programming model to be solved. Actually, this gross demand has to be estimated by the DP module. Thus, the index set jlt defines the corresponding attribute combination to be forecast in the DP module.

However, as shown above, planning is usually executed on several planning levels requesting different planning granularity. Since DP has to serve several planning levels simultaneously, forecasts also have to be made within hierarchies. Thus, even for a single attribute combination different levels of aggregation can be possible and necessary. For example, the above mentioned items Ice Tea 1, Ice Tea 2, Juice 1, Juice 2 might be aggregated to two product groups containing either ice teas or juices. Also the distribution centers DC 1 and DC 2 might be aggregated according to the regions they are located in, and days might be aggregated to weeks. Thus the attribute combination jlt has to be further refined according to the granularity of planning. While a PP/DS module might request a granularity on item/DC/day-level, an SNP module might request a granularity on group/region/week-level. Thus the *Characteristic Value Combinations (CVCs)* of SAP APO, that have been introduced in Chapter 3.3.2, actually define both the attribute combination and its corresponding granularity.

Theoretically, forecasts could be made on each planning and granularity level (i.e., for each CVC), independently of each other. Thus the forecasts of the different levels would mostly be inconsistent. Of course, this would not be desirable for planning, yet doable. However, in this case their time series also would have to be stored separately for each CVC. This would – due to the huge number of CVCs that usually have to be forecast – consume too much storage space. Thus forecasts and their time series are usually only stored on a single level. Mechanisms to consistently and reproducibly aggregate or disaggregate them have to exist in case the forecasts are needed on other levels, too. For example, SAP APO preferably stores forecasts only

at the highest level of detail. They can be stored in other granularity, too. However, it is recommended to do this only if the speed of planning would be too slow. Even then, forecasts are just stored consistently, i.e. the SAP APO system in the background enforces consistency by using appropriate aggregation or disaggregation rules.

But if the forecasts are always stored in the highest level of detail, anyway, why not also always create them on this level? The reasons lie in the typical characteristics of forecasts that can be observed and proven: firstly, the more detailed a forecast has to be made the worse it usually is. In other words, the more aggregate a forecast can be made, the better it usually is because in this case randomness can balance out (as is known from risk pooling and portfolio effects in finance). Secondly, as we have already mentioned above, the longer the forecast horizon is the worse forecasts usually get. This is the reason why planning is often very detailed on the short-term, but gets more aggregate for longer planning horizons. Therefore, if, for example, a mid-term master planning module like SNP can anyway only handle aggregated data for reasons of complexity, only aggregate forecasts should be generated.

Thus, depending on the level at which the original forecasts have been made “top-down”, “bottom-up” or “middle-out” aggregation and disaggregation rules have to be defined in order to consistently generate forecasts for other planning levels in lower or higher granularity. This functionality is called “multilevel planning”¹ in SAP APO (see e.g. Hoppe 2007, Chap. 2.3.1). Further information on forecasting hierarchies can, for example, be found in Kilger and Wagner (2008, Chap. 7.2).

4.1.3 Basic Forecasting Approaches

A further open question is how forecasts should actually be made. Chapters 4.2 and 4.3 will take care of this in some detail. Here we only want to give a first and rough overview of basic approaches to generate forecasts.

According to Wagner (2005, Chap. 7.1 and Fig. 7.1) basically statistical, judgmental and collaborative/consensus-based forecasting techniques can be distinguished.

Statistical (or “quantitative”) *techniques* have mainly been developed to support business and econometric forecasting processes. They use mathematical models and methods to automatically generate forecasts on the basis of some observations of the past. They can further be subdivided into approaches based on time-series-analysis and causal approaches. As already mentioned models of *time-series-analysis* try to detect a regular pattern in the time series of the past, assume that this pattern will be further repeated and thus extrapolate the pattern into the future. On the other hand, *causal* (or “explanatory”) *models* assume that there also might be other factors (so-called “leading indicators”, “explanatory”, “independent” or “predictor”

¹ Please note, that the term “multilevel planning” is also used in other contexts of SAP APO like ATP or SNP – and has a different meaning there.

variables) than just time influencing the object to be predicted (the “dependent” variable). Thus they try to identify and estimate reliable dependencies between these leading indicators and the dependent variable. An example for a leading indicator might be the amount of money that has been spent on advertisements, which is influencing the dependent variable “customer demand”. If one is able to measure (or at least estimate) the money spent, the knowledge about this dependency might help to generate good forecasts. Various forecasting methods have been proposed to estimate the parameters expressing the respective dependencies of the different models.

In contrast to statistical forecasting, *judgmental forecasting* (or “qualitative” forecasting in general) additionally tries to incorporate information that cannot be found in the time series of the past, e.g. knowledge about promotions or advertisements that are only planned for the future, but did not happen in the past. Similarly, as already mentioned above, the time series might contain misleading information on singular events that will not repeat themselves in the future. Then the time series should be corrected accordingly. In both cases manual intervention is necessary because only human planners know about this additional, exceptional information that is not contained in the time series. However, since human planners might be opportunistic and biased, it is not trivial to integrate this additional information with statistical forecasts. Judgmental forecasting aims at offering structured ways to combine both sources of information by simultaneously avoiding bias and over-valuation of some information.

Collaborative (also called *consensus-based*) forecasting actually addresses the same problem. Information, which is obtained from several sources, has to be integrated into a combined forecast in the best possible manner. The term just emphasizes the fact that in supply chains many different partners – which might belong to the same company (e.g. different regional departments of a sales hierarchy) or various companies (e.g. a sales department of a supplier and a purchasing department of a buyer) and have different “local” information – should contribute to a joint forecast in a collaborative process (they can all agree on, so that consensus is achieved). The question here again is how the forecasts of the different sources should be weighted within the combined forecast. One can easily see how the forecasting hierarchies and aggregation/disaggregation processes of the last section are related: the underlying problem is actually the same. However, depending on the application context, the level of automation of the forecast generation process and the individual “negotiation power” of the human planners involved can be different.

Since the Frutado case mainly applies statistical forecasting techniques of time-series-analysis, Chapters 4.2 and 4.3 will only concentrate on describing the corresponding models and methods of time-series-analysis in greater detail. For further information on statistical and judgmental forecasting approaches the reader is referred to Hanke and Wichern (2008), Kilger and Wagner (2008) and Makridakis et al. (1998).

Usually APS support all of these techniques. For example, the SAP APO Demand Planning module (SAP APO DP) offers several time series based methods pooled in a category called “Univariate Forecasting Methods” and causal methods based on multiple linear regression pooled in a category called “Causal Analysis”. According to principles of structured judgment, the “Composite Forecasting” functionality additionally allows to combine several (manually or with these alternative methods independently created) forecasts into a single one using a pre-defined weighting scheme. Further human judgment can, for instance, be incorporated using the “Promotion Planning” functionality. This could be extended by the “Collaborative Promotion Planning” functionality, which enables inter-organisational information exchange, e.g. between consumer goods manufacturers and retailers. The consistency in (intra-organisational) forecasting hierarchies is enforced by the “Aggregation and Disaggregation” functionality of the “Interactive Demand Planning Desktop” offering different rules for top-down, bottom-up or middle-out forecasting.

4.1.4 The Demand Planning Process

Finally an overview shall be given on the subsequent phases and iterative steps that are necessary to generate and apply forecasts. As we have seen above, forecasts have to be made on several planning levels to feed different planning modules like SNP or PP/DS. These modules do not only determine the granularity of the forecasts, but also the frequencies in which the forecasts have to be made. Like these modules, DP is also subject to a rolling horizon scheme. In order to generalize, in the following we only distinguish between “initial phases”, that are necessary to establish a forecasting process, and “routine phases”, that are repeatedly executed as part of a rolling horizon planning. However, we do not go further into detail nor, for example, differentiate between mid- and short-term forecasting, etc.

Table 4.1 summarizes the initial and routine phases that are part of the demand planning process. The table is adapted from Kilger and Wagner (2008, Fig. 7.7). However, it also considers the individual steps that are necessary to select an appropriate forecasting model and forecasting method.

Initial Preparation of Demand Planning Structures

When establishing a new forecasting process, first the corresponding demand planning structures have to be prepared (1). This means, for example, that the units of measurement and the granularity (like the time structure and – if necessary – a product group) of the CVC to be forecast have to be defined. Further administrative processes of APS might be necessary, e.g. creating data views and assigning responsibilities to human process owners.

- | |
|--|
| <ol style="list-style-type: none"> 1. Initial preparation of demand planning structures 2. Initial data analysis and preparation 3. (Initial) model selection: <ol style="list-style-type: none"> (a) Set a demand model for testing (b) Set a forecasting model and method for testing (c) Determine initial values of the forecasting method (d) Set parameters of the forecasting method (e) Test the forecasting method by ex-post-simulation (f) Analyze forecasting quality 4. Routine phases: <ol style="list-style-type: none"> (a) Data preparation (b) Statistical forecasting (c) Judgmental forecasting (d) Consensus forecasting (e) Release of the forecast (f) → Observation of demand realization (g) Analyze forecasting quality (if satisfying go to 4a) (h) Alert generation (if necessary go to 3 again) |
|--|

Table 4.1
Initial and routine
phases of the demand
planning process

Initial Data Analysis and Preparation

In order to enable forecasting, some sort of demand history is required. If a time series already exists, it has to be imported, analyzed for a regular pattern and possibly adapted so that singular events, that will not be repeated in the future, or other misleading exceptions are eliminated (2). Unfortunately, such time series will not always be available or might be too short to contain meaningful information. This happens, for example, when a new product is introduced. In this case sometimes at least a “like-profile” can be created. This means that the demand history of a similar product, which has already been on the market before (like a predecessor product), might be defined as a surrogate.

Model Selection

Finally, an appropriate forecasting model and a corresponding forecasting method have to be selected. Phase 3 of [Table 4.1](#) shows the individual steps that are necessary to test a single forecasting method. If we take time-series-analysis as an example, on the basis of the data analysis of phase 2 a fitting demand model has to be chosen (see also Chap. 4.2). Next, a corresponding forecasting (optimization) model and method have to be selected. As Chapters 4.2 and 4.3 will show, usually several alternative forecasting models and methods exist for a single demand model. Such a

method often needs one or several initial values and forecasting parameters to be defined before it can be run the first time. When these have been determined, the forecasting method can be tested. For this, usually an “ex-post-simulation” is executed. This means that forecasts are generated for some part of the (already known) demand history. This way, the procedure merely simulates what would have happened if the forecasting method had been applied in the past. However, since demand realizations have already been observed for this period of time, the quality of the forecasting method can easily be analyzed (see Chap. 4.1.1) and compared to other potential forecasting approaches. Examples for the data analysis and data adaption of phase 2 and for the individual steps of phase (3) are given by Meyr (2008).

Finding the “best” forecasting method is a difficult problem. There are countless ways to generate forecasts by varying the demand and forecasting models, forecasting methods, procedures to create initial values and forecasting parameters of phase (3). Additionally several measures to judge the quality of a forecast exist (see Chap. 4.1.1) and a good forecasting method should be advantageous with respect to all (or at least most) of them. However, APS usually support this selection process. They offer so-called “Pick-the-best” functionalities, which automatically choose between these alternatives and determine the method(s) and parameters to be used (see Kilger and Wagner 2008, Chap. 7.6.5, also concerning potential dangers). SAP APO, for example, offers two approaches called “Automatic Model Selection Procedure 1 & 2” for the automated selection between some time series models (see Hoppe 2007, p. 135ff.). The first one applies autocorrelation tests to select a model. The second one also applies tests to select a model (and method), but additionally determines appropriate parameters of the method by varying them on a discrete grid. The corresponding quality measure can be defined by the user.

Routine Phases

The preceding phases have to be executed only once or quite seldom. In the following, the routine phases are described, which have to be passed during each rolling horizon period (4).

For the same reasons as in phase 2, the data of the preceding period – incorporating the last observation of demand – also have to be analyzed and potentially corrected (4a). Next, a statistical forecast (4b) can be generated automatically, using the method and parameter settings selected in phase 3.

In a further step, judgmental forecasting might be necessary to add additional information that has not yet been considered (4c). If we take time-series-analysis as an example, this could be information that was not contained in the corrected time series of the past, e.g. about exceptional marketing activities like promotions or advertisements that are planned for the future. Note, that the planning of these activities can again be supported by (statistical) forecasting methods and APS. For example, the effect of a certain price discount can again be estimated by analyzing and

extrapolating the effects of former price discounts of the past. Nevertheless, after deciding about the timing of the promotion and the extent of the discount, its (estimated) additional effect on demand has to be incorporated into the statistically generated forecast of phase 4b (see Chap. 4.7.3).

If several parties are involved in the forecasting process, for example, because the (decentral) knowledge of regional sales managers (who are closer to the customers) or even the assessments of some important customers' procurement managers can contribute to a company's overall forecast of world-wide sales, a consensus or even collaborative step should follow (4d).

After a final agreement has been found, the forecast can be released for usage in other planning modules like SNP or PP/DS (4e). Until the next planning will be due, the realization of demand has been observed for (some periods of) the frozen horizon (4f). This new information can be used to control whether the forecast quality is still satisfying (4g). The quality measures of Chapter 4.1.1 help to judge this. As we have seen, by using tracking signals like SIG this control process can even be automatized. In this case alerts are automatically sent to the responsible human planner if some pre-defined thresholds are exceeded (4h). Then it is up to the human planner to decide whether the current forecasting model and method are still appropriate or new ones have to be selected by going back to phase 3 again.

Questions and Exercises

1. Why do we need forecasts?
2. What are the reasons for holding safety stocks? What drivers influence the safety stock levels?
3. How can forecast quality be measured?
4. How can we assure to use the "correct" forecast model?
5. In general it is a bad idea to create forecasts on the highest level of detail. Explain this statement and give at least two reasons.
6. Which kinds of forecasting techniques do you know?
7. What are possible phases of the demand planning process?

4.2 Demand Planning Models in the Literature

In the following we want to demonstrate the basic idea how DP usually models demand. Because the Frutado case mainly uses methods of time-series-analysis, we will only concentrate on this branch of statistical forecasting techniques and there on some selected models. For general introductions to DP models the reader is referred to Hanke and Wichern (2008), Kilger and Wagner (2008), Makridakis et al. (1998), Silver et al. (1998, Chap. 4) and Tempelmeier (2008, Chap. C). We start with the simplest assumption about

demand that can be made, namely that it is on a constant level, and extend our overview to more complicated demand models in Chapter 4.2.2.

4.2.1 Level Demand

The simplest assumption one can make about demand is that it actually is on a constant (“steady”) level a . However, in each period k there might be some random influence u_k (also called “random noise”, “white noise” or “error term”), which prevents that real demand a from being seen. Thus the observation x_k , that can be seen or measured, is assumed to consist of a and u_k according to the “*level demand model*” shown in Equation (4.3):

$$(4.3) \quad x_k = a + u_k \quad \forall k.$$

It is further assumed that randomness diminishes over a long period of time, i.e. that the expected value of u_k is zero. The general idea of generating a forecast at the end of the current period t is then to find an estimator \hat{a}_t for a such that \hat{x}_{t+1} can be computed by $\hat{x}_{t+1} := \hat{a}_t$ (or in general $\hat{x}_{t,t+s} := \hat{a}_t \forall s \geq 1$).

To find such an estimator is an *optimization problem*. The unknown level a is the variable and \hat{a}_t is the optimum solution of this problem. \hat{a}_t minimizes the distances to all the observations x_t and thus minimizes the error terms u_t . An exemplary *optimization model* to represent this problem could be:

$$(4.4) \quad \text{minimize } MVA(a) = \sum_{k=t-K+1}^t (x_k - a)^2 = \sum_{k=t-K+1}^t (u_k)^2.$$

The model $MVA(a)$ implicitly assumes (i) that the observations x_t and error terms u_t are uncorrelated, (ii) that – similar to MSE – squaring the error terms is an advantageous distance measure, (iii) that the last K observations are sufficient to represent the constant level and (iv) that all these observations are equally important and thus should get the same weight.

As we have seen in Chapter 4.1.1, other distance measures would also be conceivable. Furthermore, it seems reasonable that newer observations can better represent future demand than older observations. Thus, also (iii) and (iv) appear crucial because due to them, recent observations receive less importance the higher K gets. On the contrary, choosing a very small K would not sufficiently take into account information of early observations.

Therefore, alternative optimization models could be more appropriate. For example, $WMVA(a)$ according to

$$(4.5) \quad \text{minimize } WMVA(a) = \sum_{k=t-K+1}^t w_k \cdot (x_k - a)^2$$

would be a more general and thus improved optimization model, where the weights w_k cannot only take on the values 1, but also more “realistic”

increasing values fulfilling $w_{t-K+1} < w_{t-K+2} < \dots < w_t$. For example, the weights

$$w_k := \alpha(1 - \alpha)^{t-k} \quad \forall k = t - K + 1, \dots, t \quad (4.6)$$

could be chosen for a parameter α fulfilling $0 < \alpha < 1$. We will come back to this later.

4.2.2 Trend and Seasonality

Of course, the assumption that demand is constant is very restrictive and will only be realistic in few practical situations. Quite often more refined – and thus also more complicated – demand models will be necessary.

A straightforward extension of the level model is to assume that demand will increase or decrease over time. This is called a “*trend*”. If a trend with a (constant) slope b adds to the level a in a linear relation like

$$x_k = a + b \cdot k + u_k \quad \forall k, \quad (4.7)$$

this will be denoted as an “additive trend (demand) model” in the following.

$WT(a, b)$ could be an exemplary optimization model to represent this demand model:

$$\text{minimize } WT(a, b) = \sum_{k=t-K+1}^t w_k \cdot (x_k - a - b \cdot k)^2. \quad (4.8)$$

Since both level \hat{a}_t and trend \hat{b}_t have to be estimated now, this model depends on the two variables a and b .

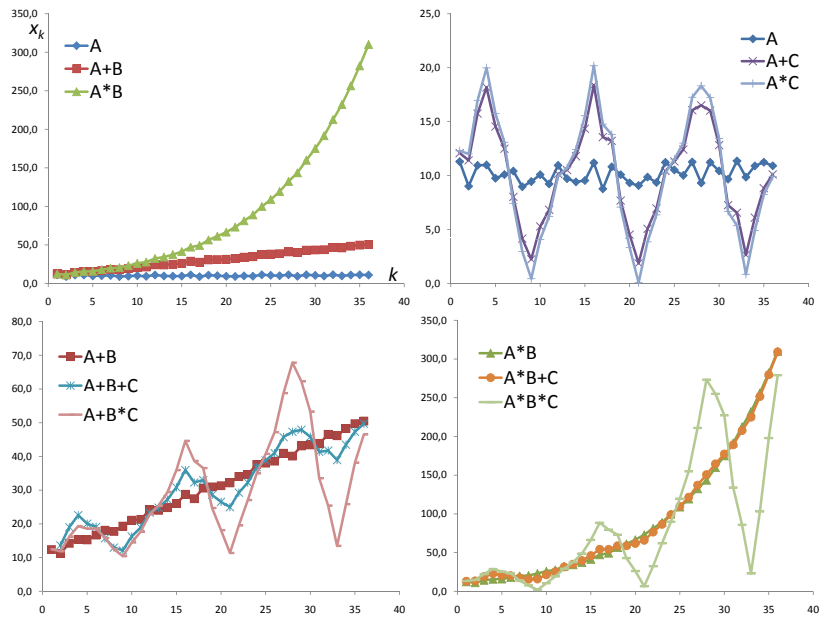
However, also other assumptions about a trend could be appropriate for certain practical applications, for example, that the trend changes the level demand in a multiplicative manner like $x_k = a \cdot b^k + u_k$.

Furthermore, in practice often seasonal effects do occur. For example, sales of bulbs are higher during winter times than during summer times because the nights are longer in the winter. Then sales are lower than average during summer, whereas a peak of sales can be observed during winter. This pattern repeats itself every year. In other words, bulbs show a “*seasonal cycle*” of one year. A “*seasonal demand model*” assumes that a seasonal coefficient c_k of a certain period k does represent the period’s seasonal influence. In our bulb example, the seasonal coefficient of the month December would clearly be higher than the seasonal coefficient of the month June.

Again, these seasonal coefficients can change the level and/or trend demand in an additive or multiplicative manner. An example of an additive trend and a multiplicative seasonal influence is given by the demand model (4.9):

$$x_k = (a + b \cdot k) \cdot c_k + u_k \quad \forall k. \quad (4.9)$$

Figure 4.1
 Additive (+) and multiplicative (*) combinations of the level (A), trend (B), and/or seasonal (C) components of a demand model. For example, $A * B + C$ describes the multiplicative trend, additive seasonal demand model $x_k = a \cdot b^k + c_k + u_k \forall k$. $A + B * C$ illustrates Equation (4.9).



Graphical illustrations of all nine possible additive and multiplicative combinations of level, trend and seasonal influences are given by Pegels (1969) and Makridakis et al. (1998, Fig. 4-1), which is replicated in Figure 4.1.

Questions and Exercises

1. Which types of demand models do you know?
2. Propose potential optimization models for seasonal demand models.
3. Formulate appropriate demand models for the plots $A + B + C$ and $A * B * C$ of Figure 4.1.

4.3 Demand Planning Methods for Time Series Analysis

In the following some selected forecasting methods for the level (Chap. 4.3.1) and additive trend (Chap. 4.3.2) demand models are presented. Additionally, two methods for additive trend and multiplicative seasonal demand will be shown in Chapter 4.3.3. This chapter concludes with some remarks on forecasting methods for other demand models of time-series-analysis.

4.3.1 Level Demand

Remember that the level demand models apply the relationship

$$(4.10) \quad \hat{x}_{t,t+s} := \hat{a}_t \quad \forall s \geq 1$$

to calculate, at the end of the current period t , a forecast $\hat{x}_{t,t+s}$ for the future period $t + s$, $s \geq 1$. Thereby, \hat{a}_t is an estimate of the demand level a that has been made at the end of period t .

Unweighted Moving Averages

One way to get such an estimate is to solve (4.4) to optimality by differentiating it with respect to a , setting the derivative to 0 and resolving this equation with respect to a (see e.g. Tempelmeier 2008, C.2.1.1). The resulting optimal solution gives the desired forecast as shown in eq. (4.11):

$$\hat{a}_t := \frac{1}{K} \sum_{k=t-K+1}^t x_k. \quad (4.11)$$

As can be seen, this forecast is just the average over the last K observations. Since this has to be done in every period and the length of the observed interval does not change, “moving averages” over the last K periods are built. Of course, this way of forecasting is quite intuitive so that it would not really be necessary to formulate and solve the optimization model (4.4) to get the idea. Nevertheless, it nicely illustrates the general principle of how forecasting methods can be derived.

Note, that K is a parameter of the forecasting method, which has to be chosen in advance (see phase 3d of Table 4.1). It can have a crucial influence on the quality of the forecasts and thus has to be set carefully. The lower it is the higher the influence of new observations are. The extreme case of $K = 1$ is called the “naive forecast” for obvious reasons. Tempelmeier (2008, p. 42) recommends $3 \leq K \leq 12$ as being typical in order to prevent a too high influence of old observations (if K is too large) and nervousness (if K is too small). Since all K observations contribute equally to the forecast, this procedure is called the “unweighted moving averages” method.

Weighted Moving Averages

As mentioned in Chapter 4.2.1, giving more recent observations a higher weight sounds reasonable and might improve the forecast. The “weighted moving averages” method

$$\hat{a}_t := \frac{1}{\sum_{k=t-K+1}^t w_k} \sum_{k=t-K+1}^t w_k \cdot x_k \quad (4.12)$$

results when solving (4.5) to optimality. Unfortunately, applying it seems difficult in the sense that obviously the increasing weights w_k have to be chosen somehow as further parameters of the method.

Single Exponential Smoothing

But actually it is not really. Setting the weights according to $w_k := \alpha(1 - \alpha)^{t-k}$ ($0 < \alpha < 1$) is an elegant way to reduce the efforts to a single parameter α again. Choosing these weights and letting $K \rightarrow \infty$, procedure (4.12) can even be simplified to the following (because $\hat{x}_{t-1,t} = \hat{x}_t = \hat{a}_{t-1}$) recursive relationship (see e.g. Tempelmeier 2008, Chap. C.2.1.2):

$$(4.13) \quad \hat{a}_t := \alpha x_t + (1 - \alpha)\hat{x}_t =$$

$$(4.14) \quad = \alpha x_t + (1 - \alpha)[\alpha x_{t-1} + (1 - \alpha)\hat{x}_{t-1}] =$$

$$= \dots =$$

$$(4.15) \quad = \hat{x}_t + \alpha(x_t - \hat{x}_t).$$

This method is very easy to use in practice because – as can be seen in (4.13) – only two values have to be added and stored: the last observation x_t and the last forecast \hat{x}_t . These are weighted by the so-called “smoothing parameter” α , which does regulate (“smooth”) the influence of the last observation x_t on the new forecast and thus the adaptability of the forecast to changes in demand. Because of the definition of w_k and the recursiveness of the old forecast \hat{x}_t (see (4.14)), information about the complete time series and increasing weights are implicitly contained. Due to the exponential character of the weights this method is called “*single (or simple) exponential smoothing*”. As (4.15) shows, the method can also be interpreted as correcting the last forecast \hat{x}_t for an α -fraction of the last forecast error e_t .

Again the parameter α has to be set in advance. It can, for example, be gained by ex-post simulation on part of the time series (if one is available), varying $0 < \alpha < 1$ on a discrete grid (e.g. $\alpha = 0.1, 0.2, \dots, 0.9$). According to Tempelmeier (2008, p. 48) values $0.1 \leq \alpha \leq 0.3$ have proven advantageous in practice.

Instead of K , an initial value for the recursion contained in \hat{x}_t is necessary if the forecasting process is started for the first time. If a time series is available, it can be used to compute this initial value (e.g. as an average of some already known observations). Otherwise, a naive forecast on the basis of the first observation could, for example, be used to start the exponential smoothing process with the second observation.

All in all, for an initial model selection and method initialization as postulated in phase 3 of Table 4.1 the time series should typically be divided into at least two segments for computing parameters (like α) and initial values (like \hat{x}_t) of forecasting methods and for executing the ex-post-simulation (e.g. to compare different parameter settings or forecasting methods) without a bias.

Adaptive-response-rate Single Exponential Smoothing

Usually α is only computed once or quite seldom, e.g. if the error tracking signal of Chapter 4.1.1 indicates in a certain period t that the forecasting

method used probably does not fit the time series any longer. However, this error tracking signal SIG_t can also be applied to adapt the smoothing constant every period t (what is emphasized by using α_t instead of α in the following).

The basic idea of putting this into practice is shown in (4.16) and (4.17):

$$\alpha_t := |SIG_{t-1}| \quad (4.16)$$

$$\hat{a}_t := \alpha_t x_t + (1 - \alpha_t) \hat{x}_t. \quad (4.17)$$

Makridakis et al. (1998, Chap. 4/3/2) denote this procedure as “*adaptive-response-rate single exponential smoothing*”. They recommend to also update SIG_t by exponentially smoothing its components e_t and $|e_t|$. The one period time lag between SIG_{t-1} and α_t is intended to prevent from overreactions to changes.

4.3.2 Additive Trend

As shown in Chapter 4.2.2, to model trend demand, estimates \hat{a}_t and \hat{b}_t for both the level and the trend component are necessary for the end of the current period t . The forecast $\hat{x}_{t,t+s}$ for a future period $t + s$ has then to be computed by

$$\hat{x}_{t,t+s} := \hat{a}_t + \hat{b}_t \cdot s \quad \forall s \geq 1 \quad (4.18)$$

for all of the following additive trend models.

Simple Linear Regression

We start again by solving the trend optimization model $WT(a, b)$ shown in (4.8) to optimality for the special case of equal weights $w_k = 1 \forall k$. Doing this leads to a “*simple linear regression*” model where the linear relationship is described by the variables a and b for the basic level and the trend slope.

To simplify the notation in the following let the abbreviation \sum_k denote $\sum_{k=t-K+1}^t$. Then, the optimal level \hat{a}_t and optimal trend \hat{b}_t of $WT(a, b)$ can be computed by (see e.g. Tempelmeier 2008, Chap. C.2.2.1):

$$\hat{a}_t := \frac{\sum_k k^2 \cdot \sum_k x_k - \sum_k k \cdot \sum_k (k \cdot x_k)}{K \sum_k k^2 - (\sum_k k)^2} \quad (4.19)$$

$$\hat{b}_t := \frac{K \sum_k (k \cdot x_k) - \sum_k k \cdot \sum_k x_k}{K \sum_k k^2 - (\sum_k k)^2}. \quad (4.20)$$

Similarly to the moving averages of the level model, linear regression is quite demanding. In every period t , all K observations of the past have to be stored and are needed for computing a new forecast. And again, the disadvantages of equal weights prevail.

Simple linear regression – and especially multiple linear regression, where not only a single but several independent variables are assumed to exist –

are actually traditional methods of causal forecasting. However, as we have seen here, also time itself can be interpreted as an explanatory factor. Thus, simple linear regression can be used for time-series-analysis, too.

Double Exponential Smoothing

Analogously, the trend optimization model $WT(a, b)$ of (4.8) can optimally be solved for weights $w_k := \alpha(1 - \alpha)^{t-k}$. Again, this ends in an exponential smoothing method. However, single exponential smoothing would lead to a systematic error in the case of a trend $b > 0$. Instead, it has to be smoothed twice. Thus, the resulting method is usually denoted as “(Brown’s) double exponential smoothing” (see e.g. Makridakis et al. 1998, p. 161 and Brown 1959 for an early description of the method).

In the following, a simpler notation than Brown’s, which has been suggested by Gardener Jr. (1984), is shown (see also Tempelmeier 2008, p. 67):

$$(4.21) \quad \hat{a}_t := \hat{a}_{t-1} + \hat{b}_{t-1} + \alpha(2 - \alpha)(x_t - \hat{x}_t)$$

$$(4.22) \quad \hat{b}_t := \hat{b}_{t-1} + \alpha^2(x_t - \hat{x}_t)$$

Note, that these are again recursive expressions showing similar advantages like single exponential smoothing has in the level demand case. However, the expressions have become more complicated than (4.15) because now also the one-period-ahead forecast $\hat{x}_t = (\hat{x}_{t-1,t} =) \hat{a}_{t-1} + \hat{b}_{t-1} \cdot 1$ is actually composed of the two components \hat{a}_{t-1} and \hat{b}_{t-1} .

Now initial values for both \hat{a}_{j-1} and \hat{b}_{j-1} are necessary if the method is to be newly introduced and applied for the first time at the end of a certain period j . Then at least two observations x_j and x_{j-1} are required to get an initial estimate, e.g. according to $\hat{a}_{j-1} := x_{j-1}$ and $\hat{b}_{j-1} := x_j - x_{j-1}$. If an even longer time series of $K > 2$ observations $x_j, x_{j-1}, \dots, x_{j-K+1}$ exists, also simple linear regression could be used to calculate \hat{a}_{j-1} and \hat{b}_{j-1} .

Method of Holt

Holt suggests to smooth \hat{a} and \hat{b} separately by two different parameters α and β (see Holt 1957 and Holt et al. 1960). This allows the following updating procedure:

$$(4.23) \quad \hat{a}_t := \alpha x_t + (1 - \alpha)(\hat{a}_{t-1} + \hat{b}_{t-1})$$

$$(4.24) \quad \hat{b}_t := \beta(\hat{a}_t - \hat{a}_{t-1}) + (1 - \beta)\hat{b}_{t-1}$$

The assignment (4.23) smooths the last observation x_t of the level with its one-period-ahead forecast $\hat{a}_{t-1} + \hat{b}_{t-1}$ by a weight α . Furthermore, (4.24) weights the most recent observation $\hat{a}_t - \hat{a}_{t-1}$ of the trend and its last forecast \hat{b}_{t-1} by the second parameter β .

Silver et al. (1998, p. 94) note, that the underlying optimization model is more general than (4.8) because it allows independent random changes of both

a and b in each period in addition to the random noise u_t . Furthermore it can be shown that the above double exponential smoothing method is a special case of Holt's method if Holt's α and β are replaced by $\alpha := 1 - (1 - \bar{\alpha})^2$ and $\beta := \frac{\bar{\alpha}^2}{1 - (1 - \bar{\alpha})^2}$ for a joint smoothing parameter $\bar{\alpha}$.

Yet, the higher degree of freedom of Holt's method is paid for by a higher effort to find suitable values for two parameters instead of a single one. If grid search is used again, now combinations of both α and β have to be tested by ex-post simulation.

4.3.3 Multiplicative Seasonal Demand

Finally we want to consider forecasting methods for the additive trend and multiplicative seasonal demand model (4.9). Thus, for representing seasonality we also need estimates \hat{c}_k for the seasonal coefficients c_k .

For this kind of demand model, the basic forecasting procedure for a single seasonal cycle² of length L reads as:

$$\hat{x}_{t,t+s} := (\hat{a}_t + \hat{b}_t \cdot s) \cdot \hat{c}_{t+s-L} \quad (s = 1, \dots, L). \quad (4.25)$$

Method of Winters

Probably the most well-known forecasting method for this demand model is an extension of Holt's method to seasonal demand by Winters (1960). In "*Winters' method*", additionally a third smoothing parameter γ is introduced to also independently update the seasonal coefficients according to:

$$\hat{a}_t := \alpha \frac{x_t}{\hat{c}_{t-L}} + (1 - \alpha)(\hat{a}_{t-1} + \hat{b}_{t-1}) \quad (4.26)$$

$$\hat{b}_t := \beta(\hat{a}_t - \hat{a}_{t-1}) + (1 - \beta)\hat{b}_{t-1} \quad (4.27)$$

$$\hat{c}_t := \gamma \frac{x_t}{\hat{a}_t} + (1 - \gamma)\hat{c}_{t-L}. \quad (4.28)$$

The relation to Holt's method is obvious. In (4.26) x_t is replaced by its so-called "de-seasonalized" representation $\frac{x_t}{\hat{c}_{t-L}}$. The computation of the trend component in (4.27) remains unchanged. Additionally, the last forecast \hat{c}_{t-L} of the seasonal coefficient of period $t - L$ is exponentially smoothed with the new observation of the seasonal influence $\frac{x_t}{\hat{a}_t}$ in (4.28).

Now also initial values for the seasonal coefficients \hat{c}_{j-L+k} ($k = 0, \dots, L - 1$) are required if Winters' method is to be applied for the very first time at the end of a certain period j . They can, for example, be gained by a so-called "*ratio-to-moving averages decomposition (RTMAD)*" (see e.g. Makridakis et al. 1998, Chap. 3/4/2). The basic idea is to compute averages of observations for complete seasonal cycles of length L in order to be sure that seasonal influences

² Seasonality is assumed to repeat every L periods in the future. Thus, a more general expression, also valid for $s > L$, would be $\hat{x}_{t,t+s} := (\hat{a}_t + \hat{b}_t \cdot s) \cdot \hat{c}_i$ for $i = t + s - L - L \cdot \lfloor \frac{s-1}{L} \rfloor$ with $\lfloor y \rfloor$ being the biggest integer number less than or equal to y ; see e.g. Tempelmeier (2008, p. 78).

have been excluded (de-seasonalized again). By dividing the original demand observations (which still contain seasonal influences) by the de-seasonalized averages, seasonal coefficients can be observed (or “decomposed”). However, these still contain random noise. Thus they have to be averaged again and finally normalized to L in order to serve as initial estimations \hat{c}_{j-L+k} of the seasonal coefficients.

Looking at this, it is obvious that quite a long historical time series should be available to initialize seasonal coefficients. Kilger and Wagner (2008, p. 147) and Makridakis et al. (1998, p. 168) recommend at least two complete seasons for this process.

Dividing the original observations of the historic time series by the initial seasonal coefficients, that result from the RTMAD procedure above, gives a de-seasonalized time series of the past. Simple linear regression applied to this de-seasonalized time series can again help to obtain initial values \hat{a}_{j-1} and \hat{b}_{j-1} of the level and the trend components. A detailed example of these initialization procedures, but also of a complete execution of Winters’ method, is given by Meyr (2008).

Linear Seasonal Regression

“*Linear seasonal regression*” – another forecasting method for additive trend and multiplicative seasonal demand models – uses exactly the same idea to forecast the future: the last K (at least $2L$) periods of the past serve as a basis to get the estimators \hat{c}_{t-L+k} ($k = 1, \dots, L$), \hat{a}_t and \hat{b}_t by applying RTMAD (or another method to estimate seasonal coefficients) first and linear regression to the de-seasonalized time series afterward. Then (4.25) can again be used to create the forecast.

4.3.4 Methods for other Demand Models

As already mentioned, only the most important demand models and forecasting methods have been presented that are necessary to understand the Frutado case. Of course, further exponential smoothing methods exist for other types of demand models. They are similar to Winters’ method presented above. Pegels (1969) summarizes them for all nine additive and/or multiplicative combinations of level, trend and seasonal demand in a unified framework. A nice and helpful overview of this work is also given by Makridakis et al. (1998, Chap. 4/3/5).

For more detailed information on exponential smoothing and time-series-analysis the reader is referred to Hyndman et al. (2008) and Box et al. (2008). Overviews of (business) forecasting approaches in general are, for example, given by Hanke and Wichern (2008) and Makridakis et al. (1998).

Questions and Exercises

1. Which forecast methods for level demand models do you know?

2. What are the advantages and disadvantages of single exponential smoothing compared to others?
3. Assume that the following values x_t have been observed for periods t of the past:

$t =$	1	2	3	4	5	6	7
$x_t =$	6	10	8	11	14	12	17

- (a) Calculate the one-period-ahead forecast $\hat{x}_{7,8}$ with single exponential smoothing for a smoothing parameter $\alpha = 0.3$. Use averaging to get the initial value.
 - (b) Plot the observations and your calculated forecast values. Do you see any problem? If yes, can you solve it?
 - (c) Recalculate the forecast for $t = 8$ with (Brown's) Double Exponential Smoothing. Use simple linear regression to get the initial estimate and discuss the new result.
 - (d) Recalculate the forecast for $t = 8$ with Holt's method ($\beta = 0.5$) and discuss the new result.
4. Do you know any further forecast models? If yes, what are they used for?

4.4 Planning Tasks and Data for the Frutado Company

We will first summarize the data that is available and relevant as input for the demand planning within the Frutado case. As we already know, this is just the historic time series of the past. The other planning modules, especially SNP and PP/DS, define which output is necessary. This allows us to finally describe what the planning tasks of the Frutado DP module actually are.

4.4.1 Available Data

As explained in Chap. 1, the Frutado case handles 60 regular customers who are served from three different distribution centers (DCs). Each DC takes exclusive care of 20 of these customers. This means that all demand of a single customer is only requested from his assigned DC, and that the customer is only delivered from this DC.³

The customer requests of all customers are known for two years in the past, i.e. for each of these customers all of his orders are known with the customer's corresponding address, desired delivery day, and desired delivery quantity (measured in hectoliters [HL]) of a certain final item. Since the

³ Nevertheless, there might be cross-shippings between the DCs.

fixed assignment of customers to DCs is also known, the total demand of a DC can be computed on a daily basis. Thus, a time series of customer demand does exist *for each DC, final item and day for two years in the past*. As we will see later on, this information is sufficient for demand planning. The concrete customer orders will mainly be necessary for deployment and transportation purposes. Note, that we do rely on customer orders rather than on actual deliveries that have been executed in the past because – due to backlogs – the latter ones might distort the real demand.

Additionally customer requests – in the same granularity – are also known for the next week of the future, i.e. all customers order with at least one week lead time in advance. Thus, no further orders will arrive with a desired delivery date within the next week, and forecasting is not necessary for the next week. Nevertheless one could interpret the already known customer orders of the next week as forecasts with best possible forecast quality (i.e. zero forecast error). Note, that only planning modules with a planning horizon of one week or less (e.g. TP/VS) can exclusively rely on already known customer orders. Planning modules with longer planning horizons (like SNP, PP/DS and Deployment) need forecasts, too.

Every final item can be delivered from every DC. Altogether 19 final items can be produced and demanded. As can be seen in [Table 1.1](#) (p. 15), six of them are ice teas. In the SAP APO implementation of the Frutado case, these are denoted as *FRU_ICETEA_xx* where the *xx* can stand for 04, 06, 08, 09, 17, or 19. The remaining thirteen are fruit juices. In the implementation, they will be denominated as *FRU_SAFT_xx* where the *xx* can take on the remaining integers ≤ 18 . Analyses of their time series indicate that items 01, 04, 06, and 10 show a seasonal demand with a peak during summer times (see [Figure 1.3](#) on p. 16). The items 07, 08, 09, 11, and 14 also show a seasonal demand, but with a peak during winter times. There does not seem to be seasonality for the remaining items. A trend cannot be recognized for any item. Thus, all items show almost a level demand with random fluctuations, some of them a further seasonality.

4.4.2 Planning Tasks and Level of Detail

In the context of the Frutado case, DP merely means forecasting. The Frutado case assumes that the necessary safety stock levels have in the past already been calculated by the sales department and then remain fixed for the whole planning horizon. Thus they do not need to be computed in a DP module any more. What-if-analyses and long-term forecasting would also go beyond the scope of the case. Therefore only mid- and short-term forecasting for operational planning remain as the main planning tasks of the DP module.

In order to be more concrete, we have to take a look at what the other planning modules expect as output of the demand planning process.

The SNP module needs mid-term demand forecasts for each of the 19 final items, for each of the three DCs and for each week of the next half year.

Thus forecasts have to be available at the granularity of DCs, final items, and weeks for a forecasting horizon of at least 26 weeks. However, it is up to the DP module to find out *how to get* these in the best possible quality and with a reasonable effort. For example, it has to be found out whether it is more convenient to forecast on the more aggregate level of product groups first and disaggregate the results to final items afterwards instead of forecasting each final item separately. And it has to be found out which demand model and forecasting method fits best.

Additionally, PP/DS needs short-term forecasts at the granularity of DCs, final items, and days for a forecasting horizon of 4 weeks. This is also the most detailed level of granularity that is necessary. Deployment uses the same granularity with a planning horizon of only two weeks. As we have seen, the first week can be excluded from forecasting because customer orders are already known. Thus, the Deployment module can mainly rely on customer orders. The forecasts of the second week are just used to get some idea about what will be going on after the first week. Therefore, it is not crucial that these forecasts only exist on DC-granularity instead of customer location granularity.

Judgmental and consensus-based forecasting are not necessary for the basic Frutado case. Nevertheless, at least to some extent, judgment will be demonstrated in an extension of the basic case.

4.5 Modeling the Frutado Planning Tasks

Now it will be demonstrated how the Frutado case is modeled with SAP APO DP.

4.5.1 Introduction to SAP® APO DP

Before this is possible the basic introduction to SAP APO DP of Chapter 3.1 has to be extended with some further information.

SAP Demand Planning Process

Similar to the one shown in [Table 4.1](#), SAP also recommends a typical demand planning process for the usage of SAP APO DP (see SAP 2011a). A – for Frutado’s needs adapted – version of it is shown in [Figure 4.2](#). The relation to the general process of [Table 4.1](#) is indicated by the numbers in brackets. Since the general process has already been described in Chapter 4.1.4, only some specifics of SAP’s implementation will be discussed in the following.

As part of the initial preparation of demand planning structures (1), the planning object structure containing the characteristic values and characteristic value combinations (CVCs), a *calendar* to plan the demand and a *storage buckets profile* to save the *key figures* for planning have to be defined. Key figures, for example, comprise historical data (orders that have been fulfilled), customer orders (that are still open) and demand forecasts. By

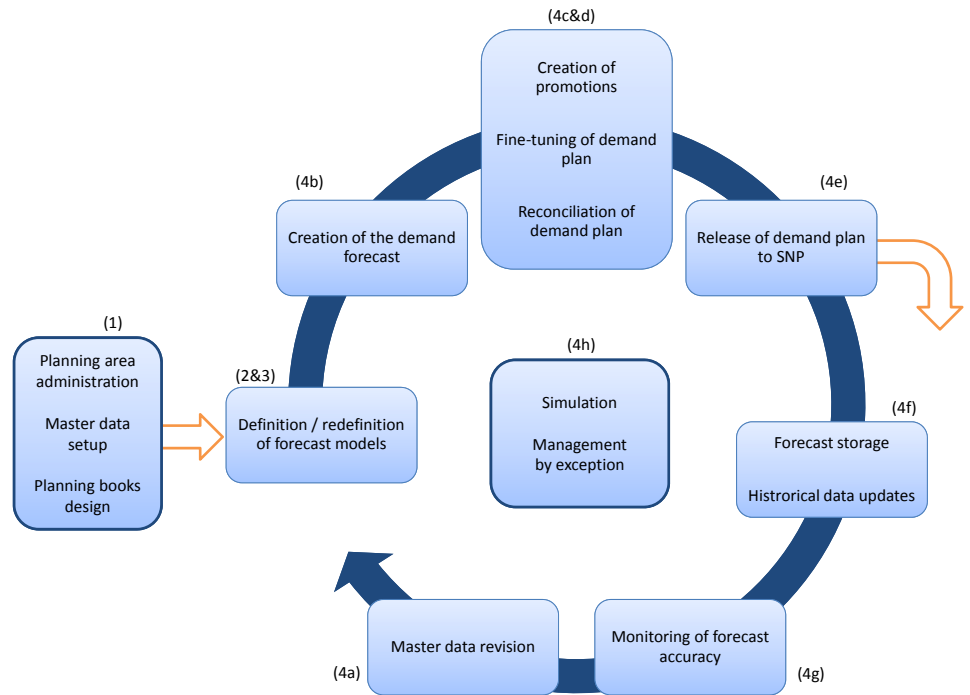


Figure 4.2
SAP APO demand
planning cycle
(adapted from SAP
2011a). The numbers
in brackets refer to
Table 4.1.

doing this, the time granularity is set, e.g. it is determined whether and when a daily, weekly or monthly precision is sufficient. As mentioned in Chapter 3.3.3 the *planning area* serves as the central data structure of the DP module enabling to define and manage the key figures. However, actual values of the key figures have to be held in *planning versions*, thus, for example, allowing to create several different forecasts from the same time series representing several different planning scenarios. Additionally, a *planning buckets profile* has to be defined. While the storage buckets profile allows to define the storage granularities, the planning buckets profile defines the granularities of the planning processes. Thus, the planning buckets define the time grid(s) which is (are) actually used for planning the demand. Multiple planning buckets profiles can be defined in a single planning book (see Chap. 3.3.3 and Fig. 3.6), e.g. in order to allow a user to test and compare different time grids and planning horizons.

Defining characteristic values and CVCs also means that all *master data*, which are necessary for (demand) planning, need to be *set up* (see Chap. 3.3.2). This includes products, locations and CVCs as feasible combinations of them, but also hierarchies, i.e. the levels of aggregation on which demand should be planned. Here aggregate groups can be generated from the CVCs of most detailed granularity. Furthermore, rules for aggregation and disaggregation between these levels can be defined. The alternatives, that are relevant for the Frutado case, will be explained in more detail below.

Since several users may be involved in DP, a planning area may contain several *planning books*. Thus each user is, for example, able to adapt the

graphical user interface (the *interactive planning screen*) to its own needs. But, even more important, different *data views* can be defined for the various users or – as we have seen above – within a single planning book to model different time grids etc. This way, only the data that are necessary (and permitted) for a certain user or a certain planning task are made available. There are two modes of generating forecasts, either interactively by the user or in the background. Both can be configured in the planning book. If calculations (and especially complex calculations) have to be done repeatedly, they can conveniently be automated by macros. The planning book also allows to define so-called *advanced macros* for this purpose.

Forecast models and methods can be tested and selected using the interactive planning screen (2&3). In order to choose and parameterize a forecast method for a key figure and forecasting horizon, a *master forecast profile* has to be defined. Both the so-called “*forecasting model*” (univariate forecast, multiple linear regression or composite forecasting) and the “*forecasting strategy*”, i.e. the forecasting method together with the necessary forecasting parameters, have to be specified in the master forecast profile. Thus different master forecast profiles can be defined and later on assigned to various CVCs. An overview of univariate forecasting strategies, which are important for the Frutado case, will be given below.

The actual forecast itself is created by combining a master forecast profile with a CVC (4b). Depending on the aggregation of the CVC, forecasts on different hierarchical levels can be made. For storage purposes, they are (in the background) automatically disaggregated to the highest level of detail according to the disaggregation rules defined when setting up the master data. The results of the statistical forecasting process can be viewed in the interactive planning screen.

They can be refined in the following. For example, human judgment can be incorporated (4c). The planning of promotions can itself be supported by SAP APO DP. But other measures of judgment like manual modifications can be applied to “fine-tune” the demand plan, too. Furthermore, forecasts of different sources can be reconciled in a subsequent collaborative, consensus process (4d). If the forecast is satisfying for all decision makers involved, the demand plan can be released to SNP (4e). Thus, it will build the basis for subsequent planning modules and planning processes like SNP, PP/DS and Deployment.

The forecasts have to be stored in order to be able to calculate forecast errors and control the quality of the forecasts. At some point in time actual demand (of the first forecasting period) will be observed. Thus, the historical data have to be updated before a new forecast cycle can start (4f). Now the forecast accuracy can in fact be calculated and monitored (4g). This can automatically be done in the background. If the forecast quality is not satisfying for some key figures and CVCs, the demand planner should be warned by (again automatically generated) alerts and triggered to redefine the corresponding forecast models if necessary (3). This alerting process is

called “management by exception” (4h). Independently on a change of the forecast model, the time series should be corrected if an one-time exceptional event has disturbed the regular pattern of the last observed time period. Furthermore, the master data should be revised, e.g. if some new products or selling points have been introduced or some old ones delisted (4a).

Forecasting Strategies

SAP APO DP offers a variety of forecasting methods for different types of demand models. Time series based methods are called “univariate” methods. Besides, causal methods are available and a combination of different methods is possible, denominated as “composite forecasting”. Table 4.2 shows an incomplete overview of the univariate forecasting methods. They are numbered as different “forecasting strategies” within APO. If some information is recognizable, the underlying demand model is also shown in the table. For further details – also on the remaining, not mentioned strategies – the reader is referred to Hoppe (2007, Chap. 4.2.3) and SAP (2011b).

Strat.	Model	Denomination	Remark
10, 11	A	Constant Model with 1st Order Exponential Smoothing	single exponential smoothing
12	A	Automatic Adaptation of the Alpha Factor	single exponential smoothing; best α is selected from a grid
13	A	Moving Average Model	
14	A	Weighted Moving Average Model	
20, 21	A+B	Trend/Seasonal Models with 1st Order Exponential Smoothing	method of Holt
22	A+B	Models with 2nd Order Exponential Smoothing	(Brown’s) double exponential smoothing
23	A+B	Models with 2nd Order Exponential Smoothing & Automatic Adaptation of the Alpha Factor	like 22, but α adapted according to error tracking signal (see Chap. 4.3.1)
30, 31	A*C	Trend/Seasonal Models with 1st Order Exponential Smoothing	like method of Winters, but without trend
35	A+B*C	Seasonal Linear Regression	RTMAD for seasonal coefficients first
40, 41	A+B*C	Trend/Seasonal Models with 1st Order Exponential Smoothing	method of Winters
50-53	A((+B)(*C))	Automatic Model Selection Procedure 1	uses autocorrelation to test for demand model
54, 55	A((+B)(*C))	Automatic Model Selection Procedure 1	season or trend are manually set, remaining pattern tested
56	A((+B)(*C))	Automatic Model Selection Procedure 2	α, β, γ are varied on a grid; best models, methods, and parameters selected
60	A((+B)(*C))	Copy History	naive forecast
70	A((+B)(*C))	Manual Forecast	$\hat{a}, \hat{b}, \hat{c}$ are set manually
94	A+B	Simple Linear Regression	

Table 4.2
(Incomplete) overview of the SAP APO DP forecasting strategies and demand models (A((+B)(*C)) abbreviates A, A+B, A*C, A+B*C; see also Fig. 4.1)

Most of the strategies are self-explaining. It is noteworthy that the automatic adaptation of the smoothing constant α according to the error tracking signal (see Chap. 4.3.1) is applied to a double exponential smoothing method which also expresses the trend.

The “Automatic Model Selection Procedure 1” applies autocorrelation tests in order to determine fitting demand models. If demand is not recognized as being sporadic, strategy 51 tests (only) for a trend, strategy 52 (only) for seasonal influences, and 53 for both simultaneously. If the tests are successful the corresponding demand models and 1st order exponential smoothing methods are selected, otherwise a constant model is applied. Strategy 54 assumes a season given and tests for a further trend. Strategy 55 assumes a trend given and tests for a further season. Strategy 50 tests for all of these demand models without preceding assumptions. The smoothing constants are not optimized.

The “Automatic Model Selection Procedure 2” given by strategy 56 varies the α , β , and γ of the different 1st order exponential smoothing models and methods on a discrete grid (per default between 0.1 and 0.5 in intervals of 0.1) and then selects the best variant (per default according to the MAD). This ex-post simulation is said to be more precise than Automatic Model Selection Procedure 1, but of course also more time-consuming. However, also some other non-exponential smoothing models like seasonal linear regression are checked (see Hoppe 2007, p. 140).

Aggregation and Disaggregation Strategies

SAP APO stores the time series and the forecasts of a forecast hierarchy in the most detailed granularity that has been defined. If they are necessary in an aggregate manner, they are aggregated, processed somehow (e.g. by generating a forecast) and then disaggregated again. Thus both aggregation and disaggregation rules have to be defined. For the demand planning of the Frutado case it will be sufficient to simply aggregate by summation. Thus in the following we mainly concentrate on more complex rules for disaggregation.

SAP APO differentiates between the disaggregation of “key figures”, that are defined by combining master data objects like “Product” and “Location” (remember the definition of CVCs before), and disaggregation of time. A simple reason for this distinction is that the key figures of the forecast hierarchy have to be updated only seldom, whereas time and the working calendar have to be updated very often, usually every day. Rules for the disaggregation of key figures are denominated as the “calculation types” of the key figures. In contrast, rules for the disaggregation of time are called “time-based disaggregation types”.

Some important calculation and time-based disaggregation types are sketched in the following. The SAP APO specific identification code that marks the respective disaggregation rule is stated in brackets after the name of the type. We start with the *calculation types* of key figures:

Pro rata (“S”): If experience on the proportional shares of an aggregate group’s individual members already exists from the past, these shares are taken as a basis for the disaggregation process.

Based on another key figure (“P”): Again shares build the basis for the disaggregation. However, these are now gained from the key figure of another group. For example, if a new product is introduced in a certain region, the regional customers’ shares of a similar product, that has already been sold in the past, might be applied.

Average of key figures (“A”): This type is, for example, used for percentage measures or per unit measures like sales prices. For aggregation purposes, they are averaged at runtime to constitute the aggregate value of the forecast hierarchy’s next higher level. When disaggregating, this aggregate value is taken over for all lower level members (Hoppe 2007, p. 109f.).

Similar rules are implemented in the *time-based disaggregation types* when disaggregating an aggregate period into several shorter ones:

Proportional distribution (“P”): If possible, time-based weighting factors are calculated and used for disaggregation. For example, if already existing daily forecasts shall be updated, but only an aggregate new forecast of a whole week is available, this would be assigned to the days of the week according to the old proportional shares.

Based on another key figure (“K”): Another key figure helps to calculate the time-based weighting factors. E.g., in the example percentages of daily sales in the past could be used instead.

Equal distribution (“E”): The aggregate period’s total quantity is equally spread over the corresponding sub-periods.

Examples for these different rules and types, respectively, are given by [Figure 4.3](#), which has been adapted from Christ (2003, Fig. 19). Since the rules show similar effects, the example does not differentiate between the disaggregation of key figures and time. It just differentiates between the total quantity of a single, aggregate group and the detailed quantities of three individual members of this group. In the past (i.e. *before* the disaggregation process) individual quantities – e.g. forecasts – of 100, 150, and 150 units had been observed for the members of the group, summing up to a total of 400 units. The new forecast estimates that only 200 units will be demanded for the group as a whole in the future. Thus, after the disaggregation the individual forecasts of the members of the group are only allowed to sum up to 200. For the type “based on another key figure” the characteristics of a second group, e.g. representing sales instead of forecasts, with the same number of members, but a total (past) quantity of 300 have been used.

4.5.2 Modeling with SAP® APO

After this basic introduction to SAP APO DP it now shall be demonstrated how the demand planning tasks of the Frutado case are modeled with SAP APO DP.

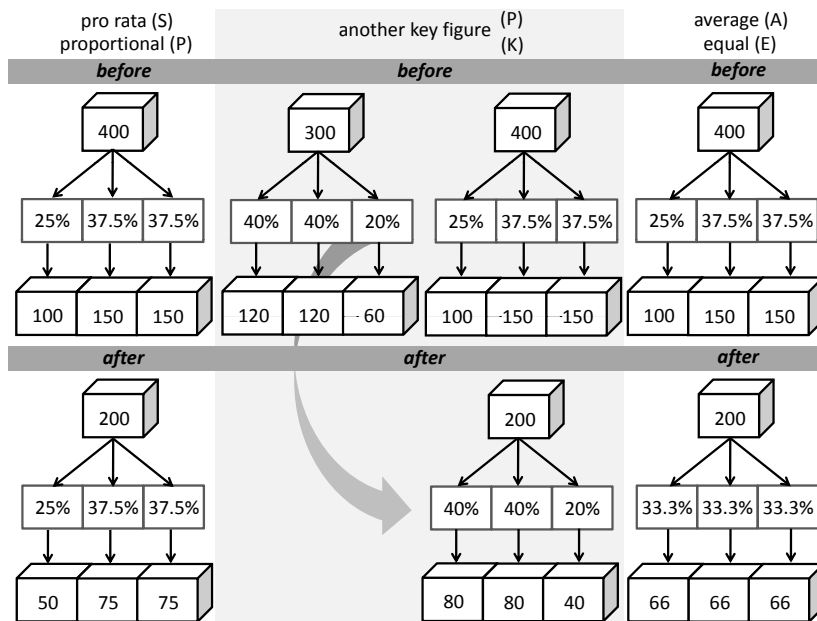


Figure 4.3
Examples for different disaggregation types of SAP APO.

Modeling Time and Characteristic Value Combinations

A working calendar has to be established that reaches two years into the past and one year into the future. Weekends and holidays are excluded such that customer orders can only arrive or be due on working days. It consists of the following parts:

- (i) The *two years of the past* are subdivided into time buckets of *days* in order to be able to make the detailed time series available, i.e. the (from a demand planning perspective) already processed customer orders.
- (ii) The *first week* of the future also consists of *daily* buckets to model customer orders accurately that already have arrived, but have not been delivered.
- (iii) *Weeks 2-4* of the future are also represented on a *daily* time grid to build a detailed basis for the short-term planning modules Deployment and PP/DS.
- (iv) Finally, for *weeks 5-52* a granularity of *weeks* is enough because they only serve the mid-term aggregate planning within SNP. Note, that weeks 27-52 would actually not be necessary for SNP. They are just considered for demonstration purposes in order to be able to show forecasts for a full seasonal cycle of 52 weeks.

This is also the most detailed granularity that will be possible for all subsequent planning activities. However, if necessary, these time buckets can

further be aggregated, e.g. to weeks, months, quarters, and years. For example, for usage within SNP the first month has to be aggregated from days to weeks. Period (i) will be available as key figure “Historical Data” (9AVHISTORY), period (ii) as key figure “Customer Orders” (9ADMD1) and periods (iii) and (iv) as key figure “Demand Forecasts” (9ADFCST) within the APO planning area (FRU_AREA_PLAN).

As we have seen, there are 19 final items in the Frutado case. These are modeled in APO as master data objects of the type “Product”. The 60 customers and three DCs are modeled as master data objects of the type “Location”. They actually constitute 1197 “Characteristic Value Combinations (CVCs)” of the granularity location-product.

However, because of the unique assignment of customers to DCs and because we do not need more detailed forecasts than on DC level, it will be sufficient to just consider the three DCs (named FRU_DC_01, FRU_DC_02, FRU_DC_03) in the following. By summing the demand of all customers of a DC, we get $3 \cdot 19$ historical time series with the daily demand of the last two years. If needed, these can be further aggregated by time, product or location.

Aggregation

To give SNP the necessary input, it would be sufficient to further aggregate the time from buckets of days to buckets of weeks. However, statistical analyses of these time series (outside of APO) have shown that the quality of the forecasts will become even better if there is a further aggregation with respect to locations and products.

Accordingly, total demand of all three DCs together should be used as a basis to make the forecasts. Furthermore, the 19 final items should be aggregated to 5 product groups according to the following rules:

Group **FRU_CONST1_DP**: final items 02, 05, 12, 13, 15, 16, and 17

Group **FRU_CONST2_DP**: final items 03 and 19

Group **FRU_CONST3_DP**: final item 18

Group **FRU_SEASON1_DP**: final items 01, 04, 06, and 10

Group **FRU_SEASON2_DP**: final items 07, 08, 09, 11, and 14

The last two groups aggregate the final items showing seasonal peaks in summer and winter, respectively. Furthermore, except for final item 18, for final items with level demand an aggregation to groups also appears beneficial. The reason again is that the variance can be reduced and forecast errors rather balance out if more observations belong to a time series.

Summing up, 5 time series remain, that have to be forecast, each of them aggregating demand over all DCs and all products of their respective group. Aggregation over time has to be defined separately in SAP APO in a so-called

“forecast profile”. Here, days shall be aggregated to weeks. Doing this, the actual granularity of the CVCs has been defined. In the end, 5 time series with a history of 104 weeks are left (see Fig. 4.4).

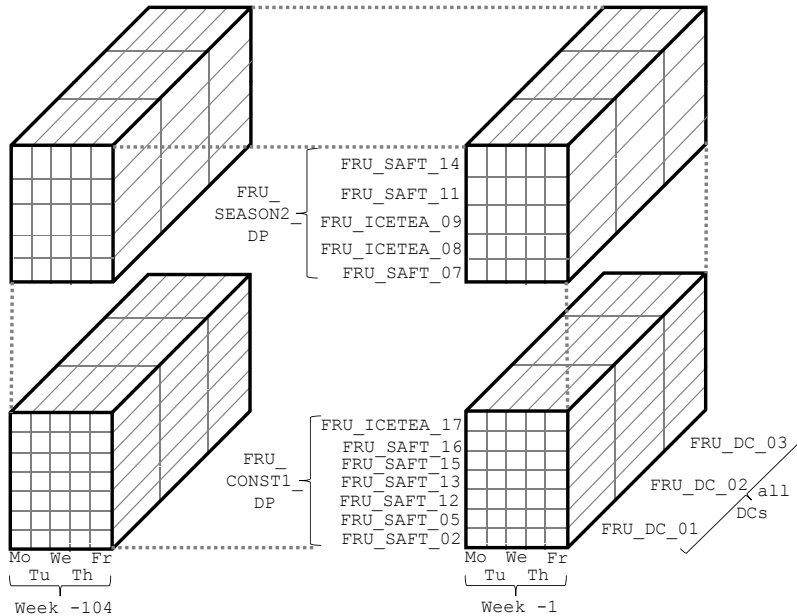


Figure 4.4
“Demand cube”
illustrating the
aggregation of
historical demand
from final
items/individual
DCs/days to product
groups/all DCs/weeks.

Note, that the simplest aggregation rule possible has been applied: we just had to add up the objects of a higher granularity group. For many applications this might not be sufficient. For example, when regional forecasts of sales prices have to be aggregated, they should be weighted with the estimated demands of the regions (see e.g. Roitsch and Meyr 2008). Then more refined rules have to be selected or defined.

Disaggregation

For this aggregate granularity of the 5 time series, forecasts are made. However, later on they are needed in a higher detail again by SNP, PP/DS and Deployment. Besides we have learned that APS in general and SAP APO especially prefer to store the forecasts in the highest possible detail, anyway. Thus, strategies have to be defined how to disaggregate forecasts to lower planning levels of higher granularity again.

As shown above, SAP APO offers several standard strategies to disaggregate data. Default strategies can be specified for both a disaggregation of time and a disaggregation of the CVCs, which have been defined on basis of the location-product characteristics. In APO, the disaggregation of the forecasts itself is kicked off by “calculating proportional factors” (see the corresponding datango learning unit). For modeling the Frutado forecast disaggregation within APO the options “calc. proportions for = proportional

factor”, “basis for proportion = history” and “proportion calc. type = fixed proportions” have been chosen.

This means that the total forecast of all DCs is distributed among the DCs according to the ratios (“proportions”) of this spread that have been observed for the historical times series of demand in the past. The same rule applies for the disaggregation of product groups. Thus it is assumed that the ratios of final items’ demand within a group, that were valid for a certain DC in the past, will also hold for this DC in the future. For the disaggregation over time, a uniform disaggregation is chosen such that the demand of a week is equally spread over the working days of this week.

Altogether, five forecasts for five groups are necessary. They are disaggregated to individual DCs and final items. For weeks 2–4 they are also disaggregated to days. SAP APO already does this in the background. After release of these forecasts to SNP (for each DC), the forecasts are also available for other planning modules like PP/DS and Deployment. If SNP or PP/DS need the forecasts, the forecasts are aggregated to the desired granularity again.

Alternatively, it would have been reasonable to generate the aggregate forecasts of the five groups only starting from week 5 on. Instead, weeks 2–4 could have been forecast directly on the highest level of detail (assuming that the time-series are more reliable on the short-term). For doing this, further forecast profiles had to be defined and fitting forecasting methods had to be selected. However, since we only wanted to demonstrate the basic principle, we disregarded this opportunity.

Finally it will be explained, which demand models and forecasting methods have been chosen for the five aggregate groups.

Selecting Demand Models and Forecasting Methods

SAP APO can automatically generate proposals for selecting a demand model, forecasting method and/or the necessary parameters. Unfortunately, these proposals do not perform very well in the Frutado case. Thus other methods and parameter settings have been selected “manually” and tested by ex-post simulation. Some of them clearly show a better performance. In the end, the following ones have been implemented:

As already mentioned groups FRU_CONST1_DP, FRU_CONST2_DP and FRU_CONST3_DP show an almost constant demand. Thus, choosing a demand model is not a big issue. Among the appropriate forecasting methods for level demand, the SAP APO DP strategy 12 has been applied. This is a single exponential smoothing method where the smoothing constant α is automatically optimized by varying it on a discrete grid. The lower bound of the grid has been set to 0.1, the upper bound to 0.5 and the step size to 0.05. The best α is then selected by an ex-post simulation on basis of the RMSE error measure. As mentioned above, the time granularity of the forecast, which is weeks in our example, has to be defined in the same forecast profile of APO. Alternatively, also SAP APO DP strategy 23, which is an

implementation of the adaptive-response-rate single exponential smoothing method (see Chap. 4.3.1), would perform very well in the Frutado case.

For the remaining two product groups a level demand model with a (multiplicative) season seems appropriate. Within the appropriate forecasting methods, SAP APO DP strategy 35 performs best, which is a seasonal linear regression (see e.g. Hoppe 2007, p. 129 and Chap. 4.3.3). Thus, only the length of the seasonal cycle $L = 52$ has to be defined as a parameter.

Questions and Exercises

1. Which disaggregation rules do you know? What do you think are their advantages and disadvantages?
2. Why do we have to build five different product groups?

4.6 Implementation and Disaggregation of Results

The results of the DP process are the forecasts that are – as mentioned above – already stored on the highest level of detail according to the disaggregation mechanisms that have been defined in Chapter 4.5.2. When the demand plan has been released to SNP as shown in step (4e) of Figure 4.2, it is also available to PP/DS and Deployment. Access is enabled for these planning modules to the respective time periods and aggregation levels that are necessary there.

4.7 Demand Planning Learning Units

4.7.1 Overview

The Frutado CD offers learning units for usage of SAP APO DP that can be accessed via a regular WWW-browser. The learning units explain the different steps that are necessary to execute Frutado’s DP processes within SAP APO. As shown in Figure 4.5 the processes are roughly subdivided into the categories “Preparation of Demand Planning Structures and Historic Data”, “Statistical Forecasting”, “Judgmental Forecasting” and “Release of the Forecast”.

Learning units in oval boxes prepare information and planning results that are input for the other planning modules of the Frutado case. Thus they are basic for the case as a whole. In contrast, learning units in rectangular boxes provide supplementary information on the application of SAP APO DP, which is not necessary to execute and understand the learning modules and planning logic of the other planning modules like SNP, PP/DS etc. Consequently, two streams of learning units exist, a “*basic stream*” (see also the introduction to this book and Haub 2008) containing the seven essential (oval) units and an “*in-depth stream*” (see also Zier 2009) containing the four supplementary (rectangular) units. The first one should be worked through before proceeding to other planning modules. Here, also the sequence of the learning units is important and should not be changed. The second one can

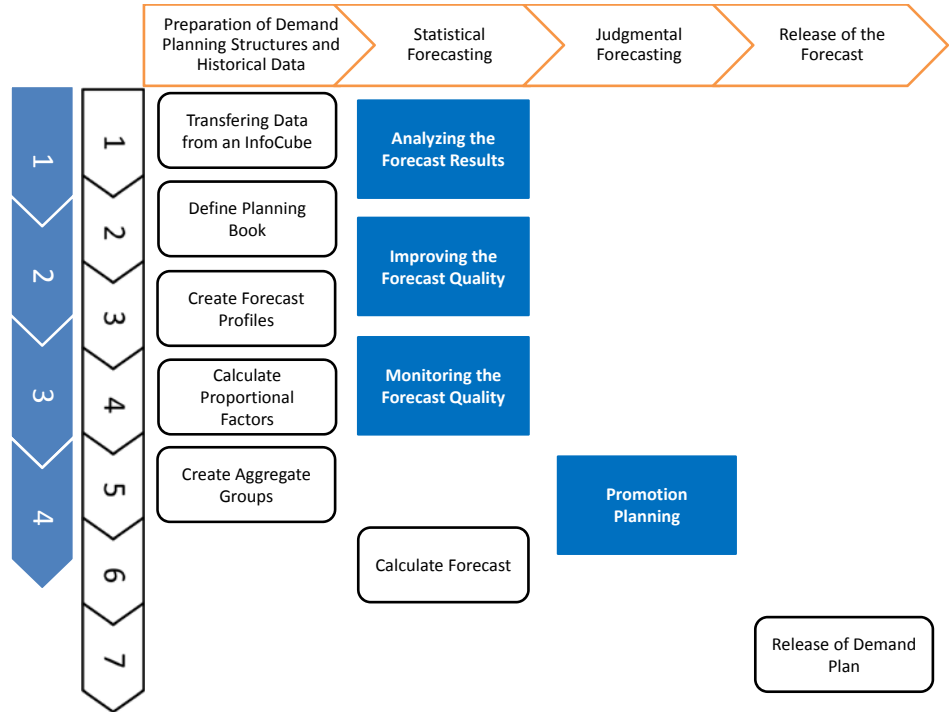


Figure 4.5
 Overview of the DP learning units for both the basic stream and the in-depth stream (see Zier 2009, Fig. 21).

be executed at any time. The sequence of its learning units follows the logic of the demand planning process. However, it is not necessary to stick to this sequence to understand their learning targets.

The learning units can be used in two different modes, a demo and practice mode, which are already offered by the datango trainer as the underlying training software. The *demo mode* acts like a movie showing the mouse clicks and typings that are necessary to configure and use the APO DP module for Frutado. Additional explanations on these actions are shown in bubbles overlaying SAP’s graphical user interface. The movie has a predetermined speed, but can be paused at any time. To enable some exercises without real access to the SAP APO system also an interactive *practice mode* exists. Here the same learning units are provided. However, the user has to execute the necessary actions, i.e. clicks or typings that have been shown in the demo mode, by himself. The datango trainer merely simulates the SAP APO system. It gives some basic support for the user and checks whether the clicks and typings are correct.

In the next section we give an overview of the learning units of the basic stream of the DP module before proceeding to the in-depth stream in Chapter 4.7.3.

4.7.2 Basic Stream

In a first step (learning unit 1 of the basic stream) the *data have to be loaded from an InfoCube* into the planning area and a planning version of DP. It is assumed that the InfoCube already exists and is filled with data. The

InfoCube contains the information on the time series, i.e. the stored key figures “historical data” (9AVHISTORY) and “customer orders” (9ADMDP1). In the Frutado case, the Infocube is denoted as FRU_ICUBE. It has the planning version 000 assigned. The key figures are loaded into the planning version FRU_PV_DP of SAP APO DP’s planning area FRU_AREA_PLAN.

To create a customized environment and data view for further usage, a *planning book should be defined*. All key figures and further characteristics of the planning area can be made available in the planning book. However, the key figures should be adapted to the needs of the respective users. In our case only a single user is working, and thus only a single data view FRU_DP_VIEW_DP is assigned to the newly created planning book FRU_DP_PLAN_DP. The data view defines the appearance of the interactive planning interface and the planning horizon. Additionally, it contains information about the structure of the time buckets. This is done by linking it with identification tags of planning buckets profiles of the future (FRU_DP_FUT) and the past (FRU_DP_PAST) that already have been specified in an earlier stage (not shown as a learning unit; see Chap. 4.5.1 and step (1) of Fig. 4.2). In the Frutado case, all key figures of the planning area are assigned to both the planning book and this single data view.

In a third step, the *master forecast profiles have to be defined* as already indicated in Chapter 4.5.1. First, the master forecast profile FRU_CONST_DP for products with almost constant demand is created. It is marked as a key figure of the type “forecast” (9ADFCST). The forecast horizon, the section of the time series to be used, and the length of the storage time buckets (weeks) have to be specified. Additionally, the type of forecasting model (univariate) has to be selected. A name has to be chosen (here FRU_FCST_CONST_DP) in order to allow a further specification of the forecasting method assigned to the profile. As described above the forecast strategy 12 is selected for this type of products, and the corresponding parameters and error measures are set. Secondly, the master forecast profile FRU_SEAS_DP has to be defined for seasonal products. The proceeding is similar. However, this profile is assigned to a(n again univariate) forecasting model denoted as FRU_FCST_SEAS_DP, which parameterizes forecasting strategy 35.

Chapter 4.5.1 has shown that SAP APO offers the calculation type “based on another key figure” (P) for disaggregating the key figures. To be able to put this into practice, the corresponding *proportional factors*, e.g. the 40%, 40%, and 20% of the gray part “another key figure” of Figure 4.3, have to be *calculated* first. This calculation and the specification of the required parameters is demonstrated in learning unit 4 of the basic stream. The most important parameters have already been mentioned in section “Disaggregation” of Chapter 4.5.2. The key figure “proportional factor” of the current planning area (FRU_AREA_PLAN) and planning version (FRU_PV_DP) has to be defined as a target of the calculation, the key figure “history” has to be set as the basis for calculating the proportions and the proportion calculation type has to be specified as “fixed proportions”.

Furthermore, it has to be determined which periods of the past should serve as the basis for the calculation of the proportional factors. Additionally, the periods of the future have to be specified, in which these proportional factors should automatically be applied for disaggregation. Only after explicitly executing the calculation, the proportional factors are available for further usage.

Learning unit 5 shows how to *create* the five different *aggregate groups* FRU_CONST1_DP, . . . , FRU_SEASON2_DP that have been introduced in section “Aggregation” of Chapter 4.5.2 to make as small forecast errors as possible. Here, the different levels of the forecast hierarchy can be defined. In the interactive demand planning screen the (already created) data view FRU_DP_VIEW_DP of the planning book’s (also already existing) folder FRU_DP_PLAN_DP has to be selected. To establish the first group FRU_CONST1_DP, the planning version (FRU_PV_DP) has to be specified again. The master data “APO-location” has to be chosen and a template for locations has to be generated that contains all three DCs. Analogously, a template for the master data “APO-product” has to be built, which contains all seven products of this steady demand group. As a next step, the name FRU_CONST1_DP can be assigned to the group. Here, the link between the aggregate locations and the aggregate products is actually established as a new CVC of a higher aggregation level. The other four groups are generated the same way. However – to save typing efforts – the template for locations can be re-used again. Finally, the newly created groups have to be assigned to the current master profile in order to become available there.

The actual *forecasts* are *calculated* in the sixth learning unit. Again the interactive demand planning screen is used. To generate a forecast the aggregate group has to be opened so that the data are loaded. Next, one of the earlier defined master forecast profiles (located in the folder “settings”) has to be linked to the group. For example, the profile FRU_CONST_DP has to be chosen for the aggregate group FRU_CONST1_DP. Eventually, the forecast appears in the corresponding row of the interactive demand planning screen. [Figure 4.6](#) illustrates this for the above mentioned group and profile. The whole procedure is repeated for the other four groups.

The last learning unit of the basic stream demonstrates the *release of the demand plan*. This procedure is very simple. The current planning area FRU_AREA_PLAN, planning version FRU_PV_DP and the forecast key figure 9ADFCST have to be specified as sources. Additionally, the target planning version (FRU_PLAN2), the forecast horizon to be released, and the corresponding DCs have to be specified. After committing the release, the forecasts are available to the SNP module, but also to the other APO modules relying on them.

4.7.3 In-Depth Stream

The learning unit “*Analyzing the forecast results*” of the in-depth stream gives further information on the interactive planning desktop and its usage. By

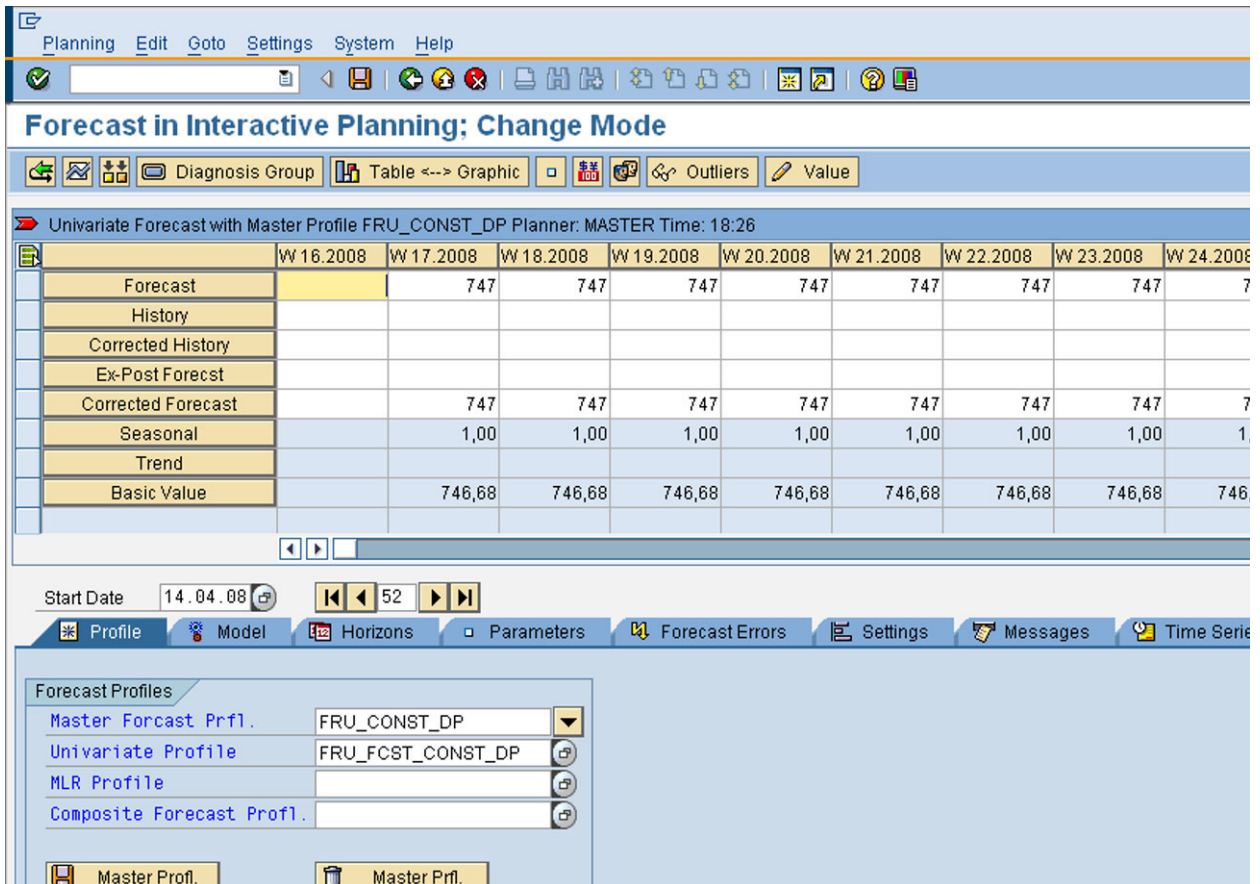


Figure 4.6
Interactive demand planning screen showing the forecast for group FRU_CONST1_DP and profile FRU_CONST_DP.
© Copyright 2011. SAP AG.
All rights reserved

entering a “change mode” the resulting forecasts can be both analyzed and adapted. A selection part on the left hand side of the interactive planning desktop allows the choice of different locations, products or groups to be considered. On the right hand side, a workspace is available as already shown in Figure 4.6. Supplementing the text-based view of Figure 4.6, also graphical representations of the historical time series, the forecast itself, an ex-post forecast etc. are offered. This learning unit uses the aggregate group FRU_SEASON1_DP as an example, thus also illustrating seasonal influences. According to the seasonal linear regression chosen in the profile FRU_SEAS_DP, the forecast constitutes of a basic value, an additive trend component (approximately 0) and multiplicative seasonal coefficients. The interactive planning desktop further allows to

- change the forecast profiles,
- select other forecasting models, methods, and their corresponding parameters,
- adjust the forecast horizon and the history horizon of the time series,
- select the forecast error measures that are used for ex-post forecasting,

- analyze the resulting forecast errors,
- and execute additional special purpose analyses, e.g. for outlier detection and correction due to singular events or to define like-profiles for products without history.

The second learning unit of the in-depth stream tests whether the *forecast quality* can be *improved* by using alternative forecasting strategies. Again the group FRU_SEASON1_DP serves as an example with seasonal linear regression (strategy 35) as a benchmark. The forecast quality shall be judged using MAPE as the primary quality measure. As alternative strategies the 1st Order Exponential Smoothing (strategy 30 for steady demand without trend; $\alpha = 0.1$ and $\gamma = 0.1$ are set manually), Automatic Model Selection Procedure 1 (strategy 54 with the options “seasonal model + test for trend”) and Automatic Model Selection Procedure 2 (strategy 56; α, β , and γ are varied on a discrete grid between 0.1 and 0.5 with a step-size of 0.1) are considered. The tests show that the time series is too short for applying 1st Order Exponential Smoothing so that no ex-post forecast can be generated for strategy 30. Strategy 54 performs worst. Strategies 35 and 56 show exactly the same forecast errors. This indicates that strategy 56 automatically selects seasonal linear regression as its best choice. Probably, the time series is also too short for the exponential smoothing methods that are further checked by strategy 56.

The learning unit “*Monitoring the Forecast Quality*” is subdivided into the three lessons “Creating an alert-forecast-profile”, “Creating a diagnosis group” and “Using the alert-feature” that need to be executed consecutively. The first lesson defines an alert profile FRU_ALERT_DP that allows to check forecasts with respect to quality measures like MAD, MAPE or RMSE. These checks may also be defined for some or all of the aggregate groups of the forecasting hierarchy. The second lesson shows how default values for alert thresholds are defined in master forecast profiles via so-called “diagnosis groups”. For example, a diagnosis group FRU_DIAGN_GROUP has been created as part of the master forecast profile FRU_SEAS_DP. This diagnosis group is assigned to the forecast model FRU_FCST_SEAS_DP (see learning unit 3 of the basic stream) so that an upper limit of 400 can be set as a default threshold for the MAD, of 1 for MAPE etc. In the third lesson the control of the thresholds and visualization of alerts are demonstrated in the interactive demand planning desktop. Both the alert profile and the diagnosis group have to be assigned to the planning book first such that the thresholds are monitored and alerts are displayed by the system. Later on, in regular use, alerts appear as warning symbols in the forecast error section of the interactive planning desktop if the pre-defined thresholds are violated. More detailed reasons for the violations can be accessed through additional masks.

The last learning unit “*Promotion Planning*” is ought to demonstrate how judgment can be supported by APS. Promotion planning serves as an example, i.e. for a certain product group (FRU_SEASON1_DP) a temporary price

discount of 10% is offered for a duration of eight weeks. From experiences of the human planners in the past it is known that a price discount of 10% usually increases demand by 5%. This human knowledge shall supplement the ordinary statistical forecast of SAP APO DP that we have seen in the preceding learning units. Thus, additionally to the basic statistical forecast a supplementary forecast of the promotion's extra-demand shall be generated if such a promotion has been defined and activated.

In order not to affect the basic stream, the planning area FRU_AREA_PLAN and its elements have been copied to a new planning area FRU_AREA_PLAN_PROMOTION. The aggregate product group FRU_SEASON1_DP has accordingly been transformed into a corresponding product group FRU_SEAS1_DP_PROMO. New key figures "9APROM1 – Promotion 1" and "9ATOTFC – Total Forecast" had to be introduced in order to represent the additional forecast for promotions and the total forecast, which is the sum of the basic statistical forecast and the additional forecast for promotions. They have also been made available in the copied planning book, planning version, and data view. The necessary steps are not explicitly shown in the learning unit. They should already be known from the preceding learning units.

The "Promotion Planning" learning unit is again subdivided into several lessons that have to be executed subsequently. The first lesson "Define the promotion level for the planning area" basically determines for which master data (location or product) a promotion should be valid. Since we want to allow the promotion for all locations, but only for the product group FRU_SEAS1_DP_PROMO, the so-called "promotion level" of the new key figure 9APROM1 has to be set to "9AMATNR – APO product". The second lesson "Create a promotion base" introduces the concept of promotion bases. A promotion base bundles information about the promotion key figure and additional characteristics like the corresponding promotion level and time span of validity in a single denominator that may be referenced several times, for example, to simulate the effects of multiple promotion alternatives (e.g. decreasing the price by 20% instead of 10%). In our example, a promotion base called FRU_PROMO_BASE is generated.

As the third lesson "Create a promotion" shows, the promotion base helps to create the promotion itself, i.e. to announce SAP APO where and when a certain promotional activity might take place. For doing this, we have to change from the interactive planning desktop of demand planning to the interactive planning desktop of promotion planning. There, a new promotion can be created by specifying its underlying promotion base (FRU_PROMO_BASE), its type (whether it shows absolute or percentage effects), starting date and periods of validity (8 weeks). Note, that the time span of validity of the promotion base is set for the promotion base in general (in the example more than ten years), whereas here the usually quite short validity of a concrete promotion has to be defined.

The promotion has been created, but not yet been fully specified. It still has to be made known, which concrete products or product groups are involved, and which effects on demand the promotion has. This is the task of the last lesson “Edit the planning data for the promotion”. In the interactive planning desktop of promotion planning, the aggregate group FRU_SEAS1_DP_PROMO has to be assigned to the promotion. After doing this the corresponding products of the group are displayed and the percentage increase of demand, which will presumably be generated by the promotion, can manually be specified for each time period of the promotion. The promotion has to be marked as being “Planned, in the future” in order to be activated and to show effects in the interactive planning desktop of demand planning. There the total forecast could be calculated and released to SNP.

Questions and Exercises

The following questions should be answered while you are working through the learning units.

1. Which quality measures does an alert profile offer?
2. Which parameters have to be set for the forecasting strategies 12 and 35?
3. Where can the seasonal coefficients be read that SAP APO has created during its forecasting process?

Bibliography

- Box, G. E. P.; Jenkins, G. M.; Reinsel, G. C. (2008) *Time Series Analysis - Forecasting and Control*, Wiley Series in Probability and Statistics, Wiley, Hoboken, NJ, 4th ed.
- Brown, R. G. (1959) *Statistical Forecasting for Inventory Control*, McGraw-Hill, New York
- Christ, S. (2003) *Implementierung der Supply Chain einer Modellfirma in SAP APO 4.0 - Demand Planning und Supply Network Planning*, Studienarbeit, Technical University of Darmstadt, Germany
- Gardener Jr., E. S. (1984) *The strange case of the lagging forecasts*, Interfaces, vol. 14, no. 3, 47–50
- Hanke, J. E.; Wichern, D. E. (2008) *Business Forecasting*, Pearson/Prentice Hall, New Jersey, 9th ed.
- Haub, C. (2008) *Entwicklung eines Lehrkonzepts zur Vermittlung von Modellierungsmöglichkeiten in SAP APO und Umsetzung im Rahmen des Demand Planning Moduls*, Diplomarbeit, Technical University of Darmstadt, Germany

-
- Holt, C. C. (1957) *Forecasting seasonals and trends by exponentially weighted moving averages*, Research Memorandum 52, Office of Naval Research
- Holt, C. C.; Modigliani, F.; Muth, J. F.; Simon, H. A. (1960) *Planning Production, Inventories, and Work Force*, Prentice Hall, Englewood Cliffs, New Jersey
- Hoppe, M. (2007) *Absatz- und Bestandsplanung mit SAP APO*, Galileo Press, Bonn
- Hyndman, R. J.; Koehler, A. B.; Ord, J. K.; Snyder, R. D. (2008) *Forecasting with Exponential Smoothing – The State Space Approach*, Springer Series in Statistics, Springer, Berlin et al.
- Kilger, C.; Wagner, M. (2008) *Demand planning*, in: H. Stadtler; C. Kilger (Eds.) *Supply Chain Management and Advanced Planning*, Springer, Berlin et al., 4th ed., 133–160
- Makridakis, S. G.; Wheelwright, S. C.; Hyndman, R. J. (1998) *Forecasting: Methods and Applications*, Wiley, New York, 3rd ed.
- Meyr, H. (2008) *Forecast methods*, in: H. Stadtler; C. Kilger (Eds.) *Supply Chain Management and Advanced Planning*, chap. 28, Springer, Berlin et al., 4th ed., 505–516
- Pegels, C. C. (1969) *Exponential forecasting: some new variations*, Management Science, vol. 15, no. 5, 311–315
- Roitsch, M.; Meyr, H. (2008) *Oil Industry*, in: H. Stadtler; C. Kilger (Eds.) *Supply Chain Management and Advanced Planning*, Springer, Berlin et al., 4th ed., 399–414
- SAP (2011a) *SAP Help Portal*, URL http://help.sap.com/saphelp_scm70/helpdata/en/78/bca554465f11d3983a0000e8a49608/frameset.htm, date: January 2011
- SAP (2011b) *SAP Help Portal*, URL http://help.sap.com/saphelp_scm70/helpdata/en/ac/216b6e337b11d398290000e8a49608/frameset.htm, date: January 2011
- Silver, E. A.; Pyke, D. F.; Peterson, R. (1998) *Inventory Management and Production Planning and Scheduling*, Wiley, New York, 3rd ed.
- Tempelmeier, H. (2008) *Material-Logistik – Modelle und Algorithmen für die Produktionsplanung und -steuerung in Advanced Planning Systemen*, Springer, Berlin, Heidelberg, New York, 7th ed.
- Trigg, D. (1964) *Monitoring a forecasting system*, Operational Research Quarterly, vol. 15, 271–274

- Wagner, M. (2005) *Demand Planning*, in: H. Stadtler; C. Kilger (Eds.) *Supply Chain Management and Advanced Planning*, chap. 7, Springer, Berlin et al., 3rd ed., 139–157
- Winters, P. (1960) *Forecasting sales by exponentially weighted moving averages*, *Management Science*, vol. 6, no. 3, 324–342
- Zier, S. (2009) *Erstellung von Lerneinheiten über vertiefende Möglichkeiten der Modellierung im Demand Planning Modul des SAP APO*, Studienarbeit, Technical University of Darmstadt, Germany

Chapter 5

Master Planning - Supply Network Planning

Hartmut Stadtler¹

¹ University of Hamburg, Institute for Logistics and Transport, Von-Melle-Park 5, 20146 Hamburg, Germany

There are a number of decisions which have to be made in the medium-term because these have to be prepared some weeks in advance (like an additional shift on Saturdays which has to be agreed to by worker's representatives) or take some time to become effective (like building up seasonal stocks). Another reason for placing decisions at the medium-term planning level (instead of short-term) is the scope of consequences. As an example consider a manufacturer with several production sites with similar capabilities. Decisions regarding the allocation of production quantities to sites and directives which site will serve which market will have consequences on the profitability of each site and should be made centrally. Note that short-term production planning and scheduling is often done decentrally.

In the following (Section 5.1) we will give an introduction to medium-term planning as discussed in the literature. Also, a very simple linear programming (LP) model is introduced to get a first idea of what later on will be extended to the medium-term planning model of the Frutado company.

Section 5.2 provides some details of the modeling features of today's commercial mathematical programming software and its solution capabilities, because these are the backbone of solving medium-term planning models to optimality. Often a user of medium-term planning models may have a choice of either using optimization or a heuristic. Hence, this section starts with a simple exact algorithm – named backward scheduling – which turns into a heuristic if applied sequentially in case of several bottlenecks. The solution approaches for solving linear or *mixed integer linear programming*

(MIP) models is beyond the scope of this textbook. Consequently, references to other textbooks are given.

Section 5.3 introduces the planning tasks of the Frutado company which have to be modeled in the medium-term. An important feature is the level of detail at this planning level – also called the level of aggregation. Furthermore, we will look at the sources of the input data needed for medium-term planning. As regards the output of the model some “decisions” will become directives for subordinated planning modules in the *supply chain planning matrix*.

In Section 5.4 we present the basic Frutado model as an LP. Subsequently, the model is extended to deal with overtime in discrete units as well as lot-sizing. These extensions result in a MIP model.

Results of the optimization run are usually not implemented directly but transferred to subordinate modules as directives (see Section 5.5).

The last Section 5.6 shows the structure of the the learning units for medium-term planning i.e. Supply Network Planning as it is called in SAP APO. We explain what you will be able to observe and include some results of the planning runs of the SNP model.

5.1 Medium-Term Planning Models in the Literature

Medium-term planning is sometimes split into two sub-levels – *Aggregate Planning* and *Master Planning* (these two terms should not be mixed with the term Master Production Schedule (MPS) which is used for the medium-term level of an enterprise resources planning (ERP) system and which usually does not consider capacity constraints). Both sub-levels take an aggregate view of the company and assume that resources are given – i.e. investment decisions have already been made at the strategic or tactical level. However, Aggregate Planning does not only consider the manufacturing part but also incorporates the transportation and distribution network, financial aspects, the work force as well as supply and demand side. Hence, *Aggregate Planning* not only addresses the manufacturing costs but is also concerned with maximizing revenues (or profits) and thus determines which suppliers to select and which products to sell on which markets. Master Planning assumes that the question “which customers to serve” has been answered, and thus demands are assumed to be given. As a result the task of Master Planning is to find a rough purchase plan, production schedule (for all sites of a company) that meets (forecasted) demands at minimum costs.

When looking at models applied in practice, there is no sharp frontier between Aggregate Planning and Master Planning despite the theoretical differences. As an example consider the objective to minimize costs while fulfilling demands. Still, backordering and even lost sales may be allowed if there are bottlenecks and penalty costs taken into account. These penalty costs often relate to the loss of contribution margins and thus conform with maximizing revenues or profits. Hence, in the following we will not

discriminate between the two sub-levels and just use the expression medium-term planning or Master Planning.

In the context of supply chain management (SCM) medium-term planning is concerned with coordinating flows among supplier sites, production sites, warehouses, and customer groups or regions. Hence, the name Supply Network Planning is justified and taken as a name for the corresponding APO module.

Medium-term planning models have been advocated since the early sixties (e.g. Hanssmann and Hess 1960). These models have been extended to multi-level production systems (von Lanzener 1970) and lot-sizing (Billington et al. 1983).

In recent years these general models have been adapted and extended to meet the needs of specific industries, like automotive (Wassermann et al. 2006), electronics (Shirodkar and Kempf 2006), pulp mills (Gunnarsson and Rönnqvist 2008) or food industries (Kanyalkar and Adil 2005) as well as product recovery networks (Walther and Spengler 2005). Often these models now include planning tasks either at the tactical level (e.g. location-allocation decisions) or more detailed decisions (e.g. lot-sizing and scheduling). An overview of models incorporating production *and* transport planning is presented by Mula et al. (2010). A further stream of research is concerned with uncertainty, especially demand uncertainty. This may result in a stochastic programming model (Sodhi and Tang 2009) or robust optimization (Kanyalkar and Adil 2005), which is beyond the scope of today's APS. Instead, APS assume deterministic data. Here, given *safety stocks* and *rolling schedules* are the means to cope with uncertainty.

We will introduce medium-term planning by a small problem statement:

The Fashion-4-You company manufactures fashion clothes comprising summer (SuSu) and winter sport suits (WiSu) as well as all seasons gymnastic suits (GySu). These product types can be split into many different styles, sizes and colors. But with respect to manufacturing the product types represent the capacity requirements very well. Mainly there are two types of capacities to consider, available working hours in the cutting process and the sewing machines. Demand forecasts are available for the next four quarters. There is only one production site. Now the question is, when to produce which product types such that demands are fulfilled, and inventory holding costs are minimized.

The next step is to convert the planning tasks of the Fashion-4-You company described above into an abstract mathematical model formulation (that is independent of actual input data and the scale of the model, e.g. the number of product types or the number of periods to be considered). In SAP APO there is no need for the user to indicate an abstract model formulation because a generic supply chain model is already provided with the software. The task of the user reduces itself to simply activate the necessary features of the model formulation either via the input data or from a list of elements. This general model formulation is based on an abstract description of planning tasks:

A company produces several products in a multistage production process. There are several product types ($j \in J$) manufactured on a few resource groups ($r \in R$) which may become a bottleneck. Available capacities c_{rt} of the resource groups are known for each period t . Semi-finished products cannot be stored over period boundaries. Demand forecasts d_{jt} are available for the next periods $t = 1, \dots, T$ and have to be fulfilled.

A Master Plan is looked for which minimizes inventory holding costs.

Before presenting the corresponding linear programming (LP) model formulation, we will state the most important assumptions underlying our model:

- Demand forecasts for product types are known for every period in the planning interval (e.g. sales forecasts/sales plan).
- Demand has to be fulfilled.
- Production coefficients are constant.
- Production quantities may take real values.
- Setup times and costs are negligible.
- Only end products may be stored (not intermediate products).
- Products can be stored for an unlimited amount of time.

Aggregation

An important issue in medium-term planning is aggregation. Usually, *aggregation* can be achieved along three dimensions:

- There are time buckets instead of continuous time. Time buckets may be of equal length or may increase the more we look into the future.
- Product types are considered instead of all variants of a product. Products are aggregated to product types having “similar” production coefficients and costs. Also, important component types or raw material types may be considered.
- Resource groups are modeled instead of single resources. Resources to be aggregated to resource groups must be “similar” in the sense that these may be capable of producing the same set of product types with “similar” costs.

There are two reasons for aggregating items. One is that this reduces the size of our model (and thus speeds up computations). The other is that detailed data may not be available over the planning interval of medium-term

planning (e.g. demand forecasts are only made for product types). Even if there were detailed forecasts, their forecasting error is reduced if detailed forecasts are aggregated having a correlation of < 1 . The drawback is that aggregation is done at the expense of exactness. I.e. an aggregate solution may be difficult or even impossible to disaggregate into detailed decisions of a subsequent planning level (e.g. module Production Planning and Detailed Scheduling (PP/DS)).

Before presenting the model formulation we have to define dimensions of our data and present the list of symbols. Note, that the dimensions of symbols are abbreviated here as follows:

- Monetary unit = MU
- Period = PE (e.g. [5 working days])
- Quantity unit = QU
- Time unit = TU (e.g. [hour])
- Capacity units = CU
- Volume units = VU

An index (e.g. t) can take certain values which can be described either by a set (e.g. $t \in T$) or by stating the range of values explicitly, whichever is more appropriate ($t = 1, \dots, T$).

Symbols

Indices

j	product type ($j \in J$)
t, s	period ($t = 1, \dots, T$)
r	resource group ($r \in R$)

Variables

X_{jt}	production quantity of product type j in period t [QU]
I_{jt}	inventory level of product type j at the end of period t [QU]

Data

hc_j	inventory holding costs of product type j (per unit and period) [MU/(QU·PE)]
a_{jr}	production coefficient of product type j on resource group r [CU/QU]
d_{jt}	gross demand of product type j in period t [QU]
I_{j0}	inventory level of product type j at the beginning of the planning interval [QU]

I_{jT}	inventory level of product type j at the end of the planning interval [QU]
c_{rt}	available capacity of resource group r in period t [CU]

Objective Function

$$(5.1) \quad \text{Min} \sum_j \sum_{t=1}^{T-1} hc_j \cdot I_{jt}$$

s.t.

Production Capacities

$$(5.2) \quad \sum_j a_{jr} \cdot X_{jt} \leq c_{rt} \quad \forall r, t$$

Inventory Balance

$$(5.3) \quad I_{j,t-1} + X_{jt} = d_{jt} + I_{jt} \quad \forall j, t$$

Non-negativity

$$(5.4) \quad X_{jt} \geq 0 \quad \forall j, t$$

$$(5.5) \quad I_{jt} \geq 0 \quad \forall j, t = 1, \dots, T-1$$

The model formulation starts with the objective function (5.1). Here, the end-of-period inventories are multiplied by the inventory holding cost coefficients and then summed up over all periods and all product types. The only exception is the last period in the planning interval: Here, the end-of-period inventory is a constant (and not a variable). The first constraints represent the (limited) production capacities (5.2). These are defined for every resource group ($r \in R$) in every period within the planning interval ($t = 1, \dots, T$). On the left hand side the production quantities are multiplied by corresponding production coefficients and summed over all product groups. The resulting capacity consumption must not exceed available capacity of a resource group r in period t . The second constraint type is called inventory balance (5.3). It says that a demand for a product group j in period t (see right hand side) can be fulfilled either by initial inventories or by the quantities produced in period t (see left hand side). Quantities not used for satisfying demands reside in the end-of-period inventory. These constraints

link decisions of adjacent periods in our model. The last set of constraints defines the set of feasible values for each variable. In linear programming (LP) these are the real numbers usually constrained further to non-negative values.

Next we present three rules which may be used to verify the syntax of an LP model:

1. Variables may only be multiplied by constants (not by another variable).
2. For every index occurring in the objective function a summation is required. An index occurring in a constraint either results in a summation or there is a constraint for all index values in the corresponding index set.
3. Dimensions, i.e. the units of measurement, must be the same on both sides of a constraint.

As an example check the capacity constraints (5.2). Note that this formal check is no guarantee that the model behaves as expected and solves the “correct” decision problem (which is called the model’s validity).

In order to get an idea of the computational efforts needed to solve the model, its dimensions should be estimated (see also Section 5.2). For an LP model this requires (at least) providing an upper bound on number of variables and constraints. The number of variables can be obtained by looking at the non-negativity constraints and the range of indices these have been defined for. Similarly, the (maximum) number of constraints can be calculated by the corresponding range of indices. (Note, the objective function – as a non-binding single row – is omitted in this estimate. Also the non-negativity constraints are bounds on variables and no explicit constraints.)

For the model above production capacities result in $|R| \cdot T$ constraints while there are $|J| \cdot T$ inventory balance constraints. For $|R| = 2$, $|J| = 2$, and $T = 4$ the model will be solved even on a slow PC with a twinkle in one’s eye.

In order to get acquainted to this type of model formulation we would like to incorporate two more features into the model. The first addition is that there is limited storage space (checked only at the end of a period). The second requirement is that there is an upper limit on the duration of storage in the warehouse for each product type in order to prevent obsolescence. Since our model is a big bucket model the upper limit may (again) only be checked at the end of a period. For simplicity the maximum duration of storage is expressed in multiples of a time bucket (here: quarter).

Before stating the constraints some new data have to be introduced:

Symbols**Data**

$\bar{\Delta}_j$	maximum duration of storage (relating to shelf life) of product type j [PE]
b_j	storage space needed for product type j [VU/QU]
cs	storage capacity [VU]

Consequently, the model has to be extended by two further constraints:

Storage Space Restriction

$$(5.6) \quad \sum_j b_j \cdot I_{jt} \leq cs \quad \forall t$$

Duration of Storage (Upper Bounds)

$$(5.7) \quad I_{jt} \leq \sum_{s=t+1}^{t+\bar{\Delta}_j} d_{js} \quad \forall j, t = 1, \dots, T - \bar{\Delta}_j$$

In addition to the decision (variables) described above one will also find the following aspects in medium-term planning models:

- Purchasing quantities and choice of suppliers
- Regular working hours and overtime needed
- Choice between different alternatives to produce a given product (on different machines with different costs)
- Transportation quantities (between sites and in the distribution network)
- Quantities to sell on different (international) markets

Some of these decision will be incorporated in the medium-term planning model for the Frutado company.

Questions and Exercises

1. Given the data exhibited in [Tables 5.1 - 5.4](#) please write down the Fashion-4-You planning model by stating each constraint explicitly. As initial and ending inventories in the planning interval consider $I_{SuSu,0} = I_{SuSu,4} = 5$, $I_{GySu,0} = I_{GySu,4} = 1$, and none for product type WiSu. A thorough analysis of the planning situation has shown that capacities on the cutting machines will never become a bottleneck in the upcoming year.

Hint: Your model should have 16 explicit constraints plus the objective function.

Product j	Period t			
	1	2	3	4
SuSu	25	25	0	0
WiSu	10	10	70	70
GySu	6	6	6	6

Table 5.1
Demand forecasts d_{jt}
in [1000 QU]

Resource r	Period t			
	1	2	3	4
SM	596	596	596	596

Table 5.2
Available capacities
for sewing machines
(SM) c_{rt} in [CU]

	Product j		
	SuSu	WiSu	GySu
hc_j	1.5	1	1.3

Table 5.3
Inventory holding
costs hc_j in [MU/(PE
* 1000 QU)]

Resource r	Product j		
	SuSu	WiSu	GySu
SM	15	8	12

Table 5.4
Production coefficients
 a_{jr} in [CU/1000 QU]

2. The optimal solution of the above planning model is provided in [Table 5.5](#). The inventory holding costs are 38.3 [MU]. Would you stock the same quantities?

Product j	Period t			
	1	2	3	4
SuSu	6 (26)	5 (24)	5 (0)	5 (0)
WiSu	0 (10)	0 (10)	0 (70)	0 (70)
GySu	0 (5)	7 (13)	4 (3)	1 (3)

Table 5.5
Master plan for
Fashion-4-You (I_{jt}
(X_{jt}))

5.2 Solution Procedures for LP and MIP

To generate a medium-term plan for a supply chain both heuristics and exact methods are available. Heuristics apply some proven rules or rules of thumb which should result in plans with an adequate or even good quality and with moderate computational efforts. In the 1990's very elaborate heuristic principles, called meta-heuristics, have been developed which often provide near optimal solutions. One such meta-heuristic is a genetic algorithm which will be explained in Chapter 6. However, the drawback of heuristics is that we do not know the gap between a heuristic's solution and the optimum. Even worse, if the heuristic does not find a feasible solution we do not know whether a feasible solution exists or not.

Since the solution capability of commercial LP and MIP solvers will be sufficient to generate medium-term plans for most supply chains we will concentrate on these solution techniques here. But, in order to enhance the understanding of the outcome of these models we will start with a solution procedure named "backward scheduling". It may be used, if there is only a single known bottleneck in the supply chain – like in the Fashion-4-You case.

The basic idea of backward scheduling to solve multi-period linear models with the objective to minimize inventory holding costs is that products are sorted (in a preprocessing step) according to monotonously decreasing relative inventory holding costs. Here, "relative" means that we calculate inventory holding costs per unit of the bottleneck resource required by a product.

The procedure starts with calculating net demands (\bar{d}_{jt}) within the planning interval given the initial and final inventories as well as demand forecasts (see Table 5.6). Next, the sequence of products has to be established where $sequence_j$ indicates the position of product j in the sequence (Table 5.7). Now, backward scheduling can start: The starting period will be the latest period (let's call it period T) in the planning interval where capacity requirements exceed available capacity. Starting with the product in the first position ($sequence_j = 1$) we will check how much net demand has to be shifted to the previous period T-1 in order to balance capacities in period T. If already a portion of the net demand suffices we are done with period T, otherwise the product relating to the next position will be shifted, etc. Note that after balancing period T we have to update the temporary inventory level ($\bar{I}_{j,T-1}$) in period T-1. This procedure is continued until we reach period 1. If a shift is needed from period 1 to a period "0" there is no feasible solution. Otherwise, we have reached the cost minimal solution.

This simple procedure is exact provided there is only one bottleneck. In

case there are two bottlenecks simultaneously the calculation of the “relative inventory holding costs” is not defined. A way out is a sequential application of backward scheduling. The above scheme is modified by applying the criterion of “relative inventory holding cost” sequentially, i.e. starting with unloading that resource which has the greatest overload in a period. Once this resource is balanced the next resource is handled in the same way. Although this may result in a feasible solution it may be non-optimal.

In most medium-term planning problems there not only exist several potential bottlenecks but further interrelated decisions, like the utilization of overtime, the outsourcing of excess demands to suppliers, or different routings for producing a product (on alternative resources). To create the best possible plan often will be too much for a human decision maker. Hence, there may a willingness to make use of a sophisticated solution procedure. Here, linear programming (LP) and mixed integer programming (MIP) solvers come into play. If only continuous (real valued) variables are used an LP solver applies while a MIP solver is required if there are one or more binary or integer variables.

The Simplex Method is used to solve LP models while Branch & Bound is applied to solve MIP models. These algorithms are well explained in many basic Operations Research (OR) textbooks. Instead of repeating this common knowledge we refer to the textbooks of Winston (2004) and Hillier and Liebermann (2005). Also, a numerical example is presented in Stadtler (2008).

Subsequently, we would like to introduce some variable types available in many commercial MIP solvers which allow to model specific features in supply chain planning. Finally, we will give some rules of thumb for estimating computational efforts.

In LP we not only have the well-known continuous (real valued) non-negative variables but also “free” variables, which may become negative or non-negative. Similar to the non-negativity constraint ($X \geq 0$) we may also have positive lower bounds on variables ($X \geq lb$) or even upper bounded variables ($X \leq ub$). These simple bounds are taken into account as part of a Simplex iteration and thus do not increase computational efforts.

In the context of MIP models we have

- binary variables ($X \in \{0, 1\}$),
- integer variables ($X \in \{0, 1, 2, 3, \dots\}$),
- partially integer variables, where integer values are required up to a threshold value c , while real values are allowed thereafter ($X \in \{0, 1, 2, 3, \dots, c\}$ and $X \in \mathfrak{R}$ for $X \geq c$), and
- semi-continuous variables defined to take either the values $X = 0$ or $X \geq c$.

As you might expect, partially integer variables are computationally less demanding than pure integer variables (provided c is reasonably small such

that some of these variables take values above the lower bound c). Partially integer variables may be used if one produces and sells indivisible units like a car or a tv set. If we sell these products in small numbers (e.g. 2, 3 or 4) integer numbers are important. However, if the numbers are large (e.g. $X \geq c = 10$) a real valued solution might provide sufficient insight (and may be rounded ex post later on). Semi-continuous variables may be applied to model minimum quantities if produced at all (like in the case of minimum lot sizes). The definition of a semi-continuous variable will save the declaration of an explicit constraint and a binary variable. A drawback is that a semi-continuous variable may not be used in case of fixed setup costs and times. In essence you should remember that

- for LP models computational efforts primarily depend on the number of explicit constraints (i.e. increases to the power of three with the number of explicit constraints) and
- for MIP models the number of binary, integer, and partially integer variables determines computational efforts (i.e. increases exponentially with the number of these variables).

Even LP models with a few hundred thousand constraints are solvable within (CPU-) minutes today. On the other hand for MIP models with only a hundred binary variables an optimal solution may not be proved within days. Hence, the use of binary, integer and partially integer variables should be limited as much as possible. As the mathematical structure of the model formulation is given by the software vendor the user only has a few (but important) levers to limit computational efforts. These four levers are

- to provide an upper limit on the CPU time,
- to relax the integer requirement for some variables (e.g. for some later periods),
- to allow certain types of decomposition such that only a (small) portion of the model contains the integer requirements while other parts of the model are either fixed or relaxed, and
- to provide sufficient hardware.

The first three levers turn the MIP solver into a heuristic. The first lever may result in the Branch & Bound search to terminate prematurely, thus the optimal solution may be missed. The second option may be advantageous if rolling schedules are used knowing that later periods are re-optimized anyway. While a relaxation of integer (production) quantities in later periods might be reasonable, a relaxation of binary variables incurs the risk of destroying the model's logic (especially if fixed costs are involved). For the third lever three options are available in SAP APO (and can be activated by the user by just clicking this option), namely

- time decomposition (together with specifying the number of periods defining the length of the gliding time window),
- product decomposition (together with a number between 1 and 99 indicating the percentage of products that are selected for optimization in one subproblem),
- resource decomposition (together with providing a priority profile).

Which of these levers (or a combination of the four) will provide the best compromise between solution quality and computational efforts has to be tested thoroughly before delivering the tool to the planner for routine use.

Questions and Exercises

1. Given the data exhibited in [Tables 5.1 - 5.4](#) please solve the model by backward scheduling. As initial and ending inventories in the planning interval consider $I_{SuSu,0} = I_{SuSu,4} = 5$, $I_{GySu,0} = I_{GySu,4} = 1$, and none for product type WiSu. While calculations are already made for period 4 (see [Table 5.8](#)) you are asked to complete the exercise for periods 3, 2, and 1.

Product j	Period t			
	1	2	3	4
SuSu	20 (0)	25 (0)	0 (0)	5 (5)
WiSu	10 (0)	10 (0)	70 (0)	70 (0)
GySu	5 (0)	6 (0)	6 (0)	7 (1)

Table 5.6
Net demand $\overline{d_{jt}}$ (I_{jt})
in [1000 QU]

	Product j		
	SuSu	WiSu	GySu
hc_j	1.5	1	1.3
a_{jr}	15	8	12
hc_j/a_{jr}	0.1	0.125	0.108
$sequence_j$	1	3	2

Table 5.7
Sequence of product
based on the relative
inventory holding cost

	Period 4			
	$a_{j,SW} \cdot (\overline{d_{j4}} + \overline{I_{j4}})$	$a_{j,SW} \cdot x_{j4}$	x_{j4}	$\overline{I_{j,3}} = \overline{d_{j4}} + \overline{I_{j4}} - x_{j4}$
SuSu	75	0	0	5
WiSu	560	560	70	0
GySu	84	36	3	4
sum	719			
$c_{SM,t}$	596			
difference	-123			

Table 5.8
Backward scheduling
period 4

2. Why is backward scheduling only a heuristic, if there is more than one potential bottleneck capacity per period?
3. Solve the planning model by using a LP solver.

5.3 Planning Tasks and Data for the Frutado company

5.3.1 Planning Tasks and Level of Detail

Medium-term planning for the Frutado company is a central planning task which coordinates production plants as well as the distribution network. The raw materials needed to produce Frutado’s product portfolio are mainly water, fruit concentrates, and additives which are regarded as commodities and thus are not considered as potential bottlenecks.

It is Frutado’s policy to fulfill demands whenever possible. To prevent stockout situations safety stocks are held for each product at each location. However, if stocks are insufficient then some backordering may be accepted by customers. Demand forecasts are used for each of the 19 products in each distribution center (warehouse) for the next half year. Cost data as well as sales prices are provided by the accounting and marketing department.

The production process consists of two stages, mixing and filling. In each of three production sites there are two production lines for filling. Since the mixing process will never become a bottleneck this stage can be omitted in the medium-term planning model. A table is available indicating which product may be produced (technically) on which filling line. Production on filling lines is done in lots because there are setup costs (e.g. for cleaning the line and pipes before starting a new product) and setup times. Products have to be transported to one of the three distribution centers of the Frutado company immediately after leaving the filling line. Customers are served solely from distribution centers (not from factory). It may turn out that the warehouse space is a limiting factor. To balance inventories it may also

be possible to transport goods between warehouses. Trucks are available whenever needed.

Based on the above description we have to decide which features of the “decision problem” to model explicitly here and those which may be omitted (since these are of minor importance or are incorporated in a subsequent planning level). Note that this abstraction is only justified if trucks will not become a bottleneck. If we are not sure we may envisage to create different versions of the medium-term planning model – called scenarios by some software vendors – and experiment with these versions in order to find out which model is most suitable. For the Frutado company we will have one scenario without any lot size restrictions while there will be a second scenario (see Section 5.6.3) where we model setup decisions explicitly.

Furthermore, we have to define the level of *aggregation*. Since there are only 19 products in the product portfolio we can do without product aggregation. This also applies to the six production lines. However, the (continuous) time axis will be aggregated to periods. Here, a period length of one week seems a good compromise with respect to accuracy of modeling demands (at the end of a week), the forecast accuracy, the production time of a lot (which takes mostly only a small portion of a week), and the model’s dimensions (and expected computational efforts). As a result we will have 26 periods covering the planning interval of half a year.

As regards the objective(s) to pursue at this planning level we are faced with the statement that it is “... Frutado’s policy to fulfill demands whenever possible”. Hence, revenues are fixed, and we can increase profits only by minimizing costs.

5.3.2 Data

Before presenting the symbols – especially data – needed to model the medium-term planning level of the Frutado company some general remarks and classifications seem necessary.

Cost data will be an input to the objective function. One source will be the data provided by accounting. However, we should be cautious which cost elements have been included in these cost coefficients. If we ask for the production cost per unit for a particular item this may include cost of input materials, cost of personnel, cost of resource usage, and some overhead costs. All these cost elements may be well justified – but may be inappropriate to find the best solution for our planning task. For instance, the cost for input material may be based on historic purchasing costs, while for planning purposes the cost for its replenishment in the future is recommended. The cost of personnel should not be a weighted sum of costs for regular and overtime as observed in the past. As regards overheads these should not be included in the model at all, except for those fixed costs which can be attributed to decisions to close or open a complete department or factory including its administrative personnel.

The leading idea for generating the cost figures for our model is that these costs can be attributed to explicit decision variables.

A second category of cost figures which are not provided by accounting are the so called “steering costs”. These are included in our model to prevent something “undesirable” from happening like penalty costs for backordering, for falling below prescribed safety stock levels etc. These costs are difficult to set and often incur some subjective elements. What is the cost of one unit supplied late to a customer? Note that at this stage we do not know which customer will face the late delivery nor do we know the consequences (will there be less loyalty in the future?).

Both costs – *accounting* and *steering costs* – will be put together in one objective function. Hence, the dimensions for the steering costs should be chosen carefully in order not to “distort” the “optimal” solution.

Further data is retrieved either from an ERP system or a data warehouse. E.g. demand forecasts are either retrieved from a data warehouse or directly from the Demand Planning module. Besides the demand forecasts also already accepted (customer) orders become an input to Master Planning. Last but not least demand data have to be aggregated into demands per period and per product (type). Available capacities, BOM coefficients, and production coefficients may be transferred from an ERP system automatically. This assumes that the data stored in an ERP system are “correct”, i.e. are kept up to date and in a precision needed for Master Planning. Note, this is not always the case. Hence, some manual inspections and corrections may be necessary.

A very useful representation of data is the *Production Process Model (PPM)*, which combines the data of a BOM with the routing of a product including production coefficients (see Vollman et al. 1997, p.804). Furthermore, timing between activities is possible (e.g. a maximum time lag between mixing of liquids and filling). Since, we have concentrated on one production stage (filling) and only have a single activity (for filling) the PPM for each *location-product* combination (or *location-product* for short) is very simple – it only contains the production coefficient a_{jr} .

Special care is also needed when specifying the *utilization rate* of production lines (e.g. us_r). The “correct” utilization rates are only known, once the detailed production schedule has been established and usually vary between periods. However, since this plan is unknown when preparing the Master Plan these coefficients have to be anticipated. Since setup decisions will not be modeled explicitly in our basic case, utilization rates have to be “estimated” according to the expected loss due to setup times.

Now, we will provide a complete list of the data used in the Frutado model:

Symbols

Indices and Index Sets

δ	interval where a specific penalty cost applies ($\delta = 1.. \bar{\delta}_{jl}$)
Δ	period (subindex for duration of storage restrictions)
j	product ($j \in J$)
$l(s, e)$	location ($l \in L$), (with s being the start and e being the end location of a transportation lane)
r	resource ($r \in R$)
t	period ($t = 1, \dots, T$)
$Dest(l)$	index set of all successor locations (destinations) e connected via a direct transportation lane from location l
$LocProd(l)$	index set of all products j that may be available in location l
$ResLoc(l)$	index set of all resources r that belong to location l
$ResProd(r)$	index set of all products that may be produced on resource r
$Sour(l)$	index set of all predecessor locations (sources) e connected via a direct transportation lane to location l

Data

a_{jr}	capacity consumption for one unit of product j on resource r (“production coefficient”) [CU/QU]
c_{rt}	available gross capacity on production line r in period t [CU]
d_{jlt}	gross demand of product j at location l in period t [QU]
$\bar{\delta}_{jl}$	maximum number of periods customers will wait for a product j backordered in location l
$\bar{\Delta}_j$	maximum duration of storage, i.e. number of periods a product j is allowed to be stocked
I_{jl0}	initial inventory of product j [QU] at location l
o_{rt}	available gross capacity during overtime on production line r in period t [CU]
ss_{jl}	target safety stock for product j at location l [QU]
us_r	utilization rate of production line r during regular time
uo_r	utilization rate of production line r during overtime
hc_{jl}	holding cost for product j per unit and period at location l [MU/(QU*PE)]
oc_r	overtime cost per hour on resource r [MU/TU] ($t \geq 1$)

pc_{jl}^{ss}	penalty cost for violating the safety stock level for product j in location l per quantity unit [MU/QU]
pc_{jl}^1	penalty cost for backordering one unit of product j in location l [MU/QU]
pc_{jl}^2	penalty cost for one unit of lost sales of product j in location l [MU/QU]
tc_{se}	cost of transportation on the lane from s to e per transport unit [MU/transport unit]
rc_{jr}	cost per unit of product j if produced on resource r [MU/QU]

To complete the list of symbols we add variables here, although variables will be discussed in the next section.

Variables, real [QU]

$B_{jlt,t+\delta}$	backordered demand of product j in location l in period t finally fulfilled in period $t+\delta$
D_{jlt}	demand fulfilled for product j in location l in period t
D_{jlt}^-	demand not fulfilled for product j in location l in period t
I_{jlt}	end of period inventory for product j in location l in period t
SS_{jlt}^-	safety stock violation for product j in location l in period t
XD_{jlt}	amount destroyed of product j in location l in period t due to violation of maximum time in warehouse
XO_{jrt}	production quantity of product j on resource r in period t produced using overtime capacity
XS_{jrt}	production quantity of product j on resource r in period t produced using regular capacity
XT_{sejt}	transportation quantity of product j from location s to e in departure period t ($t \geq 1$)

5.4 Modeling the Frutado Planning Tasks

5.4.1 Introductory Remarks

The model of Frutado's production and distribution network is an extension of the standard Master Planning model as described in Section 5.1. It not only includes decisions on the production quantities of products j on filling lines r in a period t in regular time (XS_{jrt}) as well as overtime (XO_{jrt}) but also decisions on the transport of goods (XT_{sejt}) from a start location s (warehouse or production site) to a destination location d being one of the

three warehouses of the Frutado company which are placed next to each production site.

Special attention has to be paid to demand fulfillment: Although the Frutado company intends to satisfy all customer demands, this may turn out to be impossible in case demands have been underestimated and available inventories at the location of demand (warehouse l) are insufficient. In order to limit the number of stockout events safety stocks are put in place (ss_{jl}). The safety stocks may be used in the model (variable SS_{jlt}^-) at a certain penalty cost pc_{jl}^{ss} per unit. If there is a stockout this may not result in lost sales immediately, because customers are assumed to be prepared to wait for a certain period of time (at most $\bar{\delta}_{jl}$) depending on the type of product j and the location l . This is the backorder case (variables $B_{jlt,t+\delta}$). As backorders are a poor customer service this is penalized by a (steering) cost of pc_{jl}^1 per unit and time unit. Only if the upper limit is exceeded backorders will become *lost sales* (variables D_{jlt}^-).

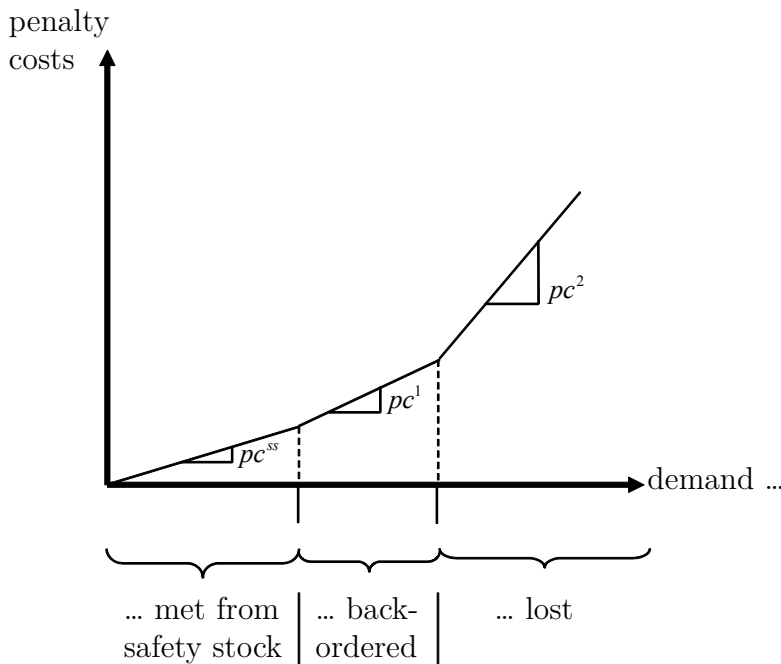


Figure 5.1
 Piecewise linear
 penalty costs

In the consumer goods industry the case of overestimating demands may cause problems too, because of the *shelf life* of goods. Let us assume a maximum shelf life of 6 months for juice (from the time of production to the time of sale) then only a small portion of this period may be used within the Frutado network for transport and storage (e.g. $\bar{\Delta}_j=3$ weeks).

Given this information we may start describing the basic Frutado model.

5.4.2 Basic Frutado Model

$$\begin{aligned}
(5.8) \quad & \text{Min} \sum_{r \in R} \sum_{j \in \text{ResProd}(r)} \sum_{t \in T} [rc_{jr} \cdot (XS_{jrt} + XO_{jrt})] \\
& + \sum_{r \in R} \sum_{j \in \text{ResProd}(r)} \sum_{t \in T} [oc_r \cdot a_{jr} \cdot XO_{jrt}] \\
& + \sum_{l \in L} \sum_{j \in \text{LocProd}(l)} \sum_{t \in T} [hc_{jl} \cdot I_{jlt}] \\
& + \sum_{s \in L} \sum_{e \in \text{Dest}(s)} \sum_{j \in J} \sum_{t \in T} [tc_{se} \cdot XT_{sejt}] \\
& + \sum_{l \in L} \sum_{j \in \text{LocProd}(l)} \sum_{t \in T} [pc_{jl}^{ss} \cdot SS_{jlt}^-] \\
& + \sum_{l \in L} \sum_{j \in \text{LocProd}(l)} \sum_{t \in T} \left[\sum_{\substack{\delta=1, \\ t+\delta \leq T}}^{\bar{\delta}_{jl}} (\delta \cdot pc_{jl}^1) \cdot B_{jlt,t+\delta} \right] \\
& + \sum_{l \in L} \sum_{j \in \text{LocProd}(l)} \sum_{t \in T} [pc_{jl}^2 \cdot D_{jlt}^-]
\end{aligned}$$

s.t.

Inventory Balance

$$\begin{aligned}
(5.9) \quad & I_{jl,t-1} + \sum_{s \in \text{Sour}(l)} XT_{sljt} + \sum_{r \in \text{ResLoc}(l)} [XS_{jrt} + XO_{jrt}] = \\
& \sum_{e \in \text{Dest}(l)} XT_{lejt} + D_{jlt} + \sum_{\delta=1}^{\bar{\delta}_{jl}} B_{jl,t-\delta,t} + XD_{jlt} + I_{jlt}
\end{aligned} \quad \begin{array}{l} \forall j \in J, \\ l \in L, \\ t \in T \end{array}$$

Demand Fulfillment

$$(5.10) \quad d_{jlt} = D_{jlt} + \sum_{\substack{\delta=1, \\ t+\delta \leq T}}^{\bar{\delta}_{jl}} B_{jlt,t+\delta} + D_{jlt}^- \quad \begin{array}{l} \forall j \in J, \\ l \in L, \\ t \in T \end{array}$$

Production Capacities

$$(5.11a) \quad \sum_{j \in \text{ResProd}(r)} [a_{jr} \cdot XS_{jrt}] \leq c_{rt} \cdot us_r \quad \begin{array}{l} \forall r \in R, \\ t \in T \end{array}$$

$$(5.11b) \quad \sum_{j \in \text{ResProd}(r)} [a_{jr} \cdot XO_{jrt}] \leq o_{rt} \cdot uo_r \quad \begin{array}{l} \forall r \in R, \\ t \in T \end{array}$$

Safety Stock

$$I_{jlt} \geq ss_{jl} - SS_{jlt}^- \quad \begin{array}{l} \forall j \in J, \\ l \in L, \\ t \in T \end{array} \quad (5.12)$$

Duration of Storage

$$I_{jlt} \leq \sum_{r \in ResLoc(l)} \sum_{\Delta=0}^{\bar{\Delta}_j-1} XS_{jr,t-\Delta} + \sum_{r \in ResLoc(l)} \sum_{\Delta=0}^{\bar{\Delta}_j-1} XO_{jr,t-\Delta} + \sum_{s \in Sour(l)} \sum_{\Delta=0}^{\bar{\Delta}_j-1} XT_{slj,t-\Delta} \quad \begin{array}{l} \forall j \in J, \\ l \in L, \\ t \in T \end{array} \quad (5.13)$$

Non-negativity

$$B_{jlt,t+\delta} \geq 0 \quad \forall j \in J, l \in L, t \in T, \delta = 1.. \bar{\delta}_{jl} \quad (5.14)$$

$$D_{jlt} \geq 0 \quad \forall j \in J, l \in L, t \in T \quad (5.15)$$

$$D_{jlt}^- \geq 0 \quad \forall j \in J, l \in L, t \in T \quad (5.16)$$

$$I_{jlt} \geq 0 \quad \forall j \in J, l \in L, t \in T \quad (5.17)$$

$$SS_{jlt}^- \geq 0 \quad \forall j \in J, l \in L, t \in T \quad (5.18)$$

$$XD_{jlt} \geq 0 \quad \forall j \in J, l \in L, t \in T \quad (5.19)$$

$$XO_{jrt} \geq 0 \quad \forall j \in J, r \in R, t \in T \quad (5.20)$$

$$XS_{jrt} \geq 0 \quad \forall j \in J, r \in R, t \in T \quad (5.21)$$

$$XT_{sejt} \geq 0 \quad \forall s \in L, e \in L, j \in J, t \in T \quad (5.22)$$

The model's logic will now be explained in detail starting with the objective function (5.8). It aims at minimizing costs within the planning interval. It is composed of direct variable costs as well as steering costs. The first term calculates the sum of production costs over all products j produced on resources r in the Frutado network throughout the planning interval both in regular time as well as overtime. While in the first term direct production costs per unit are used as a basis (like material costs) which do not differ between regular time and overtime the second term allows to add the extra labor cost to be paid per time unit of overtime. The inventory holding costs at all locations are accumulated in the third term. Transportation costs between warehouse locations are considered in term four. The next three

terms account for penalty costs due to violating safety stock levels (fifth term), for making use of backorders (depending on the duration δ a customer order is delayed), and finally for lost sales (seventh term).

These costs have to be related to the constraint set which starts with the inventory balance constraints (5.9). Like in the standard model for medium-term planning (see Section 5.1) the left hand side indicates what is available in period t for product j at location l while the right hand side shows what is used or left over. Hence, the left hand side is composed of the initial inventory, the amount that is received by transportation from other locations s as well as the production of product j during period t at location l both in regular time and overtime. These quantities may be either transported to other locations e or used to fulfill the demand of period t (second term on the right hand side). Furthermore, one can fulfill backordered demands or destroy inventories if the maximum duration of storage will be exceeded. The remaining quantity is put in the end-of-period inventory.

One might ask whether such an unfavorable situation as destroying products due to the violation of the duration of storage restriction might occur? – First the product is produced at a certain cost, then stored in inventory (at a certain cost) and finally will be disposed! Does the model not keep track of the goods and prevent this unfavorable situation from happening? The answer is, yes, the model tries to prevent this situations from happening, but there may be two reasons for this unfavorable event to occur: One is that we may allow the user to specify a minimum lot size. Now, assume there is a positive net demand for a product j in a period t which requires to produce a minimum lot size. If cumulated demands in the interval a product may stay in the Frutado network $[t, t + \bar{\delta}]$ is less than the minimum lot size then the violation of the duration of storage restriction is inevitable. The other reason refers to rolling schedules: If from one planning run to the next, demands turn out to be lower than previously estimated then products which were expected to be sold in the last period of their allowed duration of stay in the Frutado network are unsold and have to be destroyed.

The demand fulfillment constraints (5.10) relate to the inventory balance constraints. These constraints show how customer demands d_{jlt} are split into demands either being fulfilled in time or backordered until periods $t + 1$ to $t + \bar{\delta}$ or finally disposed of. Even if there is no way to fulfill demands completely due to some bottlenecks there will always be a feasible solution to the model – with penalty costs activated in the objective function. Hence, the inventory balance constraints (5.9) are termed *soft constraints* as opposed to *hard constraints* which may result in an infeasible solution.

Filling lines have a limited capacity per period within regular time c_{rt} (5.11a) and overtime o_{rt} (5.11b). Since there will be a loss of capacity due to setup times (which are not modeled explicitly here), we will reserve a certain portion of the capacity for setup activities resulting in “productive” utilization rates for these resources (u_r, u_o). Only after scheduling has

been performed (see Chap. 6) it becomes clear whether this estimate has been correct.

As already said, Master Planning of the Frutado company is achieved by a deterministic model although one is aware that at least customer demands show a certain degree of uncertainty. To be able to even satisfy demands exceeding the estimate to a certain extent the end-of-period inventory should not be depleted completely (in the deterministic model) but stay above the safety stock level to be calculated for each location-product separately (see constraints (5.12)). Its ex-ante calculation is beyond the scope of this textbook and the reader is referred to Tempelmeier (2005) or Waters (2003). However, in case a feasible solution of the Master Planning model can only be obtained by violating the safety stock level this will be allowed and penalized in the objective function.

The last explicit constraints (5.13) limit the *duration of storage* Δ_j for products within the Frutado network. This time limit is controlled by the end-of-period inventory. Due to the fact that we do not know when a product is produced within a period and when it is sold to a customer (either at the start or end of a period) the calculated maximum duration of storage Δ_j has to be further reduced (due the model's inaccuracy). To be on the safe side one should deduct 2 periods while on average the inaccuracy is 1 period. The latter will be assumed here (see the sums over Δ on the right hand side of (5.13)). To be more precise, the end of period inventory of a product j at location l in period t is limited by the quantities either produced in that location during the last $\bar{\Delta}_j - 1$ periods or transported to location l .

Although this logic seems convincing it still contains a flaw! In order to prevent a violation of the duration of storage restriction the model will send the corresponding product from one location to another where it arrives in the status "new". This behavior is advantageous for the model if the cost of transportation is less than that of disposal (which should hold in most cases). In order to prevent this behavior a reformulation of the model is necessary. One way is to allow the transportation of goods only in the period a customer demand is actually fulfilled. Another, more elaborate approach, is to define a complete path for a product from the location of production via an intermediate warehouse to the warehouse the product is finally sold (by adding a further location index l to the transportation quantity variable $XT_{slej t}$).

Finally, we have the non-negativity constraints for the variables (5.14)-(5.22).

A last remark concerns a phenomenon often observed in optimal solutions of mathematical programming models, namely *extreme solutions*. For the above model the logic runs as follows: The "cheapest" product (holding cost per capacity unit) is the first choice for building up seasonal inventory. Only if duration of storage restrictions are at its limits the second cheapest product is stored. This is known as extreme points or extreme solutions. This incurs no problem in case the setting is really deterministic. However, in real life

situations demands are to some extent uncertain. Hence, letting one product hit the duration of storage restriction is risky, because this will result in waste if demands turn out to actually be lower than expected. Hence, in order to better balance the risk one could add a further constraint over several products such that the runout time of the inventory of these products is “similar” (e.g. the difference in runout times is at most one period).

Questions and Exercises

1. Reformulate the above model in such a way that transportation between a warehouse s to the final destination warehouse e is only allowed if it is needed to fulfill customer demands immediately after arrival of the consignment (i.e. in the same period t).
2. Create new constraints for balancing the risk of obsolescence such that the runout time of products based on the end-of-period inventory differs by at most one period. You may use the mean demand \bar{d}_{jlt} at location l during the maximum duration of storage $\bar{\Delta}_j$ in the Frutado network for calculating the actual runout times RT_{jlt} of a product j at location l at the end of period t .
3. Assume we have introduced a semi-continuous variable

$$X_{jrt} \geq LS_{jr}^{min} \text{ or } X_{jrt} = 0 \quad \forall j \in J, r \in R, t \in T$$

for modeling minimum lot sizes. How can this variable be linked to the above model?

4. Have a look at the results of the optimization run of the SNP Planner: Which variables and its optimal values are documented in the “results log” learning unit?

5.4.3 Extensions

A great advantage of mathematical programming models is their flexibility to incorporate a great number of features the user may regard relevant for Master Planning. However, to be too precise may not be useful because Master Planning has to take a broad view of the supply chain and has to consider only the most important decisions necessary in the medium-term (e.g. with a planning horizon ranging from six months to three years depending on the type of industry). For instance the exact sequencing of production (lots) is usually not done in Master Planning but in Production Scheduling (see Chap. 6). In this subsection we will outline two potential extensions. The first is that overtime is not only limited to 24 hours per period, but has to be taken either completely or not at all. The second extension considers lot-sizing decisions and related setup times explicitly.

The overtime option has already been considered in the basic model (see constraints (5.11b)). There, the constant o_{rt} represents the maximum overtime per week for a resource r . This could be related to a calendar with the additional condition that in case of a holiday on Friday overtime on Saturday is prohibited. The optimal solution of the basic model may propose overtime in the range $[0,24]$ hours at the end of a week (e.g. 0.45 hours on a resource r at the end of period t). This may be impossible to implement because no employee will work on Saturdays for this small amount of time. Let us assume there is an agreement between management and workers (or factory committee) that either a resource is open for 24 hours on a Saturday or not at all.

The alteration of the model is rather easy. First, we have to define additional binary variables:

O_{rt} 1, if overtime capacity is used on resource r in period t ,
0 otherwise

This new variable definition is added to the non-negativity constraints. Also, an additional term is included in the objective function consisting of the sum of the above binary variables multiplied by the additional costs for a 24 hour capacity extension on Saturdays. Constraints (5.11b) have to be modified, too:

Overtime Production Capacities, Discrete

$$\sum_{j \in ResProd(r)} a_{jr} \cdot X_{jrt} \leq u_{or} \cdot O_{rt} \quad \forall r \in R, \quad t \in T \quad (5.23)$$

As regards the explicit consideration of the lot-sizing decisions into the model we will start with some observations. When taking a closer look on the setup matrix we will realize that setup times and costs are sequence dependent. However, our Master Planning model is a big bucket model, i.e. many products can be produced on a resource within a period. Hence, the sequence of products is not known at this planning level. As a compromise the *mean* setup time and cost for each product on a resource r is an input to the model. Obviously, the “true” loss due to setup times and the true setup costs will be determined at the scheduling level. Only then the extent of the “error” made by the “compromise” will be known.

As a first step we must remove the expected loss of a setup from the utilization rates us_r and uo_r which will be renamed us_r^- and uo_r^- . Next, we will have to define new binary variables:

Y_{jrt} 1, if product j is setup on resource r in period t ,
0 otherwise

This new variable definition has to be added to the non-negativity constraints, too. Furthermore, we will have to add the setup costs to the objective function of the basic model:

... Setup Costs (added to the objective function)

$$(5.24) \quad Min \dots + \sum_{r \in R} \sum_{j \in ResProd(r)} \sum_{t \in T} sc_{jr} \cdot Y_{jrt}$$

In order to activate the setup costs in the objective function an additional constraint is needed linking the amount produced of a certain product j to its binary setup variables:

Explicit Setups

$$(5.25) \quad a_{jr} \cdot XS_{jrt} + a_{jr} \cdot XO_{jrt} \leq [us_r^- \cdot c_{rt} + uo_r^- \cdot o_{rt}] \cdot Y_{jrt} \quad \begin{array}{l} \forall j \in J, \\ r \in R, \\ t \in T \end{array}$$

The logic is as follows: If one produces product j on resource r in period t then the left hand side of inequalities (5.25) becomes positive and thus forces the binary variable Y_{jrt} to “1”. In order not to limit the production amount to “1” the binary variable is multiplied by a sufficiently large number – which is the resources capacity in period t , here. Note, that we have at most one setup per period, product and resource. Thus, a lot may start in regular time and end in overtime (without an additional setup during overtime) which is obvious from the point of view of practice but does not usually apply in big bucket models.

The last alteration to the basic model concerns the capacity constraints: Production Capacities with Explicit Setups

$$(5.26) \quad \begin{array}{l} \sum_{j \in ResProd(r)} a_{jr} \cdot XS_{jrt} + a_{jr} \cdot XO_{jrt} \\ + \sum_{j \in ResProd(r)} st_{jr} \cdot Y_{jrt} \end{array} \leq us_r^- \cdot c_{rt} + uo_r^- \cdot o_{rt} \quad \begin{array}{l} \forall r \in R, \\ t \in T \end{array}$$

Here, the mean setup times st_{jr} have been included on the left hand side of the inequalities explicitly. We would like to add that the new production capacity restrictions (5.26) will not make inequalities (5.23) redundant, because the new restrictions do not relate overtime production XO_{jrt} with the overtime variables O_{rt} .

A last remark concerns our mathematical programming model of the Frutado supply chain presented above. Actually, our model will probably not be exactly the same as the model formulation created by the SNP Optimizer of SAP APO. But this has not been our intention here. Instead, we have shown how Frutado’s Master Planning model might look like. We have discussed its underlying logic and outlined some pitfalls which might arise before coming up with a “correct” model formulation.

The learning units relating to the above model extensions are called “in-depth streams” and are not needed for subsequent modules and learning units.

Questions and Exercises

1. How would the constraints and variables look like if the overtime option cannot be executed for each resource individually but for the whole location at the end of a period t ?
2. Which solution of the model do you expect if some setup times st_{jr} are larger than the length (i.e. capacity) of a period?
3. Please rank the three model variants “basic model”, “extended model with a discrete overtime option per resource”, and “extended model with a discrete overtime option per location” according to increasing expected computational efforts. What is the reason?

5.5 Implementation and Disaggregation of Results

Once the optimization run has been completed and accepted by the planner(s) – may be after some manual changes – planning results have to be released. Since medium-term planning is concerned with “planning” and not with “execution” the main task here is to provide *directives* for the subordinate planning levels. However, some medium-term decisions now can be realized and negotiated, e.g. negotiations with (representatives of) employees regarding planned overtime in the weeks to come or entering into a purchasing agreement with suppliers based on secondary demands calculated from planned medium-term production quantities.

As subsequent planning levels we have PP/DS, Deployment as well as TP/VS. For the Transport Load Builder (TLB) the transport quantities (XT_{sejt}) are handed over. In the Frutado case disaggregation is neither necessary for products nor for resources because these have not been aggregated.

However, for the PP/DS module we will have to disaggregate the planning results (of the basic Frutado model) due to aggregation of time (see the “periods” in Master Planning). A related reason for disaggregating Master Plans is that PP/DS requires *planned orders* as directives. In a first step, these planned orders will be placed in an infinite capacity schedule over the PP/DS planning interval (4 weeks here). Based on this – probably – infeasible schedule a feasible sequence of planned orders is generated by the PP/DS module given some short-term objectives, like minimizing the sum of setup times (see Chap. 6).

Hence, before transferring results, a disaggregation routine has to be applied that converts production quantities of products per period and production line (the sum of XS_{jrt} plus XO_{jrt} , here) into planned orders (i.e. lot sizes). If the lot-sizing policy is “lot-for-lot” then production quantities may be taken as given by the output of the LP or MIP solver. As an alternative one may choose a simple lot-sizing rule – like “fixed lot sizes” – or even design an individual algorithm. In the case of “fixed lot sizes” a disaggregation routine will convert production quantities into fixed lot sizes by a forward or backward scheduling routine.

As an example consider a product with a fixed lot size of 100 [QU] and planned production quantities of 40 [QU], 180 [QU], and 70 [QU] in periods $t=1, 2,$ and 3 . In the first period we have to start with the first lot (100 [QU]) resulting in an end-of-period inventory of 60 [QU]. The net demand in period 2 now is 120 [QU] requiring two lots with an end-of-period inventory of 80 [QU]. Consequently, no production is necessary in period 3.

Note that this has been an illustrative numerical example which also shows that in case lot size production deviates significantly from the medium-term plan it will probably be difficult to create a feasible production schedule. Hence, if lot sizes deviate that much from production quantities obtained by an LP model this would be a strong argument for modeling lot sizes *explicitly* already in the medium-term.

5.6 SNP Learning Units

5.6.1 Overview

The SNP learning units are split into five broad themes:

1. SNP master data
2. Model building
3. SNP planning run
4. Transfer of planning results
5. In-depth stream

These learning units represent a “natural” sequence of tasks from creating a model to solving it and implementing its results. There are a number of sub-learning units which provide the necessary details (see [Fig. 5.2](#)).

These learning units contain the most relevant tasks to perform. The way SNP may be used in practice is described in [Figure 5.3](#): Setting up a new model can be done in the Supply Chain Engineer of SAP APO (see the rectangle on the left) and requires to create the master data, a model, and a planning book. A planning book is composed of multiple data views. These data views show the content (e.g. key figures, characteristics, the planning horizon, bucket size) based on a certain layout for interactive planning. The customization of planning books enables the use of just one planning area for different planning tasks (see 3.3.4). For example the DP and the SNP module both may have planning books based on the same planning area.

“Building a supply chain model at the macro level” means that locations for production sites, warehouses, and customers are defined as *master data* elements, placed on a map and assigned to a model. Also, the actual distance for every transportation lane has to be provided (e.g. either manually by the user, automatically by calculating the straight-line distance based on

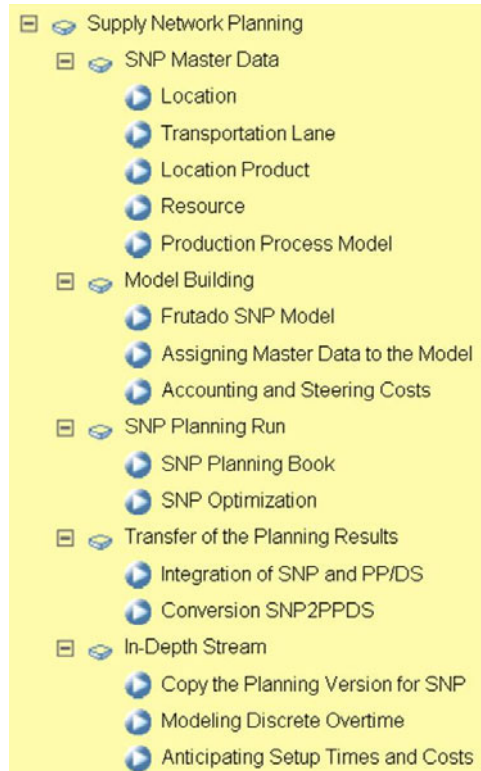


Figure 5.2
SNP outline
© Copyright 2011. SAP
AG. All rights reserved

geo-coordinates or by a *GIS system* taking into account real road distances). These two tasks are made once at the beginning. All other tasks are repeated routinely in the course of rolling schedules. At the macro level the following types of master data have been created for the Frutado company (see [Table 5.11](#)).

Master Data	Number
locations	6
products	19
location-products	86
transportation lanes	15
production process models	32
supply chain model	1
planning version	2
hierarchy	0

Table 5.11
Required master data

The routine tasks start with modeling the SNP model in detail, indicating procurement, production, storage, distribution, and sales activities. The activities may be modeled in different ways and combined such that a model results that best suits the real world Master Planning problem to be solved. [Table 5.12](#) provides an overview of these activities and the way it has been

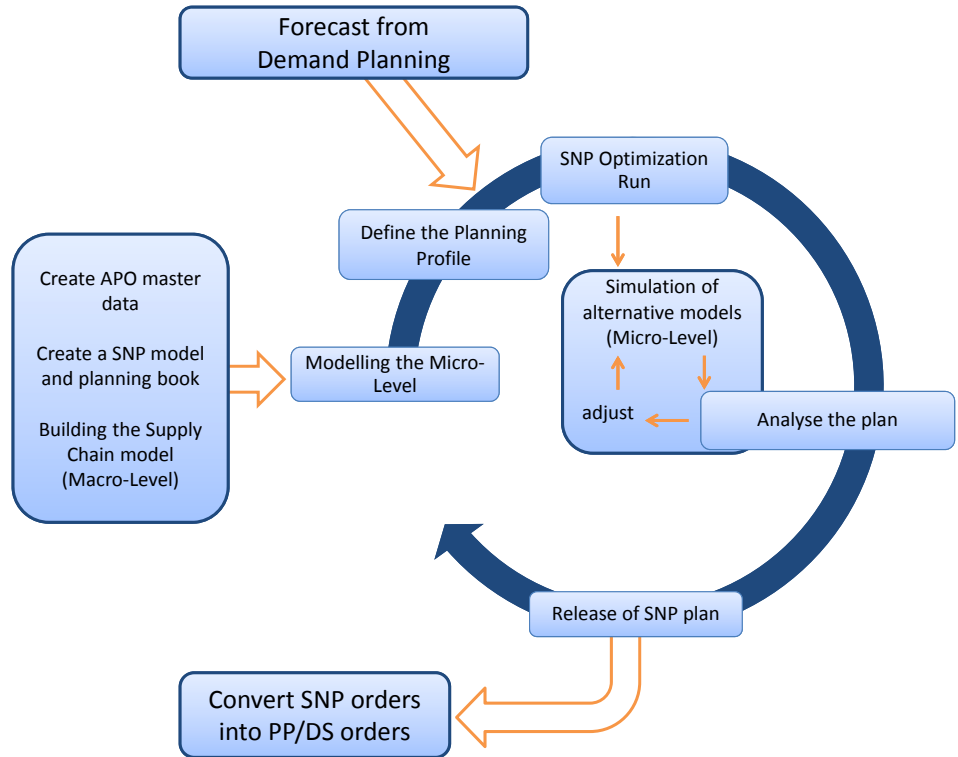


Figure 5.3
SNP planning cycle

specified in the SNP Frutado model. We will distinguish between the basic model (indicated by a “B”) and the extended model (“E”).

The planning profile defines the parameters of a planning scenario. Then demand forecasts are loaded to the model created in the Demand Planning module. Now, optimization can start. Planning results have to be analyzed. This may give rise to explore a new set of data and a new optimization run – also called *deterministic simulation*. Once an acceptable Master Plan has been created it will be released, i.e. converted into directives for the subordinate planning units. This finishes one planning cycle. It will start again when the next (rolling) schedule is needed. The last learning unit corresponds to the in-depth stream.

Now, we will briefly describe what can be observed when working through the different learning units.

5.6.2 Basic Stream

SNP Master Data

While locations and transportation lanes have already been specified at the macro level we now have to provide the remaining features of the Frutado SNP model. [Table 5.12](#) provides an overview of all features at the micro-level that have been input to SNP. Note, there is a distinction between the “Basic” Frutado model and the “Enhancements” (relating to Section 5.4.3).

Data can be input to the SNP model in different ways. There are mainly two sources, an ERP System and a data warehouse. Also, some data will be kept in the APO itself (e.g. directives from other modules). Sources that might cause security problems are flat files or spreadsheets. In SAP APO these sources of data require the use of additional interface programs – called *BAPIs*.

While at the macro level the types and number of master data have been generated (see [Table 5.11](#)) these objects now have to be filled with data. Also, these master data may be updated interactively here, if necessary.

As an example we refer to the learning unit “location-products”. Here we present a screenshot where penalty costs defined globally for a specific product – named FRU_SAFT_01 – may be written over by specifications valid for a specific location ([Fig. 5.4](#)). In this screenshot “delay” means “backordered”.

The screenshot displays the SAP APO interface for changing product data. The title bar reads "Change Product FRU_SAFT_01 for Location FRU_DC_01". The main area is divided into two sections: "Penalty Costs" and "Location-Dependent Penalty Costs". Each section contains three columns of input fields for different demand types: "For Customer Demand", "For Demand Forecast", and "For Corr. Demand Fcst". The input fields are organized as follows:

Section	Category	For Customer Demand	For Demand Forecast	For Corr. Demand Fcst
Penalty Costs	No Del. Penalty	400.000,000	400.000,000	400.000,000
	Delay Penalty	50.000,000	50.000,000	50.000,000
	Maximum Delay	7	7	7
Location-Dependent Penalty Costs	No Del. Penalty	400.000,000	400.000,000	400.000,000
	Delay Penalty	50.000,000	50.000,000	50.000,000
	Maximum Delay	7	7	7

Figure 5.4
Change product
FRU_SAFT_01 for
location FRU_DC_01
© Copyright 2011. SAP AG.
All rights reserved

A further example for master data is a production process model (PPM) which combines the information available in the bill of materials and in the routing of a location product. Among other things, a PPM contains the

production coefficients, i.e. the resource consumption for each unit of an operation (a more detailed description of PPMs is provided in Section 6.4.2).

Model Building, Model Solving and Results

Model Building

A supply chain model now will be created by linking it to the respective master data. This is either achieved by using the “Supply Chain Engineer” or directly at the creation of each master data element (which is mainly used in practice). The master data already describes the structure of the Frutado model. Which elements are available for the Frutado model in the basic and extended version has been documented in [Table 5.12](#).

Frutado Process	Parameter	Characteristics	APO-Element	Model*
Procurement	not modeled			
Production	production costs	linear	PPM	B
		piecewise linear (setup costs)	PPM + Optimizer profile	E
	production quantities	continuous	PPM	B
		minimum lot size	PPM + Optimizer profile	E
	available capacity	regular	Resource: Capacity Variant 1 & Quantities/Rates Definition + Optimizer profile	B
		overtime with linear costs per extra unit	Resource: Capacity Variant 2 & Quantities/Rates Definition + Optimizer profile	B
		discrete enhanced capacity	Resource: Capacity Variant 2 & Quantities/Rates Definition + Optimizer profile	E
	capacity consumption	linear	PPM	B
fixed + linear (setup time)		PPM + Optimizer profile	E	
Storage	storage costs	linear	Location-Product	B
	safety stock	fixed	Location-Product: method SB	B
	safety stock penalty	linear	Location-Product	B
	maximum duration of storage	fixed	Product	B

Continued on Next Page...

Frutado Process	Parameter	Characteristics	APO-Element	Model*
Distribution	cost of transportation	linear	Transportation Lane	B
		piecewise linear	Transportation Lane + Optimizer profile	E
Sales	delay	maximum delay: 7 days	Product	B
		linear penalties	Product	B
	non delivery	linear penalties	Product	B
*B = Basic model; E = Enhancement				

Table 5.12
Model features for the micro-level

Also, you can specify which constraints of the Frutado model should be hard or soft constraints. Assigning the master data to the model is required for each data type (plants, distribution centers, location-products ...). As this task is very similar for the different data types some of these lessons may be skipped.

At last the cost data – accounting and steering costs – have to be assigned. Here, also cost functions can be defined (like the one in [Fig. 5.1](#)).

SNP Planning Run

Before starting the *SNP optimizer* a *planning book* has to be created. Subsequently, all data for the basic Frutado model are loaded into the planning book. As an example the learning unit will show how to specify the data for the 3 plants, 3 distribution centers, and 86 location-products. The planning book allows to view these data in tables over the SNP planning interval (26 weeks). Once a solution is generated by SNP optimizer the planning book will also show key figures and characteristics of the solution. The planning book is the same as in Demand Planning, hence, it could be skipped if the corresponding learning unit in Demand Planning has already been worked through.

Now, everything is prepared to start the SNP optimizer. For the basic Frutado model we will use LP. An optimizer profile has to be set up. There are a number of options still available which further specify the model to be solved. For instance, one may start with an infinite model run, in order to find out what the “ideal” solution would be if there were no bottlenecks at all (see [Fig. 5.5](#)). I.e. none of the “capacity constraints” will be activated.

Once the optimization run is finished its results may be reviewed. E.g. you may have a look at the different cost components in the objective function (see [Fig. 5.6](#)) or view the receipts of all location-product. Also, graphs showing the utilization rates of filling lines may be obtained (see [Fig. 5.7](#)).

Transfer of the Planning Results

Once the medium-term plan created by the SNP optimizer has been accepted by the decision maker it can be transferred to the subordinate planning level

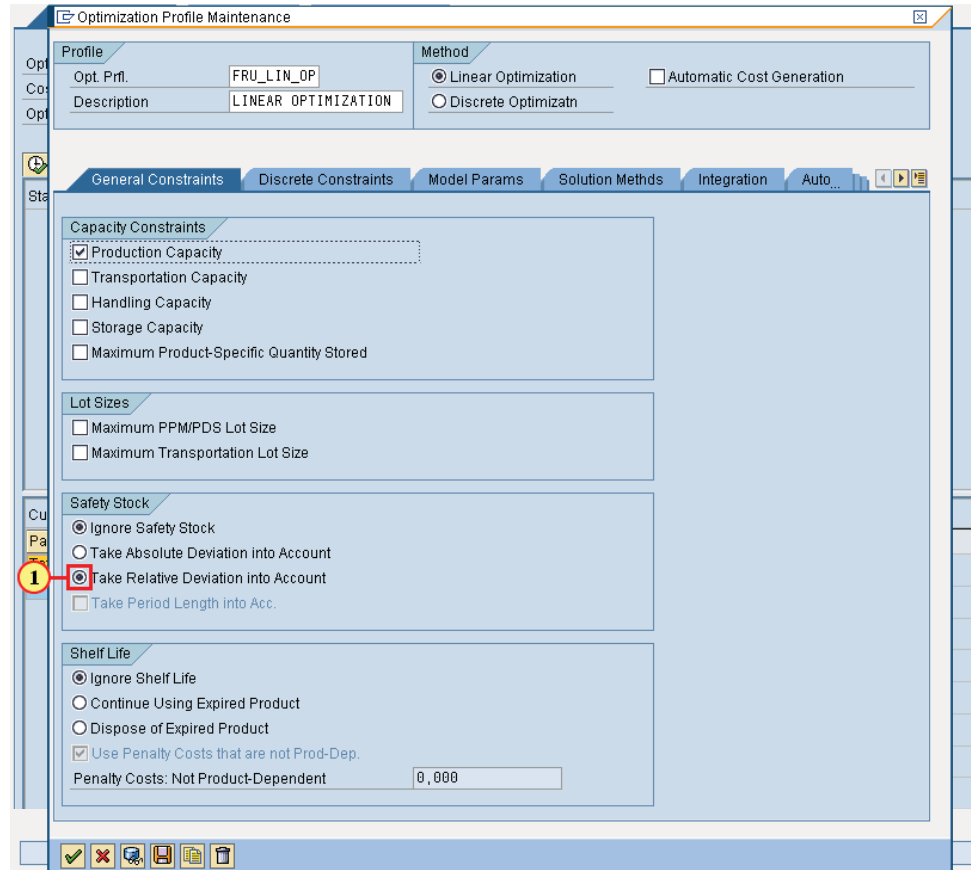


Figure 5.5
SNP optimizer profile
© Copyright 2011. SAP
AG. All rights reserved

– the PP/DS module – where the bucket oriented SNP planned orders are sequenced and scheduled on the resources (the filling lines here) in continuous time.

As a first step this learning unit shows how to generate SNP PPMs from PP/DS PPMs automatically. The procedure chosen assumes that setup times are not converted into the SNP PPMs. Instead the portion of the capacity lost due to setups is input at the SNP level manually (and may be based on the loss of capacity observed in the past on respective filling lines). This uploading of data can be regarded as a kind of *feed-forward-bottom-up* mechanism (see Section 2.1). A software specific requirement for the interface between SNP and PP/DS is that filling lines have to be defined as *mixed resources*.

A further feature of PP/DS PPMs is that there may be alternative modes of operation within a single PPM. As an example consider Ice Tea 04 which may be produced on both filling lines in plant 03. This can be modeled in one single PP/DS PPM while two SNP PPMs are needed. This has to be taken into account when creating SNP PPMs from PP/DS PPMs.

As directives we will have the SNP production orders calculated by the SNP optimizer covering the planning interval of detailed scheduling (4 weeks here). Note that these production orders may – in part – have been created

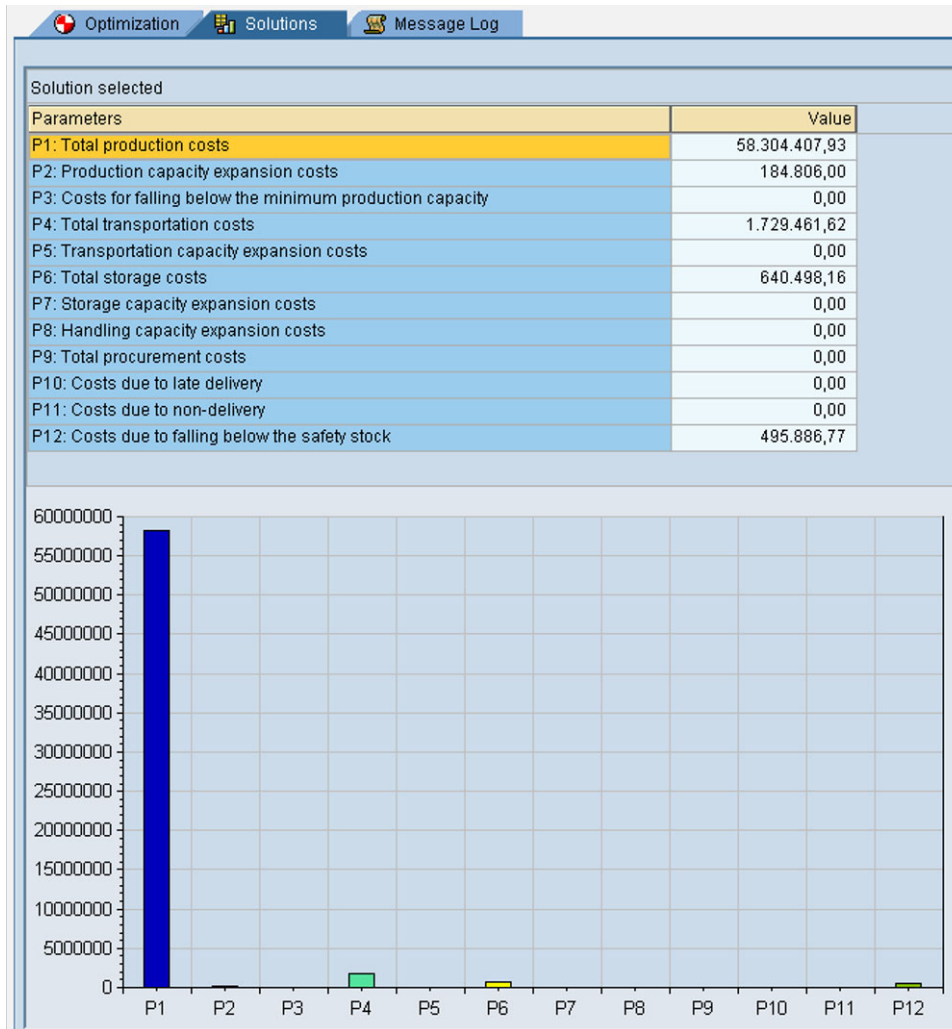


Figure 5.6
SNP optimizer result
© Copyright 2011. SAP
AG. All rights reserved

to build up seasonal stocks in order to cope with future bottlenecks on the filling lines. *Pegging* is needed to link the demand elements with the planned orders.

The “production planning” part of the tasks of the module PP/DS is already executed when converting SNP orders to PP/DS orders: Instead of transferring each SNP order within the PP/DS planning interval in a 1:1 fashion a number of alternative heuristics may be used: An SNP planned order may be converted into (several) fixed PP/DS orders (i.e. lot sizes) or an SNP planned order is split if it exceeds a given maximum lot size. It may even be possible to insert one’s own *lot-sizing* heuristic here (via a so called *BAdI* (Business Add-In)).

As a starting point for detailed scheduling a rough – and often infeasible – initial schedule is created on the filling line while transferring the SNP planned orders to PP/DS. This is done by simple heuristics. As an example for each SNP time bucket the planned and probably converted orders are

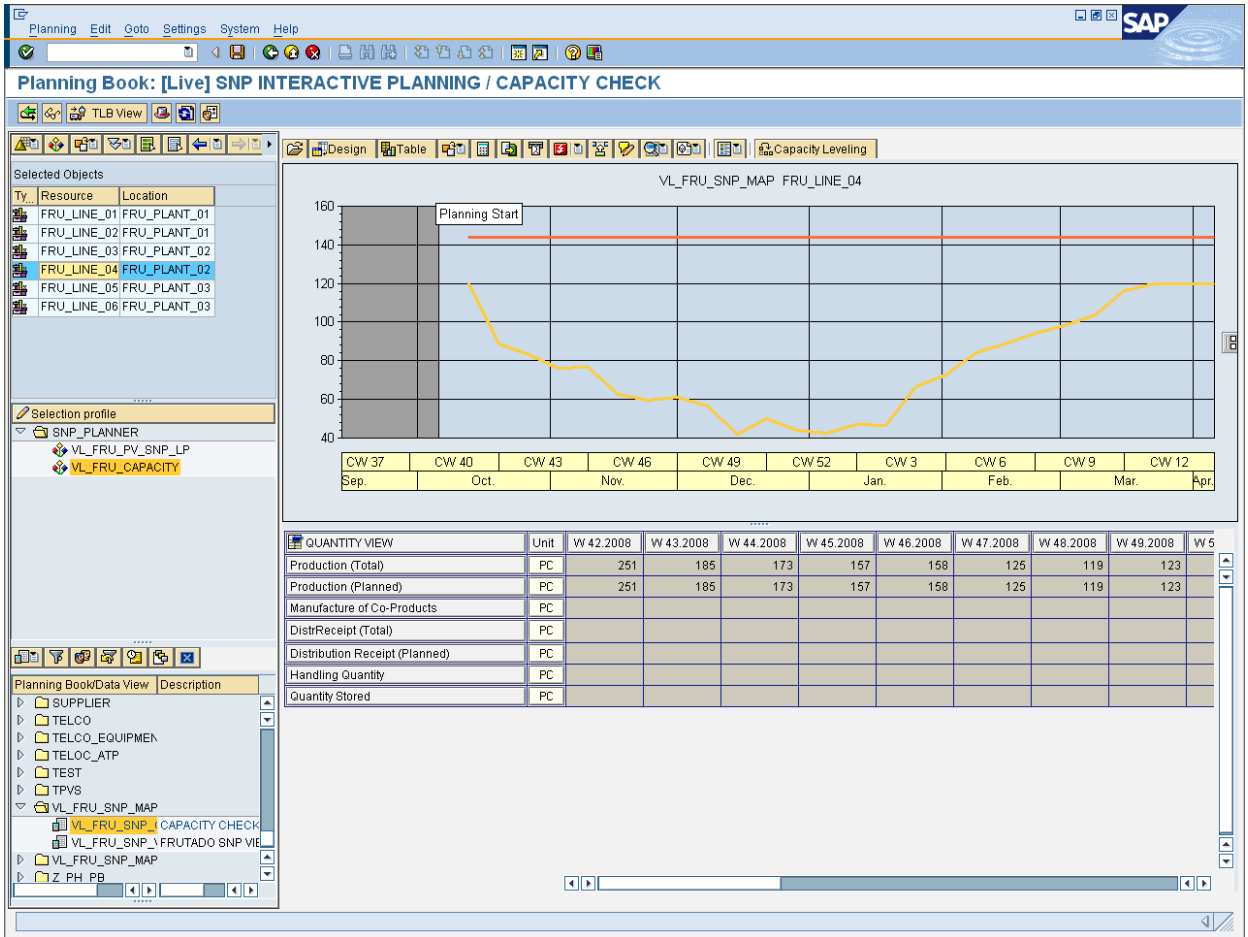


Figure 5.7
Planning book

© Copyright 2011. SAP AG.
All rights reserved

inserted into the schedule of the corresponding filling line starting at the end of the respective time bucket. Further planned orders are then inserted by backward scheduling in continuous time. In order to be able to load all SNP planned orders into the interval of the corresponding SNP time bucket *infinite* capacities of the filling line are assumed here. Note, finite capacities will be considered once the detailed scheduling meta-heuristic available in PP/DS is applied (see Chap. 6). A further way to remedy scarce resources is to insert planned orders in later periods (a kind of backordering) which is called the “backward + reverse” option.

5.6.3 In-Depth Stream

The *in-depth stream* is separated into three parts:

- The addition of a new planning version,
- modeling overtime in discrete units of Saturdays, and
- the introduction of lot size production.

In order not to delete results generated in the basic learning unit it is necessary to create a new planning version for the in-depth stream. To save efforts the new planning version starts with a copy of the current basic planning version. Alterations and extension are then executed in the new planning version. As this is more or less a bookkeeping task this part may be skipped, unless you are working with SAP APO directly.

In the basic model overtime capacities have already been taken into account. Extra costs compared to production in standard time have been added to the products' variable production costs and an upper limit per period and resource has been provided (see constraints 5.11). In reality the overtime option is not allowed in arbitrary quantities on Saturdays. Complete shifts have to be planned and manned, hence, overtime on Saturdays has to be either taken for an interval of 24 [h] or not at all (see constraints 5.23). To implement this option two capacity variants have to be defined, one for standard time and the other for overtime. The corresponding binary variables have to be defined as "discrete constraints" over the complete planning interval of 26 weeks. A special feature are "cross-period lot sizes". This option allows, that production of a product started in standard time in a given week may continue during overtime without starting a new setup activity. Note that in big bucket models two setups are required, if production of a product is continued from one period to the next. This does not apply if the "cross period lot size" option is activated (which is advisable in light of the last part of this in-depth stream).

Finally, the "discrete production capacity increase" option has to be activated in the SNP optimizer profile. Furthermore, we have to specify the number of periods (26 weeks, here) the integrality requirement must hold starting with the first period in the planning interval. Now, the MIP solver may be started. As expected, the minimal costs of this run are higher than those of the basic model. A closer look at the objective function reveals that not only overtime costs (named "production capacity expansion costs" in SNP, [Table 5.13](#)) are higher if the overtime option is discrete. Also, there are significant penalty costs for falling below safety stock levels as well as an increase in inventory holding costs by nearly 500,000 [MU] (see "storage costs" in [Table 5.13](#)).

[MU]	Basic model	Overtime, discrete	Lot sizes
Total production costs	57,820,281	58,304,408	58,307,166
Production capacity expansion costs	101,438	184,806	133,482
Total transportation costs	1,719,848	1,729,462	1,720,576
Total storage costs	158,903	640,498	640,797
Costs due to falling below safety stocks	0	495,887	495,887
Overall total	59,800,469	61,355,060	61,297,908

Table 5.13
Details comparison of the cost elements in the basic, discrete overtime model, and lot size model

In order to make the best use of producing 24 hours on Saturdays some production to stock is recommended. Furthermore, it is advantageous to

deplete safety stocks in some periods in place of a complete additional day of production.

The last extension considered here, is the explicit modeling of *lot sizes*, its costs, and setup times. Now, resource consumption due to setups no longer has to be estimated in advance as in the basic stream.

Like in the previous part we start with creating a new planning version which includes the basic model's data. Then we set capacity utilization rates at 100 percent – both for standard time as well as overtime. Cost functions now have to be modeled as a specific piecewise linear cost, namely a fixed cost once a lot is started and a variable (production) cost coefficient for each unit produced. There are two alternatives to input these cost functions – either with the help of the supply chain engineer or by the corresponding PPM. The in depth stream will present both alternatives for a single location-product.

Furthermore, setup times have to be specified for each location-product in the PPM. Note that these setup times actually are sequence dependent. However, this cannot be modeled in a big-bucket model. Hence, the average setup time for a location-product over all possible predecessors is calculated beforehand and is input to the PPM. Although this figure is usually imprecise, it may be acceptable if setup times for a location-product do not vary (regarding its predecessors). The “true” setup times will become known as soon as the schedule for each filling line is created (see the PP/DS module).

Lot-sizing incurs binary variables for setup decisions. Hence, discrete optimization has to be chosen. In the discrete constraints section of the SNP optimizer we also activate the minimum lot size requirement, i.e. if there is a lot size in a period for a certain location-product, the production quantity may not fall below the minimum lot size.

In order not to wait too long for the final result of the SNP optimization run, an (upper) time limit of 15 [min.] has been set. Consequently, it may well be, that an optimal solution cannot be found within the time limit given. The last slide of the in-depth stream exhibits the CPU-times of the SNP optimization runs: While an optimal solution for the basic model has been calculated within a few seconds the search for an optimal solution for the lot size model was terminated at the time limit of 15 [min.]. Still, a feasible solution is available.

Total production costs now include the fixed costs of lot-sizing. As in the model with discrete overtimes a solution was generated with a considerable cost increase in storage costs and costs due to falling below safety stocks (see [Table 5.13](#)). A closer look at the number of planned orders (lot sizes) reveals that there are 703 planned orders in the basic model and only 593 in the lot size model (each requiring a setup). The increase in storage and variable production costs can be explained as the SNP optimizer looks for a compromise between (the newly introduced) fixed setup costs and inventory holding cost. I.e. in comparison to the basic model the lot size model combines demands of adjacent periods to one lot size (as long as the additional holding costs are smaller than those of a setup).

Questions and Exercises

The following questions should be answered while you are working through the learning units.

1. Which options have been chosen for the SNP optimizer to solve the basic Frutado model?
2. In the Frutado model safety stocks are given. There, safety stocks may be used to satisfy demand if necessary (i.e. order to obtain a feasible solution). Is this the “usual” way safety stocks are used?
3. What does a “bucket offset” during production of 0.5 mean?

Bibliography

- Billington, P. J.; McClain, J. O.; Thomas, L. J. (1983) *Mathematical programming approaches to capacity-constrained MRP systems: Review, Formulation and Problem Reduction*, Management Science, vol. 29, no. 10, 1126–1141
- Gunnarsson, H.; Rönnqvist, M. (2008) *Solving a multi-period supply chain problem for a pulp company using heuristics—An application to Södra Cell AB*, International Journal of Production Economics, vol. 116, 75–94
- Hanssmann, F.; Hess, S. W. (1960) *A linear programming approach to production and employment scheduling*, Management Technology, vol. MT-1, no. 1, 46–51
- Hillier, F. S.; Liebermann, G. J. (2005) *Introduction to Operations Research*, McGraw-Hill, Boston, 8th ed.
- Kanyalkar, A. P.; Adil, G. K. (2005) *An integrated aggregate and detailed planning in a multi-site production environment using linear programming*, International Journal of Production Research, vol. 43, 4431–4454
- Mula, J.; Peidro, D.; Diaz-Madronero, M.; Vicens, E. (2010) *Mathematical programming models for supply chain production and transport planning*, European Journal of Operational Research, vol. 204, 377–390
- Shirodkar, S.; Kempf, K. (2006) *Supply chain collaboration through shared capacity models*, Interfaces, vol. 36, no. 5, 420–432
- Sodhi, M. S.; Tang, C. S. (2009) *Modeling supply-chain planning under demand uncertainty using stochastic programming: A survey motivated by asset-liability management*, International Journal of Production Economics, vol. 121, 728–738
- Stadtler, H. (2008) *Linear and mixed integer programming*, in: H. Stadtler; C. Kilger (Eds.) *Supply Chain Management and Advanced Planning*, Springer, Berlin et al., 4th ed., 517–527

- Tempelmeier, H. (2005) *Bestandsmanagement in Supply Chains*, Books on Demand, Norderstedt, 1st ed.
- Vollman, T. E.; Berry, W. L.; Whybark, D. C. (1997) *Manufacturing planning and control systems*, Irwin/McGraw-Hill, New York, 4th ed.
- von Lanzener, C. H. (1970) *Production and employment scheduling in multistage production systems*, Naval Research Logistics Quarterly, vol. 17, no. 2, 193–198
- Walther, G.; Spengler, T. (2005) *Impact of WEEE-directive on reverse logistics in Germany*, International Journal of Physical Distribution and Logistics Management, vol. 35, 337–361
- Wassermann, R.; Lautenschläger, M.; Reuter, B.; Steiner, C. (2006) *Automotive Case Study*, Real Optimization with SAP APO, Heidelberg, Berlin
- Waters, D. (2003) *Inventory Control and Management*, John Wiley, 2nd ed.
- Winston, W. L. (2004) *Operations Research: Applications and algorithms*, Thomson/Brook/Cole, Belmont, California, 4th ed.

Production Planning and Detailed Scheduling (PP/DS)

Hartmut Stadtler¹, Christopher Sürie²

¹ University of Hamburg, Institute for Logistics and Transport, Von-Melle-Park 5, 20146 Hamburg, Germany

² SAP Deutschland AG & Co. KG, Hasso-Plattner-Ring 7, 69190 Walldorf, Germany

The module Production Planning / Detailed Scheduling (PP/DS) of SAP[®] APO is concerned with lot-sizing decisions at the production planning level and subsequently with sequencing and timing of these lot sizes on respective resources at the detailed scheduling level.

Even if lot-sizing is already addressed in medium-term planning (see Chap. 5), directives regarding production quantities have to be disaggregated into volumes of products and their variants as well as their components and dependent items (in general: items). However, the level of detail considered refers to operations and not items. The relationship between these terms will be explained by means of an example: An item usually requires several operations for its creation, like cutting a metal plate (1st operation), grinding of edges (2nd operation), and painting (3rd operation). An operation is defined by an item *and* a (primal) resource necessary for its execution. Note that an operation might also require a secondary resource in parallel (e.g. personnel, tools). An operation may even be divided further into activities, e.g. a setup activity and a processing activity.

As we are concentrating on the planning tasks of the Frutado case we will omit the production planning level because of three reasons:

- The only (bottleneck) production stage to consider are the filling lines.

- Lot sizes are taken as given (period by period as a result of the SNP run).
- Products at the SNP level have not been aggregated to product groups, and hence there is no need for disaggregation.

The interested reader is referred to Stadtler (2008) for a description of the MRP logic used to generate secondary demands at the production planning level. A detailed description of the various options (e.g. for lot-sizing and the generation of the pegging net) is provided in Dickersbach (2009, pp 235).

The granularity of time in scheduling is nearly continuous (e.g. seconds or minutes). Also, distinct resources are considered (not groups of similar resources).

This chapter comprises six sections. In Section 6.1 we start with a differentiation of typical production processes according to some organizational principles leading to specific production segments. These production segments usually show some typical requirements regarding the planning of lot sizes and the generation of detailed schedules. Furthermore, this section introduces a model for the resource constrained project scheduling problem which forms the basis for detailed scheduling.

In Section 6.2 we present some popular solution techniques, namely we will focus on priority (sequencing) rules and a prominent meta-heuristic – a genetic algorithm.

Section 6.3 describes the data as well as the lot-sizing and sequencing decisions in the Frutado case. This gives rise to the detailed model used for Frutado with special emphasis on the production process model (see Section 6.4).

Detailed schedules are required as an input to other APS modules as well as for execution (usually via an ERP system, see Section 6.5). Finally, Section 6.6 explains the content of the learning units.

6.1 Operating Principles of Production Segments

6.1.1 Criteria

Production planning and scheduling concern the most detailed planning tasks for a firm's production and service processes which finally result in the execution of these schedules. Consequently, we have to carefully model all relevant options, restrictions, and operating principles. A closer look at industrial production processes reveals that there are specific operating principles which have become best practices across many industries. These operating principles result from four main criteria (see [Table 6.1](#)), namely

- the layout of the production process,
- a stable and high utilization rate of resources by a few similar products,
- the number and diversity of the different products to produce, and

- the number of production stages (operations) to perform.

A specific combination of these criteria gives rise to form a so called *production segment* where the most efficient production process(es) are realized. For each production segment specific planning and scheduling procedures are proposed (see Drexel et al. 1994). Next, we will provide a short description of three production segments where the PP/DS concept of SAP® APO may be applied advantageously:

- Job shops (and their variants: flow shops and open shops)
- Flow lines with setups
- One of a kind production

Finally, some remarks regarding JIT lines and paced assembly lines are added.

	Layout of production	Utilization rate	Number of different products	Number of production stages
Job shop, flow shop, open shop	Grouping of similar resources in shops	Unstable	Many, great diversity	Many (e.g. > 10)
Flow line with setups	Routing determines the ordering of resources	Stable, high (e.g. > 0.8)	Few, but many variants possible	1 to 3
One of a kind production	Resources are moved to the location of the product	Unstable, low to high	Few individual products (projects)	Large number (e.g. > 100)

Table 6.1
Three prominent production segments and four criteria for its creation

6.1.2 Job Shops

Description

Machines in a job shop, a flow shop or an open shop which are able to perform specific operations (like welding, drilling, or coating) are grouped together in specific areas. Machines in such an area (a shop) may substitute each other (i.e. parallel machines). In a job shop the sequence of shops to visit for completing an order will differ for at least two orders, while in a flow shop the sequence is the same for all orders. In an open shop there are some orders for which the sequence of shops to visit is not fixed but is a decision in the course of planning and scheduling.

If we take a look at a shop then we will observe that there are a great number of jobs either waiting or being processed (i.e. several hundreds or even thousands). Here, a *job* is a synonym for an operation (often related to a production order). The demand for items produced in a shop is neither

sufficient nor stable enough to fully utilize a machine in the long run if dedicated to this product (group) or item. Hence, different items are produced on a resource sequentially with some setup efforts in between. Consequently, lot-sizing plays an important role. Products or items manufactured show a great diversity (e.g. of dimensions). To complete an order a large number of operations has to be performed.

Planning Concept

In order to reduce complexity resulting from the large number of operations and the many resources taken into account simultaneously the proposal is to separate the planning tasks into (at least) two levels, production planning first and sequencing and scheduling second. For both planning tasks simple and elaborate heuristics are available.

In production planning the planned orders have to be exploded into dependent demands of items by means of the bill of materials (similar to the ERP logic). Furthermore, lot-sizing has to be performed for the operations created (in contrast to the ERP logic where lot sizes are created at the item level). These two tasks can either be performed simultaneously or sequentially.

For the simultaneous approach we refer to the multi-level capacitated lot-sizing problem (MLCLSP) and the solution algorithms proposed (see Sahling 2010) and the literature listed there). Although there has been great progress in solving the MLCLSP (and its variants) in recent years computational times for its solution still limit its application to small shops.

The second planning level concerns sequencing and scheduling of operations on respective resources. Solution approaches range from simple priority rules to elaborate meta-heuristics.

6.1.3 Flow Lines with Setups

Description

Machines or work stations in a flow line are arranged according to the sequence of operations to be executed for completely manufacturing a product, i.e. its routing. There is (nearly) no storage space in between two stations. As a result waiting times, and thus throughput times, are rather short. A flow line usually allows us to produce a small range of products often having a large number of variants. Due to the specific arrangement of (specialized) resources and the associated operating costs a high utilization rate is important.

Flow lines can be differentiated into paced or unpaced lines, capable of producing a single product or multiple products, with or without significant setup efforts, and a small or large number of production stages (i.e. operations to be performed to complete a product).

We will concentrate on flow lines typical for food production, like in the Frutado case. Here, we have an unpaced flow line composed of one to three production stages. A relatively small number of products is produced (often

< 10) possibly with many variants. A flow line which is discussed frequently in the literature is a make & pack system (see Neumann et al. 2002). There we have significant setup times and costs often associated with cleaning or changing the sizes of packing devices. In food production products are made to stock, i.e. based on demand forecasts.

Planning Concept

In order to minimize setup costs and inventory holding costs of end products lot-sizing and scheduling should ideally be done simultaneously. This also allows a high utilization rate of flow lines. Models are either continuous time models or small bucket models.

Several small bucket model types have been developed for single stage flow lines. The planning interval of a small bucket model is divided into a large number of small intervals of time – e.g. an hour, a shift, or a day. However, it is not the length of the interval of time (alone) which characterizes a small bucket model, but rather the assumption that at most *one* or *two* different products (lots) may be produced within a time bucket. An example is the continuous setup lot-sizing problem (CSLP) for the former and the proportional lot-sizing and scheduling problem (PLSP) for the latter model. Note there are further assumptions which constitute a specific small bucket model (like the consideration of sequence dependent setup efforts or the interruption or continuation of a setup state while the flow line is empty). For a review of the most popular model types the reader is referred to Meyr (1999) and Suerie (2005).

The single stage models described above may also be of help for two or three stage flow lines provided there is an unequivocal bottleneck stage where a single stage model can be applied. Based on the plan created for the bottleneck stage, downstream operations may be scheduled by a simple forward heuristic whereas backward scheduling is employed for upstream operations. Solution approaches range from simple heuristics to Mixed Integer Programming (MIP) solvers.

A less elegant but practice oriented solution approach is to subdivide lot-sizing and scheduling of a flow line into two planning stages – lot-sizing first and scheduling second. This may be justified in case setup efforts are not sequence dependent, and lot-sizes are either fixed or restricted to a small range due to technical reasons. Note that varying lot sizes is a rather cheap means to circumvent a temporal bottleneck situation on a resource (Stadtler 2007).

Once lot sizes have been specified the scheduling of these lots is similar to a job shop (and thus may be solved by the same procedures). However, the timing of lot sizes on successive production stages may be more restrictive (e.g. linked by minimal or maximal distances). Furthermore, limited storage space may complicate planning efforts. Sophisticated algorithms may be of great help. For instance Baumann and Trautmann (2010) have introduced a very promising MIP model for scheduling a make & pack production.

Finally we would like to mention an approach popular in industrial practice that eases the scheduling of flow lines – *cyclic scheduling* (see Mayr 1996 and Levner et al. 2010). Here, the sequence of items is fixed in the first stage. This sequence is repeated after a fixed interval of time (e.g. every fortnight). Lot sizes may vary slightly according to the items’ demands and can be adapted at short notice (second planning stage).

6.1.4 One of a Kind Production

Description

Products in a one of a kind production segment are customer specific. Hence, lot-sizing of end products is not applicable. Mostly, these end products are rather big, like a turbine, a ship or a bridge – and they are difficult to move. Consequently, personnel, machines, and input materials have to be made available at the location where the end product is finally assembled.

Often a company creates a few different end products in parallel at different locations, which require the coordination of the availability of resources at the time needed. Keeping internal and external (customer) due dates is a major issue. Usually, there are hundreds or even thousands of operations to be performed until an end product is completed. Most operations are linked to predecessor and successor operations – so called precedence relations – in the form of minimal and maximal time lags.

An end product is regarded as a project. Several projects competing for the same resources will be combined to an artificial super project and modelled by the so called resource constrained project scheduling problem (RCPSP).

Since the RCPSP is central for detailed scheduling in SAP® APO we would like to present a MIP model formulation for a single project in the most basic planning situation (see Pritsker et al. 1969 and Klein 2000 for further model formulations). For each operation j we are able to calculate a feasible time window $[EF_j, LF_j]$ in a preprocessing step: The earliest finishing time EF_j and the latest finishing time LF_j of an operation j can be calculated by forward and backward recursion (Klein 2000, p. 78).

The following model of the RCPSP results:

$$(6.1) \quad \text{Min} \quad \sum_{t=EF_j}^{LF_j} t \cdot X_{jt}$$

s.t.

$$(6.2) \quad \sum_{t=EF_j}^{LF_j} X_{jt} = 1 \quad \forall j \in J$$

$$\sum_{t=EF_k}^{LF_k} t \cdot X_{kt} \leq \sum_{t=EF_j}^{LF_j} (t - \Delta_j) \cdot X_{jt} \quad \forall j \in J, k \in P_j \quad (6.3)$$

$$\sum_{j \in JT_{rt}} a_{jr} \cdot \sum_{s=\min\{t, EF_j\}}^{\max\{t+\Delta_j-1, LF_j\}} X_{js} \leq cap_{rt} \quad \forall r \in R, t \in T \quad (6.4)$$

$$X_{jt} \in \{0, 1\} \quad \forall j \in J, t \in [EF_j, LF_j] \quad (6.5)$$

Symbols

Indices and index sets

J	set of operations j, k , where \bar{j} is the last operation that completes the project
JT_{rt}	set of operations that may be processed during period t on resource r
R	set of available resources r (e.g. personnel, machines)
P_j	set of direct predecessors of operation j
T	set of periods t, s in the planning interval

Data

a_{jr}	capacity needed per period for operation j on a resource r during processing time Δ_j
Δ_j	processing time of an operation j
cap_{rt}	available capacity of resource r in period t
EF_j	earliest finishing time of operation j
LF_j	latest finishing time of operation j

As decision variables we have (see constraints (6.5)):

Variables

X_{jt}	1, if operation j is finished at the end of period t , 0 otherwise
----------	--

The aim is to find a schedule that minimizes the throughput time of the project (constraints (6.1)) or equivalently to find a sequence of operations $j \in J$ (with \bar{j} being the last operation of the project) such that the project is completed as early as possible. The first type of constraints (6.2) secures that each operation j is finished in its feasible time window. Furthermore, all precedence relations have to be obeyed (constraints (6.3)) taking into account the duration Δ_j of an operation. No resource $r \in R$ must be overloaded for any period $t \in T$ (constraints (6.4)). For calculating the capacity required on resource r in period t (see left hand side of (6.4)) we check for each operation j that may be produced in period t on resource r (i.e. $j \in JT_{rt}$) whether it is finished in one of the periods t to $t + \Delta_j - 1$. If this is the case (indicated by $X_{jt} = 1$) then we have to load the corresponding resource r in period t with a_{jr} capacity units for processing operation j .

Note that this model formulation is based on time buckets. The associated modeling defect compared to a continuous time model may be small provided the length of periods is rather short. The MIP model presented is a *descriptive* model to precisely state the (most basic) decision situation of an RCPS. It is not intended to be solvable by a MIP solver for problems of realistic sizes.

Planning Concept

In a one of a kind production several types of decisions with different planning horizons exist. At the top we have to decide which customer orders to accept and which due dates to promise. Here, a very rough network of aggregated operations constituting a project will be considered and embedded into the rough network of already accepted but still unfinished projects. Depending on the industrial sector a planning interval of six months to several years seems appropriate.

In the short-term operations to be executed within the next two to four weeks have to be assigned to resources and put into a sequence. Thereby, detailed timing restrictions and capacities of resources have to be taken into account. Operations of several projects have to be scheduled simultaneously if they compete for the same resources.

For both levels an RCPS representation with different degrees of aggregation and possibly different objective functions is recommended.

6.1.5 Further Production Segments

A quick glance at two other prominent production segments will round off our picture of operating principles and associated planning concepts. However, since these planning concepts require a totally different approach these segments will not be considered further in this book.

The main planning task for a *JIT line* is to structure the production process such that a one-piece flow (without setup efforts) is possible. A subsequent planning task is to smooth production which can be done at the

master planning level. The sequencing of items on the JIT line may finally be done decentrally by the operating personnel (manually).

Special purpose algorithms have been developed for *paced assembly lines*, like those for the final assembly of cars. Special attention is paid to assembly line balancing in the medium-term and to finding a sequence for loading the assembly line with the different product variants in the short-term. As a result of sequencing no station in the line should be overloaded while the desired output for each shift is reached (for a detailed description of these planning tasks see Scholl 1999 and Boysen et al. 2009).

6.1.6 Conclusions and Additional Remarks

Let us recall that due to the large number of operations to coordinate in a *job shop* we proposed to divide short-term planning into two levels – production planning (including lot-sizing) and detailed scheduling. For *flow lines with significant setups* a simultaneous lot-sizing and scheduling would be ideal. However, until now powerful models and solution algorithms only exist for single stage flow lines. Hence, in order to be able to also make short-term plans for flow lines with two to three production stages a separation of the planning task into two levels is a must. As regards *one of a kind production* lot-sizing does not apply, hence detailed scheduling in the form of an RCPSP is the method of choice.

Now, it is not surprising that the solution concept of SAP[®] APO for production planning and detailed scheduling is divided into these two planning levels. In SAP[®] APO detailed scheduling for the first three production segments mentioned can be modelled by an RCPSP and solved by a powerful meta-heuristic. This approach is also well justified by the theoretical work of Drexl (1990), who has shown that some job shop scheduling problems can be regarded as special cases of a (multiple mode) RCPSP with respect to the structure of its mathematical models.

Questions and Exercises

1. Assume there is a job shop which is controlled by a simple priority rule. Can we expect throughput times which are less than double the sum of processing times?
2. What is the main characteristic of a small bucket model compared to a big bucket model?

6.2 Detailed Scheduling – Solution Algorithms

6.2.1 Overview

Solution algorithms can broadly be classified into exact and heuristic methods. Exact methods make sure that an optimal solution will be identified (if there

is only a single objective) while a heuristic comprises a set of rules that aims at finding a feasible, good, but not necessarily optimal solution with “reasonable” computational efforts. Heuristics may vary from simple rules to a set of elaborate algorithms executed until a given stopping rule is reached. In any case a compromise between solution quality and computational efforts has to be looked for. Note, that heuristics usually do not guarantee to find a feasible solution even if one exists.

An example of a simple heuristic is the *nearest neighborhood heuristic* for creating a traveling salesman tour. It starts from one city and then sequentially adds the city which is nearest to the last city included in the (sub-) tour – until all cities have been visited.

It may even be possible to use an algorithm – originally designed as an exact method (like branch and bound) – as a heuristic, namely, if the search is stopped before discovering or proving an optimal solution, e.g. due to a given upper limit on computational times.

Nearly all sequencing and scheduling decision problems cannot be solved exactly with polynomial computational efforts (with the exception of some special cases with specific objective functions: For instance Johnson’s rule provides an exact solution to a two stage flow shop problem minimizing the makespan (see the textbook Silver et al. 1998)). Hence, heuristics are used for detailed scheduling in APS.

We further distinguish heuristics into construction and improvement heuristics. *Construction heuristics* generate a solution from scratch. The emphasis lies on a feasible solution and not so much on the objective function value. The latter is the aim of improvement heuristics, which start from a given solution and try to improve it as much as possible. The nearest neighborhood heuristic is a construction heuristic. However, making swaps between two or more cities in a given tour would be an *improvement heuristic* (like the 2-opt heuristic). We would like to add that one could modify a deterministic heuristic (like those presented above) by incorporating and modifying the heuristic rule by a (pseudo) random number. This easily allows to generate a number of solutions without having to create a totally “new” heuristic.

In order to improve the quality of solutions *meta-heuristics* have been developed in recent years.

“A meta-heuristic is an iterative master process that guides and modifies the operations of subordinate heuristics to efficiently produce high-quality solutions. It may manipulate a complete (or incomplete) single solution or a collection of solutions at each iteration. The subordinate heuristics may be high (or low) level procedures, or simple local search, or just a construction method. The family of metaheuristics includes, but is not limited to, adaptive memory procedures, tabu search, ant systems, greedy randomized adaptive search, variable neighborhood search, evolutionary methods, genetic algorithms, scatter search, neural

networks, simulated annealing, and their hybrids.” (Voss et al. 1999, p. IX)

For instance a list of operations to be scheduled on a set of parallel machines may be produced by meta-heuristic principles, while a subordinate heuristic subsequently does the assignment and loading of the operations on the machines (according to the listed sequence). As the most sophisticated method available for detailed scheduling in SAP[®] APO is a genetic algorithm (GA) we will provide its basic principles.

6.2.2 An Example

In order to achieve a deeper understanding of the way detailed scheduling can be done, we will explain it by means of a simplified numerical example – “Frutado light”. The example will be solved by the *remaining slack time* priority rule and by a GA. In addition the basic steps a GA consists of will be introduced.

Detailed scheduling in the *Frutado light* case only considers a single filling line and six operations that have to be scheduled within the next four days (see Table 6.2). For operations “Ice Tea 3” to “Ice Tea 6” the due date is at the end of day two (equivalent to time 48 [h]) while the due date for operations “Ice Tea 1” and “Ice Tea 2” is placed at the end of day four. Due to the importance of customers the penalty costs for a delay differ between operations.

Operation	Ice Tea 1	Ice Tea 2	Ice Tea 3	Ice Tea 4	Ice Tea 5	Ice Tea 6
Production time [h]	12	24	9	12	15	6
Due date [h]	96	96	48	48	48	48
Penalty cost [MU/h]	2	1	2	1	3	2

Table 6.2
Production time, due dates and penalty costs for operations

from/to operation	Ice Tea 1	Ice Tea 2	Ice Tea 3	Ice Tea 4	Ice Tea 5	Ice Tea 6
Ice Tea 1	-	4	2	5	3	4
Ice Tea 2	5	-	6	2	4	3
Ice Tea 3	2	6	-	5	8	3
Ice Tea 4	5	2	5	-	4	2
Ice Tea 5	3	3	8	4	-	6
Ice Tea 6	3	6	3	2	6	-

Table 6.3
Sequence dependent setup times [h]

The filling line operates in three shifts (i.e. 24 [h/day]) without any break. It is in the state “cleaned” at the beginning of the planning interval (i.e. there is no setup effort for the first operation). Setup times are sequence dependent (see Table 6.3).

The objective is to find a schedule with minimal total penalty costs. Next, we will introduce some basic ideas underlying priority rules and then show how to create a solution by the *remaining slack time* priority rule.

6.2.3 Solution by a Priority Rule

Introduction

A priority rule assigns a number (named priority) to each operation in a set of operations. Then the operation with the highest priority is selected from this set for further processing. In the sixties and seventies research into the quality of schedules generated by these priority rules started. However, the quality of solutions still remains limited.

There are certain rules which tend to perform well for some objectives but worse for others. For instance the *shortest processing time* rule results in rather small throughput times in general, but falls short in keeping due dates. The *remaining slack time* rule on the other hand is good at keeping due dates but does not yield small throughput times. A straightforward idea is to build a combined rule of the two priorities. However, the more priorities are combined the less the quality of solutions can be controlled. Still, there are some promising computational experiments with more sophisticated priority rules for the RCPSP (see Hartmann and Kolisch 2000).

While the *shortest processing time* rule is a *static rule* (i.e. the priority of the operation does not change in the course of processing – assuming that setup times are not sequence dependent), the *remaining slack time* rule is *dynamic* and has to be recalculated whenever a new decision for selecting an operation has to be made.

For the *shortest processing time* rule the operation with the shortest processing time waiting to be executed on a machine will be selected. If the *remaining slack time* rule is applied to a set of operations waiting to be processed, then the operation with the smallest remaining slack time will be selected. The priority for an operation i is calculated as follows:

$$\begin{aligned} \text{Remaining slack time } (i) &= \text{due date } (i) - \text{potential start time } (i) \\ &- \text{sum of remaining processing times on the critical path } (i) \end{aligned}$$

Here, we assume that processing times will also include setup times (if relevant). We would like to add that one should not expect too much from the application of priority rules as the following line of thought shows:

Consider a job shop with many machines and a great number of operations to be processed. We will concentrate on just one of these machines including the operations waiting to be processed in front of the machine. To simplify matters we further assume that the processing time of an operation (including setup times) always takes one time unit.

Now, at a certain point in time the machine has finished an operation and there is another operation waiting to be processed. This operation is instantaneously loaded and processing starts. At the same time a new operation enters the queue. It has to wait until processing is finished, i.e. it has to wait one time unit before it is processed for another time unit. Hence, the share of waiting time on the throughput time is 50 percent. If the situation mentioned is valid for all machines and for any time instance then half of the throughput time of the system is “wasted” by waiting. Note, that this performance is valid irrespective of the priority rule employed because there is only one operation waiting and there is no choice.

Next, we assume the same situation as above except that there are always two operations waiting once the machine becomes empty. Now, we can select the next operation (out of two) and priority rules will come into play. However, the portion of waiting time is 66.7 percent of the total throughput time.

Similarly, if there are three operations waiting when the machine becomes empty waiting time has a share of 75 percent – a figure which has been observed by some researchers for a job shop.

Although the decision situation described above seems rather artificial, it shows that priority rules will only have an effect on throughput times if there are at least two operations waiting to be processed on a machine. Consequently, we cannot expect waiting times to be less than 50 percent of throughput times when applying simple priority rules. Hence, if we wish to reduce throughput times in job shops drastically, more sophisticated procedures than simple priority rules must be employed. Meta-heuristics are the method of choice in many cases. Still, despite the criticism regarding the quality of solutions priority rules may be very useful to generate initial solutions quickly as a basis for improvement methods. As regards industrial practice, priority rules are easy to understand and to implement. They may be applied for sequencing of operations both *decentrally* (at each individual machine by the workers on the shop floor) as well as *centrally* by the IT department generating a schedule for the whole production segment.

Application of the *Remaining Slack Time* Priority Rule

To start with we calculate the slack of each operation at time 0 (Table 6.4, columns 3 and 4). Since each operation consists of just one operation no summation of remaining processing times on the critical path is necessary (see column RPT). It turns out that the minimum remaining slack time (highest priority) relates to Ice Tea 5 which now starts the sequence. Processing of Ice Tea 5 is finished at time 15, where the next operation to be loaded on the filling machine has to be chosen (Table 6.4, columns 5 and 6). Now, sequence

dependent setup times have to be taken into consideration to calculate the remaining processing time. Ice Tea 3 has the highest priority and is second in the sequence and will be ready at time 32. Table 6.4 shows that the Ice Tea 4 will be chosen next, however its completion will be late by 1 [h]. Similarly the remaining operations are put into a sequence.

Operation	Due date	Time 0		Time 15		Time 32		Time 49		Time 57	
		RPT	Slack	RPT	Slack	RPT	Slack	RPT	Slack	RPT	Slack
Ice Tea 1	96	12	96-12 =84	12+3 =15	96-15- 15=66	12+2 =14	96-32- 14=50	12+5 =17	96-49- 17=30	12+3 =15	96-57- 15=24
Ice Tea 2	96	24	96-24 =72	24+3 =27	96-15- 27=54	24+6 =30	96-32- 30=34	24+2 =26	96-49- 26=21	24+6 =30	96-57- 30=10
Ice Tea 3	48	9	48-9 =39	9+8 =17	48-15- 17=16	-	-	-	-	-	-
Ice Tea 4	48	12	48-12 =36	12+4 =16	48-15- 16=17	12+5 =17	48-32- 17=-1	-	-	-	-
Ice Tea 5	48	15	48-15 =33	-	-	-	-	-	-	-	-
Ice Tea 6	48	6	48-6 =42	6+6 =12	48-15- 12=21	6+3 =9	48-32- 9=7	6+2 =8	48-49- 8=-9	-	-

Table 6.4
Dynamic calculation of slack at times of decision (Abbreviation: RPT = remaining processing times)

Finally, we recall the solution generated (sequence: 5-3-4-6-2-1) and calculate the associated total penalty costs (Table 6.5) which is 35 [MU].

Operation	Ice Tea 5	Ice Tea 3	Ice Tea 4	Ice Tea 6	Ice Tea 2	Ice Tea 1
Setup time	-	8	5	2	6	5
Finished by	15	32	49	57	87	104
Penalty costs incurred	0	0	1	18	0	16

Table 6.5
Evaluation of the sequence 5-3-4-6-2-1

6.2.4 Solution by a Genetic Algorithm

Basic Principles

Heredity of characteristics from parents to their children has proved to be a very good procedure for species to adapt to a changing world and to survive. Even more, it results in the “survival of the fittest” (Darwin). This observation has led to the formalization of the principles of (sexual) reproduction and its transfer to artificial systems (due to Holland 1975) – named genetic algorithms (GA).

These principles can be associated to the six steps of a GA as shown in Figure 6.1. We will introduce the most basic principles while we are aware that a good GA will require much more “flavor” (the reader is referred to Reeves and Rowe (2003) and Klein (2008) for further information on GA).

In a preparatory step we have to find a way to represent the decision problem at hand by *genes* which are ordered in *chromosomes* (see Figure 6.2). Mostly, genes are either digital or decimal numbers. Whether it is advantageous to have a small number of chromosomes with many genes or a

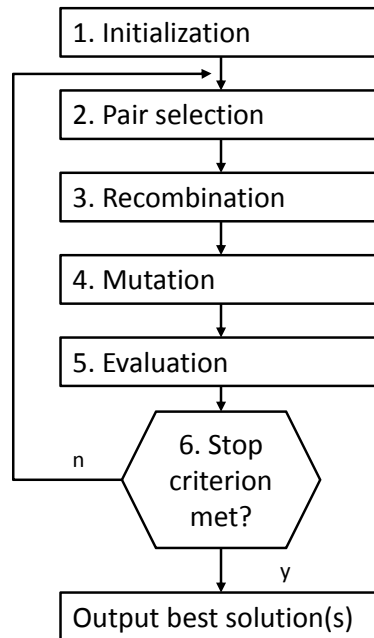


Figure 6.1
General procedure of a
GA

subdivision of genes into several chromosomes (e.g. a chromosome for each “machine”) is subject to the design of a specific GA and cannot be generalized. A set of chromosomes defines an *individual*. This abstraction is called *encoding*. One way of encoding a scheduling problem with one machine (and one chromosome) has already been exercised in the *Frutado light* example: There, a solution – sequence 5-3-4-6-2-1 – has been generated by a priority rule (note, each position in the sequence represents a gene).

Obviously, some information of the decision problem “gets lost” by encoding – like the start and finish times of each operation. This information has to be added before evaluating an individual – in our context a *solution* (see Figure 6.1, step 5): Hence, *decoding* usually incurs some further problem specific heuristic elements (like the choice of starting an operation before a break and completing it later or to process the operation completely after the break). Evaluation is also known as providing a *fitness value* to an individual. In the *Frutado light* example the objective function value (abbreviated Obj) is 35 [MU]. Since we are minimizing penalty costs we will invert the objective function value in order to obtain a fitness value (abbreviated FV) resulting in $1/35$ [1/MU]. Now, the smaller the penalty costs of a schedule is the larger its fitness value will be. Note that once we find a schedule without delays we are done and there is no need to calculate a fitness value.

At certain points in time several individuals exist – called a *population*. Individuals of a population now can give birth to new individuals by reproduction (on the basis of predefined rules). This is the start of a new iteration beginning with “pair selection” (see Figure 6.1, step 2). Given these introductory remarks we will provide some more details regarding the general steps of a GA.

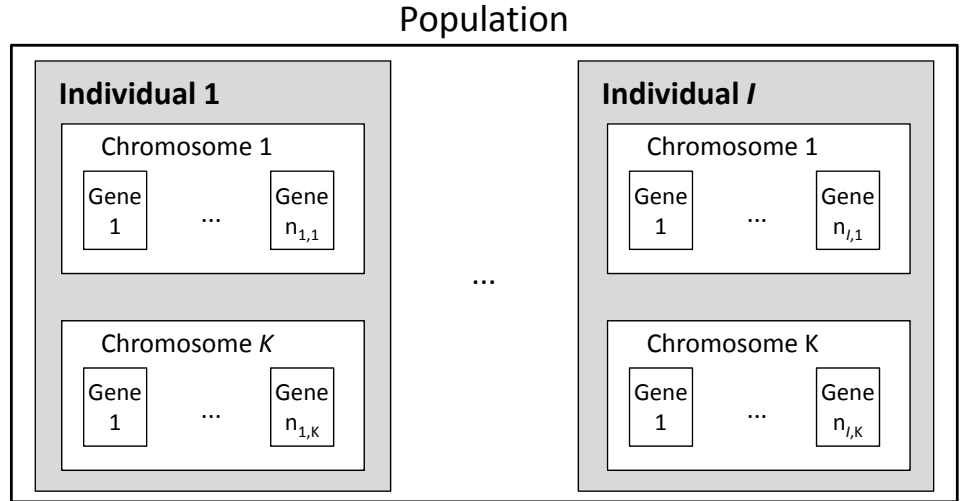


Figure 6.2
Context of population,
individuals,
chromosomes, and
genes

In the first step – *initialization* – we create the individuals that form the *start population*. Individuals will be created by a construction heuristic or even randomly. Designing such a heuristic incurs finding a compromise regarding (computational) efforts and the quality of solutions. Quality does not simply mean a high fitness value for each individual but, equally important is a high diversity of individuals in the initial population. If the GA performs well, then the fitness value of individuals of the start population should have little effect on the quality of the final solution(s) generated at the end of the GA. At the end of the first step of the GA we have to evaluate the fitness value of each individual. The number of individuals forming a population is one of the parameters of a GA.

In the second step – *pair selection* – we will have to select pairs of individuals who will pass on their genes to the next generation. Which individuals to select will depend on their fitness values. There are both *deterministic* und *stochastic* selection methods. In deterministic methods the n -individuals with *best fitness values* are selected (in pairs). Two stochastic methods will be presented next. In the *roulette wheel strategy* each individual is assigned a selection probability which corresponds to its fitness value, i.e. the larger the fitness value the higher the selection probability of the individual will be. In a *rank based strategy* each individual in the population is ranked according to its fitness value. Then an individual obtains a selection probability corresponding to its rank.

In the third step – *recombination* – we will have to define rules of heredity, i.e. the exchange of genes from parents to a child or children. One such rule is – *one point crossover*: A single position in a chromosome is chosen (i.e. between two successive genes). Now, the first child receives the genes of the first parent up to this position while the remaining genes are taken over from the second parent. A second child may be created simply by taking the other genes (see [Figure 6.3](#) for genes with digital values).

The other rule presented here is – *order crossover*. This requires to

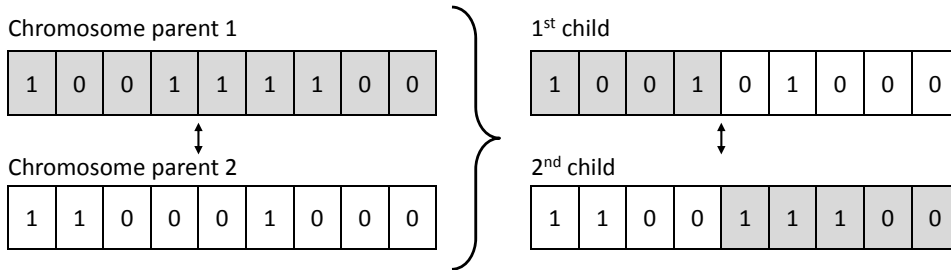


Figure 6.3
Applying the one point crossover strategy at position 4

(randomly) determine two positions separating a chromosome in three subsets. The genes in between these two positions are exchanged between parents and subsequently form the two children (see Figure 6.4).

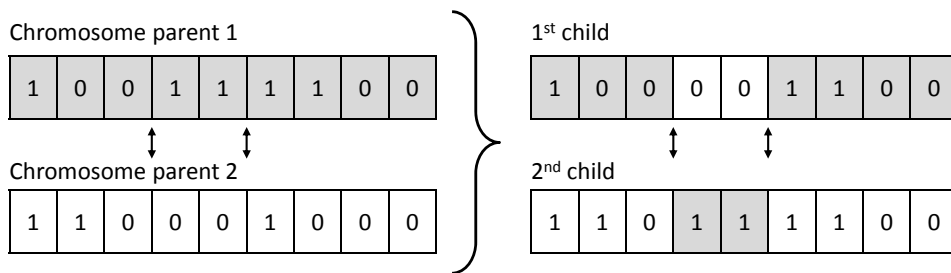


Figure 6.4
Applying the order crossover strategy at positions 3 and 5

In the fourth step – *mutation* – we exploit the observation that in the course of reproduction some defects may occur which may sometimes result in advantageous features of the child. In any case there is a chance to enlarge the diversity of the population (and in the sense of an optimization model to explore new areas of the decision space). Mutation can be achieved by randomly selecting a gene and *inverting* or altering its “value”. As an example consider the second child of the order crossover strategy, where the second gene is inverted (see Figure 6.5).

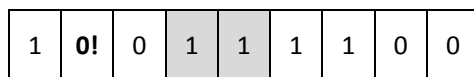


Figure 6.5
Mutation of gene two of the second child

In the last step we will check given *stopping rules*. In the case where a stopping rule is met, the best solution(s) will be presented to the decision maker. Otherwise we will continue with a further iteration, i.e. the creation of a new population (step 2). We may choose both *absolute* and *relative* stopping criteria. As an example for an absolute stopping criterion an upper limit on the number of generations or on computational times may be given. A relative stopping criterion is to require a *minimum percentage improvement* of the fitness value of the best two individuals created over the last *i* iterations.

The quality of solutions and computational efforts resulting from a GA largely depend on the encoding and decoding (heuristics) as well as the various parameters available in a GA. Usually, a number of experiments have to be conducted before a reasonable, satisfying quality will be reached. However,

GA have been used favorably both in research as well as in commercial projects. In case you are interested in the design of the GA developed by the SAP AG for PP/DS we recommend the PhD thesis of Scheckenbach (2009).

A very simple numerical example has been attached in the following in order to better grasp the ideas of a GA.

A Numerical Example

The six steps of the GA used to “solve” the *Frutado light* example have been specified as follows (Table 6.6):

1. Initialization:	The starting population consists of 4 individuals. Each individual is characterized by one chromosome with six genes.
2. Pair selection:	2 parents will be selected by a roulette wheel strategy.
3. Recombination:	Order crossover.
4. Mutation:	Random selection of a child to mutate with probability 0.1, the position to mutate is uniformly distributed over the genes of the chromosome.
5. Evaluation:	The fitness value equals the inverted penalty cost of the associated schedule.
6. Stop criterion:	After 1 generation.

Table 6.6
Specifications for the GA for the *Frutado light* example

The following random pseudo numbers will be used in the course of the GA:
0.9519; 0.3482; 0.9036; 0.1686; 0.4713; 0.0895; 0.8055; 0.5429

Now, we show the different steps in detail:

1. Initialization:

The start population is given:

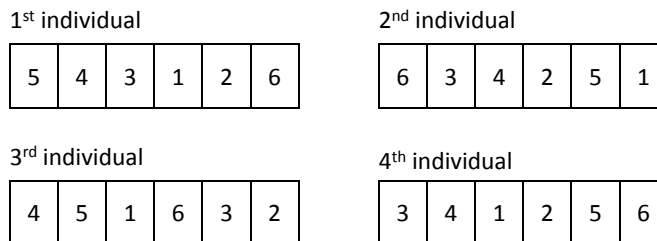


Figure 6.6
Start population of the GA

Tables 6.7 to 6.10 show the evaluation of each individual, respectively.

Operation	Ice Tea 5	Ice Tea 4	Ice Tea 3	Ice Tea 1	Ice Tea 2	Ice Tea 6
Setup time	-	4	5	2	4	3
Finished by	15	31	45	59	87	96
Penalty costs incurred	0	0	0	0	0	96

Obj(1) = 96

Table 6.7
Evaluation of the 1st individual (sequence 5-4-3-1-2-6)

Operation	Ice Tea 6	Ice Tea 3	Ice Tea 4	Ice Tea 2	Ice Tea 5	Ice Tea 1
Setup time	-	3	5	2	4	3
Finished by	6	18	35	61	80	95
Penalty costs incurred	0	0	0	0	96	0

Obj(2) = 96

Table 6.8
Evaluation of the 2nd individual (sequence 6-3-4-2-5-1)

Operation	Ice Tea 4	Ice Tea 5	Ice Tea 1	Ice Tea 6	Ice Tea 3	Ice Tea 2
Setup time	-	4	3	4	3	6
Finished by	12	31	46	56	68	98
Penalty costs incurred	0	0	0	16	40	2

Obj(3) = 58

Table 6.9
Evaluation of the 3rd individual (sequence 4-5-1-6-3-2)

Operation	Ice Tea 3	Ice Tea 4	Ice Tea 1	Ice Tea 2	Ice Tea 5	Ice Tea 6
Setup time	-	5	5	4	4	6
Finished by	9	26	43	71	90	102
Penalty costs incurred	0	0	0	0	126	108

Obj(4) = 234

Table 6.10
Evaluation of the 4th individual (sequence 3-4-1-2-5-6)

2. Pair selection:

For selecting the two individuals to pair we will use the roulette wheel strategy (Table 6.11).

Individual	Obj	Obj inverted	FV, cumulated	Probabilities cumulated and transposed in the interval [0;1]
1	96	1/96=0.0104	0.0104	0.0104/0.0423=0.2459 → [0;0.2459[
2	96	0.0104	0.0208	[0.2459;0.4917[
3	58	0.0172	0.0381	[0.4917;0.8991[
4	234	0.0043	0.0423	[0.8991;1]

Table 6.11
Roulette wheel strategy

Random number 0.9519: → Selection of the 4th individual

Random number 0.3482: → Selection of the 2nd individual

3. Recombination:

Here we have to determine the two positions in the chromosome for the exchange of genes by the order crossover rule. These points will be calculated with the help of random numbers 0.9036 and 0.1686.

$$s_1 = 0.9036 \cdot 6 = 5.4216 \approx 5$$

$$s_2 = 0.1686 \cdot 6 = 1.0116 \approx 1$$

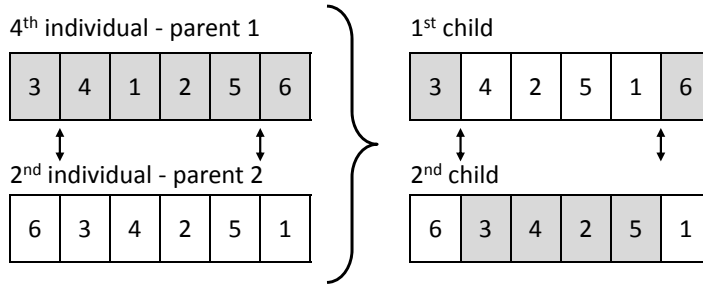


Figure 6.7
Applying the order crossover strategy at positions 1 and 5

Applying the order crossover requires an additional logic compared to the strategies with digital genes as explained in the general principles of a GA. The problem is that our encoding is based on decimal numbers referring to operations and that a simple exchange of genes often will result in a duplication or a “loss” of operations (making the solution infeasible). Consequently, recombination is done here as follows: For the 1st child we fix information of genes in the first and third subset (i.e. genes 1 and 6) according to the information provided by the parent (i.e. the 4th individual). For the genes of the second subset (i.e. genes 2 to 5) we insert the genes of the 2nd individual starting with the first gene, unless the operation indicated by the gene has already been fixed (from the 4th individual). The same procedure is applied to generate the 2nd child, which by chance results in the same chromosome as the parent (2nd individual).

4. Mutation:

Examining whether an individual will be mutated:

1st child: Random number: 0.4713 > 0.1 → No mutation

2nd child: Random number: 0.0895 < 0.1 → Mutation:

Selection of the two genes using the random numbers 0.8055 and 0.5429:

$$s_1 = \lceil 0.8055 \cdot 6 \rceil = \lceil 4.833 \rceil = 5$$

$$s_2 = \lceil 0.5429 \cdot 6 \rceil = \lceil 3.2574 \rceil = 4$$

6	3	4	5!	2!	1
---	---	---	----	----	---

Figure 6.8
Mutation of genes 4 and 5 of the second child

Note, the mutation here is not simply an inversion of a digit but an exchange of positions of two genes.

5. Evaluation:

Operation	Ice Tea 3	Ice Tea 4	Ice Tea 2	Ice Tea 5	Ice Tea 1	Ice Tea 6
Setup time	-	5	2	4	3	4
Finished by	9	26	52	71	86	96
Penalty costs incurred	0	0	0	69	0	96

Table 6.12
Evaluation of the 1st child (sequence 3-4-2-5-1-6)

Obj(C1) = 165 → Rank 5

Operation	Ice Tea 6	Ice Tea 3	Ice Tea 4	Ice Tea 5	Ice Tea 2	Ice Tea 1
Setup time	-	3	5	4	3	5
Finished by	6	18	35	54	81	98
Penalty costs incurred	0	0	0	18	0	4

Table 6.13
Evaluation of the 2nd child (sequence 6-3-4-5-2-1)

Obj(C2) = 22 → Rank 1

An improvement is achieved with the new generation, as the 2nd child possesses the best objective function value (Obj = 22) generated so far (Table 6.13). The 1st child is merely the fifth best individual (Obj = 165, Table 6.12).

6. Stop Criterion:

The first iteration has been completed and the stop criterion applies. The 1st child is the individual with the best fitness value and will be shown to the decision maker.

Questions and Exercises

1. Generate a best solution for the *Frutado light* example by one iteration of the GA (as described above) but now with the objective to minimize total setup times.
2. Name (at least) three parameters of a GA.
3. Is there a guarantee that a GA will find a feasible or optimal solution?

6.3 Planning Tasks and Data for the Frutado Company

6.3.1 Planning Tasks

Short-term planning for the Frutado company is a local planning task which is coordinated by the central Master Planning covered by the SNP module. As a result of SNP, the allocation of the production orders to plants, the planned overtime shifts per production (filling) line, and the weekend target-stocks have been determined.

Each plant has two filling lines, which can produce a subset of Frutado's products each. Only these filling lines are relevant for production planning, because the other production step (mixing) does not represent a bottleneck. Also, the input materials do not constrain the production system, because they are either stocked well in advance to avoid run-outs or contracts with suppliers allow procurement of these materials within short lead times. All plants produce in three shifts per day on five days per week with the option to use up to three additional shifts on Saturdays depending on the utilization situation. Whether or not these overtime options are available for short-term production planning for one or more of the filling lines, has been determined by the central Master Planning.

As there are no coupling constraints for short-term production planning between the plants, each plant can be planned separately. Coordination of plans has been achieved by the super-ordinate planning level. In two of the three plants, even the filling lines can be planned individually, because the allocation of products to filling lines is unique there. However, in one of the plants, the same products can be produced on both filling lines, even though with different production coefficients and at different costs. In this plant, both lines (5 and 6) have to be planned jointly to allow for a balanced decision. Short-term production planning has to cover a planning horizon of four weeks with exact continuous timing.

With respect to the definition of production segments in Section 6.1, the Frutado scenario can be classified as a flow line with setups. Thus, the short-term planning task of Frutado is to find a good schedule for each of the filling lines. Finding a good schedule has to consider several hard and soft constraints. While soft constraints can be violated, hard constraints have to be met. The following constraints are relevant for the Frutado company:

- Capacity of filling lines (hard)
- Working and non-working time of filling lines (hard)
- Shelf-life of products (hard)
- Validity of production orders (hard)
- Setup characteristic of production orders (hard)
- Due dates of production orders (soft)

The production schedule which is based on exact continuous timing has to consider the capacity of each filling line. This means, it has to take into account the production coefficients (speeds) of the scheduled products as well as the efficiency of the filling lines. Furthermore, the calendar representing the valid shift model to distinguish between working and non-working times has to be considered. As fruit juices have a limited shelf-life, this is an additional scheduling constraint for the corresponding orders. From time to time recipes for the production of fruit juices or ice teas change, e.g. because suppliers for certain ingredients get replaced. Thus, for production orders based on a recipe a validity period exists which has to be respected during planning in the absence of explicit modeling of input materials. If two products are produced subsequently on a filling line, the filling line has to be cleaned in between. The cleaning effort depends on the sequence of products, e.g. less effort is required for changing from orange juice to mixed juice than vice versa or changing from an ice tea blend to a fruit juice. Different sequences therefore imply different unproductive resource times.

One objective will be to minimize the total setup time. The second objective will be the minimization of due date violations, because the products shall reach the customers as requested. In Plant 3 (Production Lines 5 and 6) there exists a third objective that has to be considered while creating the production schedule. The objective here is to find a reasonable assignment of production orders to filling lines. While the super-ordinate planning level (SNP) has made the assignment of production orders to plants, this yields a unique assignment of production orders to filling lines in Plants 1 and 2. In Plant 3, this decision is left to the local production planner, because both filling lines in Plant 3 allow production of the same products. The assignment decision however is influenced by the different characteristics of the filling lines with respect to the production coefficient and production cost.

6.3.2 Data

The data required to perform the short-term planning tasks of the Frutado company can be distinguished in master data, transactional data, and configuration data that allows to influence the planning process flexibly. Master data is considered to be those data that does not change frequently (see Section 3.3.2). For short-term production planning relevant master data are related to the filling lines, products, and production process.

The *master data* of the filling lines (e.g. efficiency, planned downtimes for maintenance) is available locally in the plants while their temporal availability is provided from the central planner in form of valid shift models. Master data for products consists of e.g. base units of measurement for consistent calculation results or shelf-life information. Most of this information is maintained centrally at the Frutado company and therefore valid for all plants. However, there are also some location specific product data, like minimum lot-sizes which are plant-dependent and therefore maintained locally. Recipes of the different products are also plant specific and therefore maintained

individually in each plant. The main factor distinguishing the recipes between the plants is the production coefficient and for Plant 3 the option to produce a subset of the products on both resources, which basically means that the recipe contains both production methods. Finally, setup matrices are required. As each filling line has different characteristics with respect to production speed and capabilities with respect to products being able to process, the setup matrices have to be maintained individually per filling line locally in each plant taking into account the experience and knowledge of the local workforce.

Transactional data required for the short-term planning process are the detailed demand forecast (daily granularity) which is provided by the demand planning department. To be able to anticipate demand peaks outside the short-term planning horizon, the Master Planning result has to be considered, which takes into account these effects by looking further into the future and considering a longer planning horizon and providing seasonal stocks at the end of periods (weeks). Finally, safety stock requirements are important for short-term production planning.

Configuration data influences the planning process. The planning horizon and frozen zone are examples for this kind of data. For short-term production planning the processes of the Frutado company require a frozen zone of one week to be able to replenish material stocks in due time. Thus, weekly short-term production planning covers weeks 2-4 (see Section 3.4) based on exact continuous timing. Other configuration data are the weights for the different objectives. Similar to the discussion of cost elements in Section 5.3.2 the question for a planner is, how the different objectives outweigh each other. Is it more important to save one hour of setup time or to avoid the delay of one production order for one day? How is the choice of the filling line, implicating additional or less setup time as well as causing different production costs, related to this decision?

The decisions taken in short-term production planning are for each production order, which production mode is to be chosen in case several modes are available in a production order (only Products 3 to 5 in Plant 3) and when to schedule the order (start time). Additionally the planning department of Frutado wants to monitor actual setup times per filling line and per week to verify and potentially adapt the global capacity reduction for setup which is used in Master Planning.

Questions and Exercises

1. Which restrictions have to be considered when creating a production schedule for one of Frutado's plants?
2. Which are the three objectives that need to be weighted against each other in short-term production planning in the Frutado scenario?

3. Why is certain data maintained centrally and other data maintained locally in the Frutado company? Provide some examples.

6.4 Modeling the Frutado Planning Tasks

6.4.1 Basic Frutado Model

The objective function of the short-term production planning problem of Frutado contains 4 objectives that have to be minimized:

- Sum of delays of production orders
- Delay of the most delayed production order
- Sum of setup times
- Sum of mode costs

While the first three objectives are all measured in dimension time, the mode costs which represent the different production costs of products that can be produced on both filling lines in Plant 3 (Products 3 to 5) have to be converted into the time domain to be able to use a common unit of measurement for calculating the objective. This is done in SAP[®] APO by defining a cost function, that multiplies the duration of the order with a constant factor and adding a constant value on top. This way the choice of the mode, i.e. filling line, is dependent on the filling line itself and the duration the production order will occupy the filling line.

Using two different measures for the delay of production orders in the objective, calculating the delay itself as well as paying special attention to the most delayed order, is intended to have the following effect. The total delay of all production orders shall be distributed between several production orders to avoid outliers which may scare away customers. This can be illustrated by a small example: If there are five production orders for five different customers, there could be two equivalent production schedules with respect to the above mentioned objective. In one production schedule each order is delayed one day, whereas in the second schedule one order is delayed for five days and the other four orders are on time. With respect to the first objective, both schedules are equivalent. The total delay is five days in both cases. However, considering also the objective to minimize the delay of the most delayed production order, the first solution is superior, because the delay of the most delayed production order is one day compared to five days in the second solution.

Minimization of setup times will be achieved by finding a reasonable sequence of production orders for Filling Lines 1 to 4, while on Filling Lines 5 and 6 an additional means of influencing setup times is the assignment of production orders to a filling line. Other means, e.g. changing the assignment of products to filling lines in order to have a smaller variety of products per filling line, will not be considered in the basic Frutado case. Less products per

filling line would yield less setups and therefore an increase in plant output, but at the same time require additional transport between the plants/DCs. Consequently, this is a decision that would have to be made centrally on a super-ordinate planning level.

Finally weights have to be assigned to the different objectives in order to specify that one hour of setup time has a different importance than one hour of delay. For Plants 1 and 2 no weight for mode costs is required in the absence of choice between filling lines. The objective to minimizing inventory holding costs cannot be chosen in SAP® APO. One reason is that in the short term (e.g. 1 week) it is rather difficult to calculate the “true” costs of inventory.

The constraints in Frutado’s production system merely stem from the filling lines. The capacities of the filling lines determine the output and have been identified as Frutado’s bottleneck for long. The other resources like the raw material silos, the mixers, and the intermediate storage tanks that are depicted in Figure 6.9 are sufficiently available in all three plants.

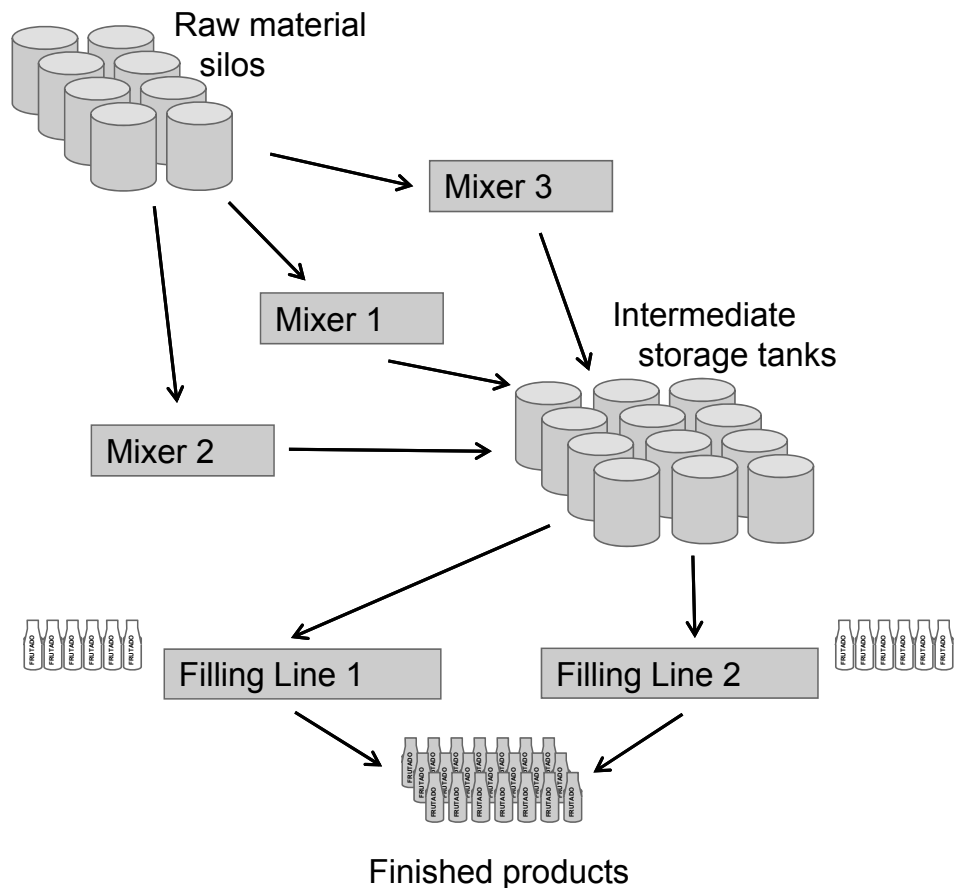


Figure 6.9
Frutado’s production process

However, capacity consumption is modeled differently in PP/DS compared to SNP. Here, capacity is considered with exact continuous timing, i.e. on a time stream the state of the resource (setup from one product to

another, production of a specific production order) is determined continuously. This calculation is based on 24h availability of the filling lines, whereas in SNP the resource availability was reduced by a constant factor to take into account unproductive times due to change-overs which have not been modeled explicitly in SNP. Moreover, this calculation is more precise compared to the SNP result which provided production orders on a weekly granularity level (buckets). Finally the available capacity considers whether overtime shifts on Saturdays are allowed or not depending on the proposal of SNP and the outcome of the negotiations with the workforce representatives.

Based on the pegging link between each production order and its demand each production order has an exact due date. As fruit juices have a limited shelf life, this maximum time constraint has to be met on the pegging link. Shelf life constraints also apply to intermediate products that are stored in tanks (see [Figure 6.9](#)). However, as neither the mixing process nor the storage capacity in intermediate storage tanks does constitute a bottleneck in any of Frutado's plants, production in the mixers follows the requirements of the filling lines. Issues with respect to shelf life of intermediate products only arise on short notice, if a filling line breaks down. As these situations usually only affect the frozen period (first week) and require immediate reaction, this is dealt with manually on the spot and is not modeled in the basic Frutado model.

Recipes of Frutado's products are composed of a bill of material that lists all the ingredients required to create the product as well as a machine routing that specifies which resources (raw material silos, mixers, intermediate storage tanks, and filling lines) can be used. Not all theoretically possible routings can be used in practice, because the pipelines connecting the raw material silos, mixers, intermediate storage tanks, and filling lines do not have full connectivity. Nonetheless, the experience of the workforce using the right silos, mixers, and tanks and the absence of bottlenecks other than on the filling lines yield a very simple recipe structure with respect to the routing, i.e. the only planning relevant resource in PP/DS is the filling line. The frozen period of one week and favorable contracts with suppliers that allow replenishment of raw materials with short lead times make the planning of raw materials needless at this stage. However, generally speaking the consideration of material availability is equally important as the consideration of resource capacity, because production of any product cannot commence if either required materials or required resources are missing.

6.4.2 Extensions

The complexity of the production process determines to a large extent, how treatable the corresponding planning problem is for a human planner and/or a supporting software system. While a limited number of bottlenecks and resources as well as a limited number of input materials that have to be considered gives the planner the chance to understand and accept the solution provided by a software system, this task becomes more demanding the more

complex the production process is. As explained above, recipes in the basic Frutado model (represented by production process models (PPMs) in the case study, see also Section 3.3.2) are fairly simple, but the real production process is somewhat more complex (see [Figure 6.9](#)). However, as long as the relevant constraints of the production system are modeled, there is no necessity to plan at this level of detail. Nevertheless, the Frutado company may face three requirements in the future that would require changes in the structure of the PPMs.

First, there is the check of material availability. This check is currently done in the ERP system, because it was not considered relevant for production scheduling in the past. This was due to very favorable contracts with suppliers that allowed replenishment of raw materials with very short lead times. However, the management of Frutado has realized as a result of another project that substantial cost savings can be achieved by following another stock strategy and reducing inventories as well as by negotiating longer lead times to allow suppliers more flexibility to be able to reduce the purchase price of the raw materials. Consequently, material availability is no longer a given and actual available inventories, planned inflows (from already existing purchase orders) and new purchase requisitions considering planned delivery times have to be considered in the future. From a modeling perspective, this means that those relevant input materials have to be part of the recipe, i.e. PPM.

Second, an engineering project has analyzed the feasibility of upgrading the filling lines. The filling lines being the bottleneck in Frutado's production process for years have been in the focus of several projects before and new advances in filling technology have opened up the chance to increase filling speed significantly. However, before taking such an investment Frutado's management wants to assess that the expected gains in filling capacity will be realized and not distorted by the emergence of new bottlenecks on the mixing stage or in one of the warehousing stages (raw materials or intermediate products). Thus, also resources that have not been identified as bottlenecks before, but are potential bottlenecks need to be modeled to assess such a scenario. While raw material silos and intermediate storage tanks are fairly large and show rather modest utilization rates and their size can be adapted with rather low effort compared to the mixers, only those should be modeled explicitly. This second scenario requires the definition of a second operation for mixing (in addition to the filling operation) in the PPM. Similar to filling not all mixers are capable of producing all products which is also reflected in the definition of the PPM.

Third, the network of pipelines connecting the raw material silos, mixers, intermediate storage tanks, and filling lines in the three plants requires a major overhaul. Considering the significant cost difference between a pipeline system that connects all resources from one stage with each resource of the subsequent stage and a pipeline system with limited connectivity, an engineering company has offered a solution with limited connectivity

to Frutado. As Frutado's management is eager to learn, whether these limitations in connectivity have an impact on the output of any of the plants, they decide to have this connectivity constraint explicitly modeled in their planning system. While this constraint can in principle be modeled using PPMs, a more elegant solution is to make use of resource networks. Resource networks are a master data object in SAP® APO that allows to model the material flow between resources (see Section 3.3.2 for details).

In summary, the complexity of the production process is expressed in SAP® APO to a large extent in the form of PPMs. The PPM in the Frutado case (basic and extended) stores the information

- which operations have to be executed to produce a product,
- how these operations are related by precedence relationships,
- which resources are used in these operations,
- how much these resources are used in an operation (production coefficient),
- whether alternative modes (resources) are allowed for an operation,
- what the priority of the different modes are,
- what the setup key for an operation is,
- which input materials are required for the production of a product,
- which quantities of these input materials are required, and
- which quantities of output materials result.

Questions and Exercises

1. Explain the effect of the two different objectives dealing with delay in the basic Frutado model. Can you think of an use case in which penalizing the most delayed order is counterproductive?
2. Is it required to extend the basic Frutado model in order to deal with a huge additional customer order? Why or why not? What kind of data is changing in this situation?
3. Is it required to extend the basic Frutado model in order to deal with a temporarily downtime of one of the filling lines? Why or why not? What kind of data is changing in this situation?

6.5 Implementation and Results

As the representation of data is considerably different in SNP and PP/DS (see also Section 3.1), the integration between these two modules is crucial for a successful implementation of SAP® APO. The concept of PP/DS is founded on an order based representation of data, whereas in SNP all information within a bucket is consolidated into one aggregate figure. For SNP this means in the Frutado case that the complete weekly demand per product shows up as one number per week in the DC, is transported as one consolidated shipment per week from the plant to the DC and is produced as one production order per week in the DC (see upper part of [Figure 6.10](#)).

However, the weekly demand / customer orders are spread over the week. Using this untreated information would require in PP/DS to have everything ready at the end of the preceding week, thus losing one week of the already short product shelf-life due to insufficient planning methods. Therefore, the following procedure is followed according to the planning concept: The weekly demand forecast is replaced by a daily demand forecast (including already received customer orders) in the DCs. Then the weekly transportation orders between the plants and DCs are split into daily transportation orders according to the daily demand figures. Finally, PP/DS planned orders are created upon the transportation requirements in the plants by following a lot-for-lot strategy (see lower part of [Figure 6.10](#)).

To achieve this logic the following steps have to be performed in SAP® APO:

1. Create safety stock requirements for PP/DS planning.
2. Replace weekly DP forecast by daily DP forecast.
3. Convert weekly SNP stock transport orders into weekly PP/DS transport orders.
4. Fix pegging relationships between weekly PP/DS transport orders and daily DP forecast.
5. Split PP/DS transport orders according to pegging relationship.
6. Create PP/DS production orders using e.g. a lot-for-lot strategy.

The first bullet is not really part of this logic, but rather a technical requirement and mentioned only for the sake of completeness to describe the Frutado planning scenario.

If SNP has created more transport quantities than there is demand in the planning horizon of PP/DS, then PP/DS planned orders have been created which do not have a pegged requirements element in the planning horizon. These are those orders that SNP has anticipated production to build up e.g. seasonal stocks. In PP/DS optimization planned orders without pegged requirement elements are considered to be due at the end of the planning

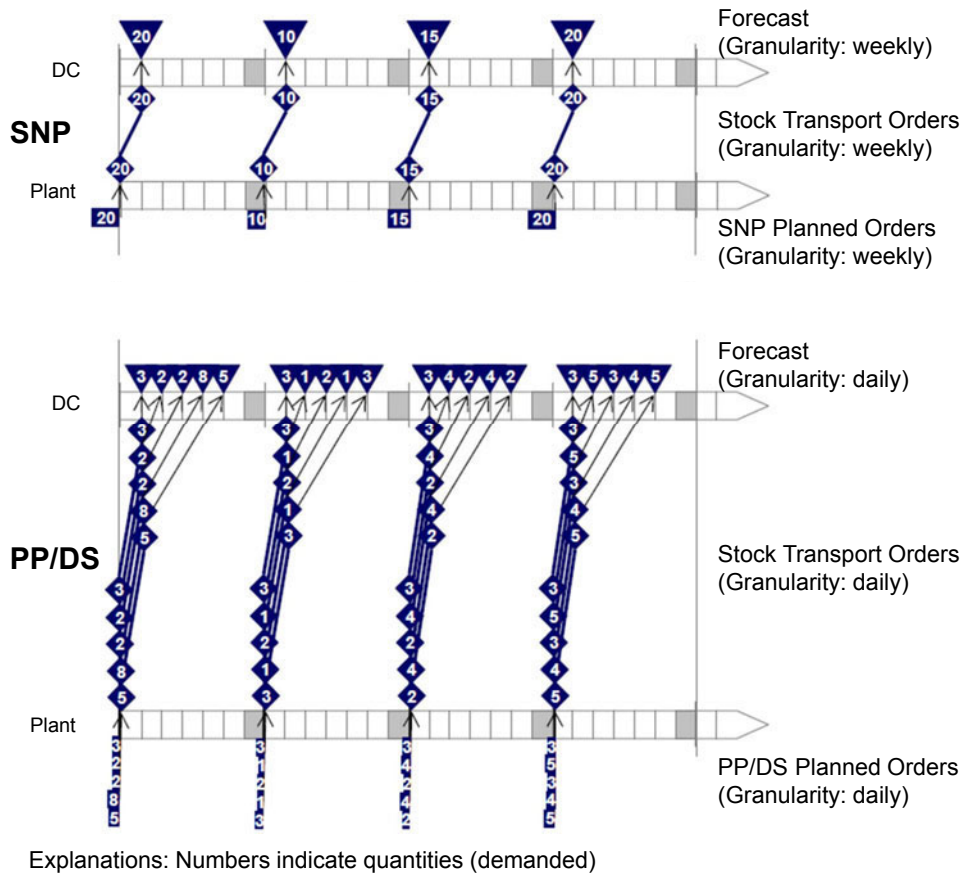


Figure 6.10
SNP - PP/DS
integration

horizon which is compatible with the requirement in the Frutado case. At last, optimization can conclude the PP/DS planning run.

The feedback integration from PP/DS to SNP is a manual process. For each week and each filling line, the responsible planner measures actual setup times. These figures are used to adapt the global setup reduction used in SNP, but also to verify the accuracy of the setup matrix that has been used for planning in PP/DS.

6.6 PP/DS Learning Units

6.6.1 Overview

The PP/DS learning units¹ are divided into five areas:

- PP/DS master data
- Interactive production planning tools
- Integration of DP data into PP/DS

¹ We are indebted to our colleague, Hans-Otto Guenther, Dept. of Production Management, TU Berlin, and his staff for preparing the datango learning units for the PP/DS module.

- PP/DS production planning
- In-depth stream

The learning units in the area of PP/DS master data explain the most important master data objects required for PP/DS in the Frutado company, i.e. production process models and the use of setup matrices to model sequence dependent setups. The product view, product planning table, and detailed scheduling planning board are different interactive production planning tools that can be used in PP/DS to review the demand situation, to analyze the capacity utilization and to visualize production schedules. These different user transactions are explained in the learning units on interactive production planning tools.

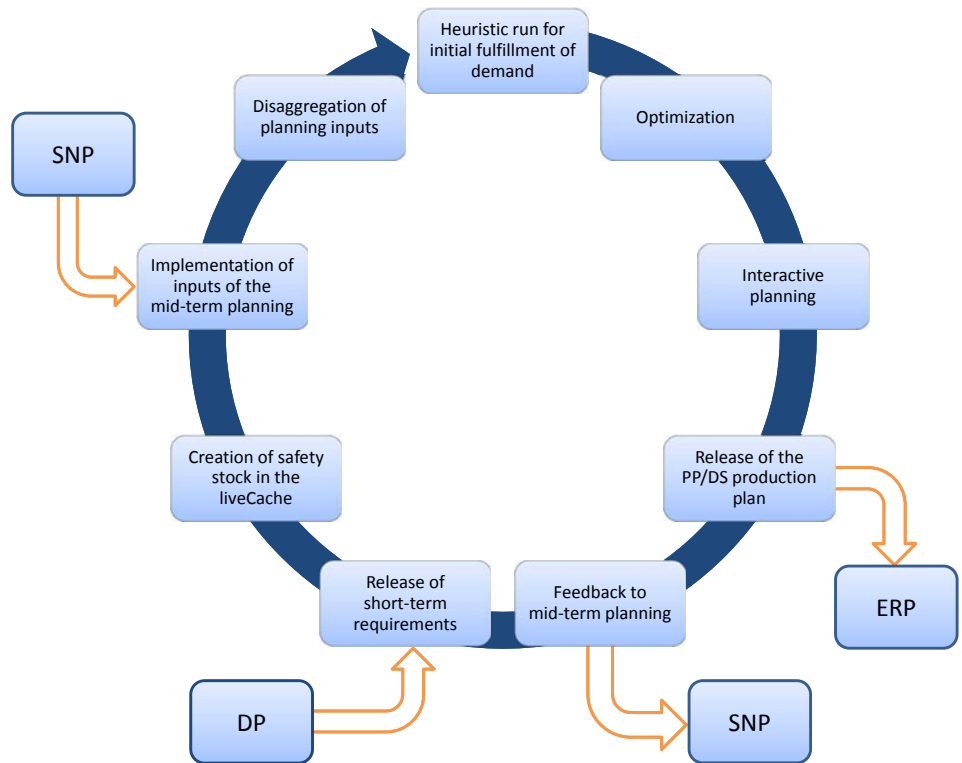


Figure 6.11
PP/DS planning cycle

The learning units in the area of integration of DP data into PP/DS and PP/DS production planning complete the PP/DS planning cycle in Figure 6.11. Anytime a PP/DS planning cycle is started, the requirement data (demands) needs to be arranged. This includes three different sets of data. First, short-term (daily) requirements from DP need to be released. Second, safety stock requirements need to be generated. This is necessary, because PP/DS relies on LiveCache (LC) data for planning, while the safety stock (method and amount per location-product) is part of the master data. Therefore, LC requirements are created in a separate step. Finally, the SNP planning results (stock replenishment orders for the three DCs of the Frutado

company), which have been obtained on a weekly granularity need to be disaggregated into daily requirements (not part of this learning unit). The conversion of orders from SNP to PP/DS has already been shown in the SNP learning units (see Section 5.6.2).

Furthermore, the capacity availability situation is also an input of the mid-term planning from SNP that needs to be implemented in PP/DS. While SNP allowed to use up to 18 shifts per week (including overtime capacity) compared to 15 shifts per week (standard capacity) the planned usage in the mid-term horizon needs to be set as a constraint in the short-term PP/DS planning horizon. One way of adapting resource availability in PP/DS is shown in the in-depth stream learning unit that shows reactions on a machine breakdown.

Now that both the demand situation and resource availability situation are correctly reflected in the SAP[®] APO system, production planning can begin. The learning unit PP/DS production planning shows the different steps that are usually required in a short-term planning scenario. These are

- creation of planned orders based on demands,
- scheduling and sequencing of planned orders (optimization),
- interactive evaluation of the planning result.

While in practice the first two steps are often run in background, these are executed here explicitly. First, a heuristic is called to generate planned orders for the Frutado products (juices and ice teas). The heuristic generates the planned orders such that they fulfill their demands just-in-time. This will obviously overload resources, because all demands for a day have the same requirement time after disaggregation of orders from SNP. Thus, the optimizer is used to generate a capacity feasible production plan that also reflects a good setup sequence (based on the setup matrix which has been defined in the learning unit on master data). Finally, the result of automatic planning using the optimizer is evaluated and can be adapted interactively in the detailed scheduling planning board.

The last learning units (in-depth stream) are devoted to situations that require manual adjustments of the plan, because of unexpected events. The first event, a machine breakdown, requires plan changes, because the expected resource availability is less than anticipated. The second event, a huge order from a new customer that needs to be matched, requires plan changes, because demand is higher than anticipated. Next, we will briefly describe the content of the different learning units.

6.6.2 Basic Stream

PP/DS Master Data

Consistent and complete master data is very important in short-term production planning. Therefore master data required by PP/DS needs to be

accurate. As sequence dependency of planned production orders is a crucial requirement in the Frutado case, the representation of the production process model (PPM) containing bill of material and routing information as well as the representation of setup characteristics is important and therefore presented in the two learning units.

In the first learning unit, the concept of setup groups, setup keys, and setup matrices is explained. In the Frutado case, a setup is required whenever the flavor of the beverage is changed on a production line. If all 19 products of Frutado could be produced on one production line, then 361 different change-overs would potentially need to be defined. However, each production line is only capable of producing a subset of the products (see Chap. 1). Each product is associated with one setup key. As certain products (juices and ice teas) can be grouped, because they have similar setup/cleaning requirements, not all possible change-overs have to be defined for each production line. In this case, the setup can be defined by using the setup group in case of the more specific setup key.

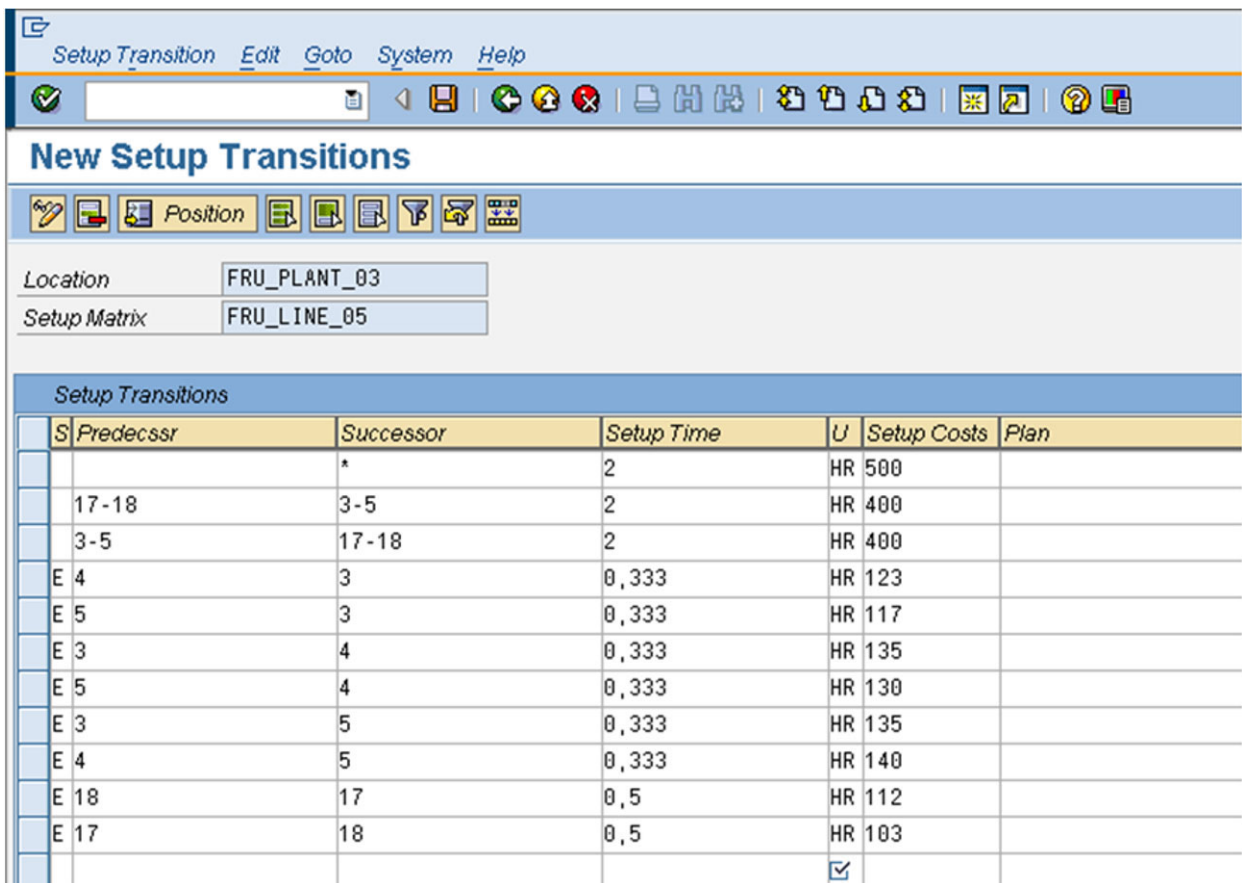


Figure 6.12
Frutado setup matrix
(Line 5)
© Copyright 2011. SAP AG.
All rights reserved

Figure 6.12 shows the setup matrix for production line 5. The setup matrix exhibits the possible changeover defined by predecessor and successor activity as well as the corresponding setup times and setup costs. In this

example, the setup matrix is based on two setup groups (3-5 and 17-18) with exceptions (indicated by an ‘E’ in the left-most column) for certain setup keys. For products 3 to 5 (setup group 3-5 with setup keys 3, 4, and 5), which are the three A-products that run on production line 5, specialized setup kits are available to allow for fast setups between these products. Therefore the corresponding setup times and setup costs are lower within this setup group than for changing to this setup group.

The second learning unit describes the creation of a PP/DS PPM. PPMs have already been used in SNP (see Section 5.6). However, due to the way of planning in SNP (based on buckets), SNP-PPMs do not contain sequence dependent setups and allow only one mode per production activity. PP/DS-PPMs on the other side are much more detailed due to the planning scope at this stage. The learning unit shows the creation of a PPM for Product 3 (juice). The PPM consists of two activities (setup and production) which can be executed in two different modes (either on Filling Lines 5 or 6). The setup key ‘3’ which has been defined in the previous learning unit is assigned to the setup activity of the PPM to allow for sequence dependent setup times and costs (if setup activities were not sequence dependent, setup groups and matrices would not be required, and setup times and costs could be maintained directly in the PPM). In practice, the PPM would not be created manually in SAP® APO, but automatically by using the bill of material and routing information from ERP. The same holds true for SNP-PPMs which would be created automatically within SAP® APO based on the PP/DS-PPMs.

Interactive Production Planning Tools

While a lot of planning tasks can be automated in PP/DS, interactive production planning tools are required to bring user experience and decision making into play. Three different transactions that can be used for this purpose in PP/DS are presented in separate learning units:

- Product view,
- Product planning table,
- Detailed scheduling planning board.

The product view displays information on requirements and receipts for each location product separately. In [Figure 6.13](#) the situation prior to PP/DS planning is shown for Product 1 at Plant 1. One can see, that each requirement element (category PReqRel) representing a weekly stock replenishment order from the DC is matched exactly by one receipt element (category SNP:PL-ORD) created by SNP. Separate tabs show this situation graphically (“Quantities”-tab) or display e.g. the pegging situation (“Pegging Overview”-tab). The product view is not only used to display data, but also to create planned orders manually and automatically. However, this will be shown in a subsequent learning unit in a different transaction.

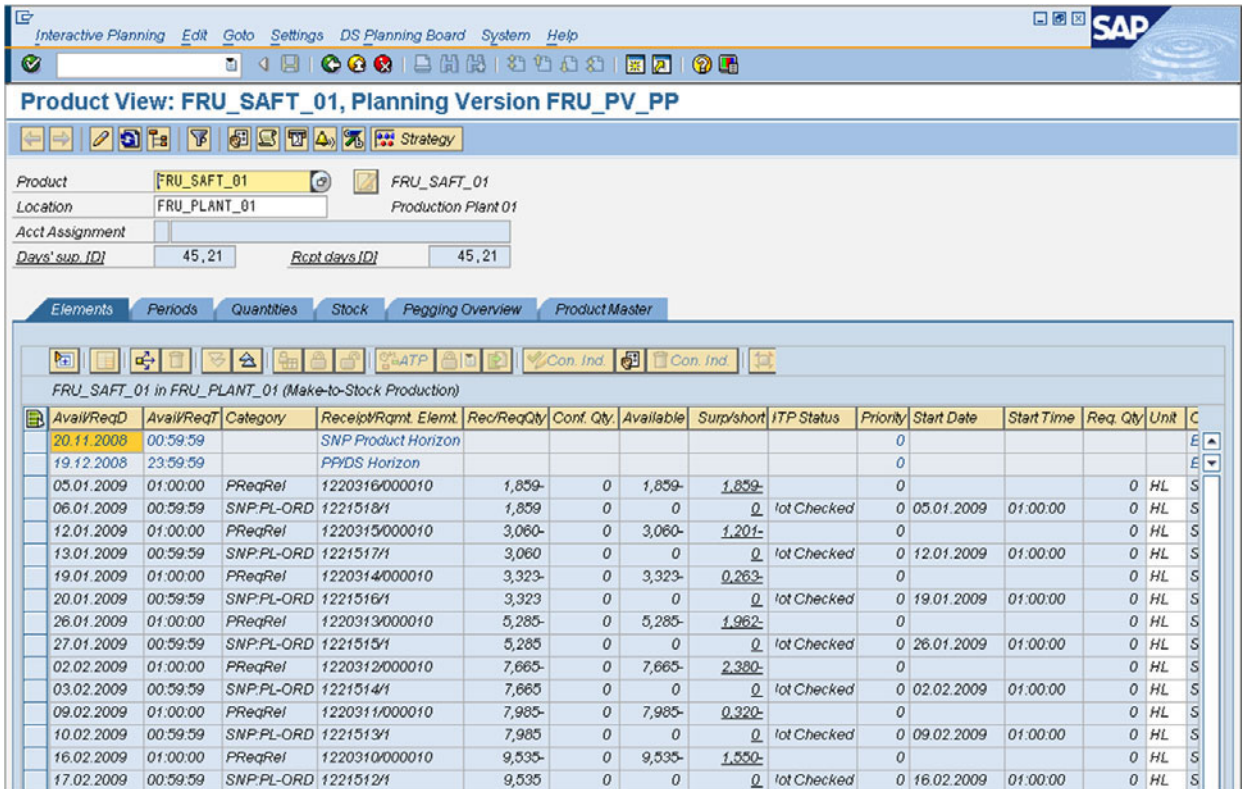


Figure 6.13

Product view

© Copyright 2011. SAP AG.

All rights reserved

The product planning table is a multifunctional and flexible interactive planning tool, which provides different views on the data. It can be used to analyze the results generated by automated planning tools like the PP/DS optimizer or PP/DS heuristics in a table format. Alerts are displayed for resource overloads or unmet demands. Like in the product view, the product planning table is not used in display mode, but heuristics and the optimizer can be started from here directly in order to modify the production plan. Of course, manual adaptations of the plan are also possible.

The detailed scheduling planning board is used for the graphical representation of the planning situation. The scope of planning is defined by resources which have to be selected and are displayed in a gantt like view. Figure 6.14 shows the detailed scheduling planning board for Plant 3 (Production Lines 5 and 6). The upper part of the screenshot shows the schedule for Production Lines 5 and 6. While Production Line 6 is fully occupied for the entire displayed interval, and only one short setup activity is visible, Production Line 5 is utilized less and shows more setups (setup activities are visualized with more narrow rectangles compared to production activities).

The product chart of the detailed scheduling planning board shows the planned orders product-wise compared to the resources chart. All products that can be produced on the selected resources are displayed. Also pegging links (see Section 3.3.3) can be displayed here which is helpful, if the bill of

material modeled in SAP® APO contains more stages than in the Frutado case. Finally, the product stock for each location product can be displayed graphically to detect inventory peaks and/or unmet demands. Various scheduling functions and heuristics are available in the detailed scheduling planning board. This is the place, the human planner can use to merge his expert knowledge gained e.g. from experience, that has not been made available to the system for any reason, to the plan.

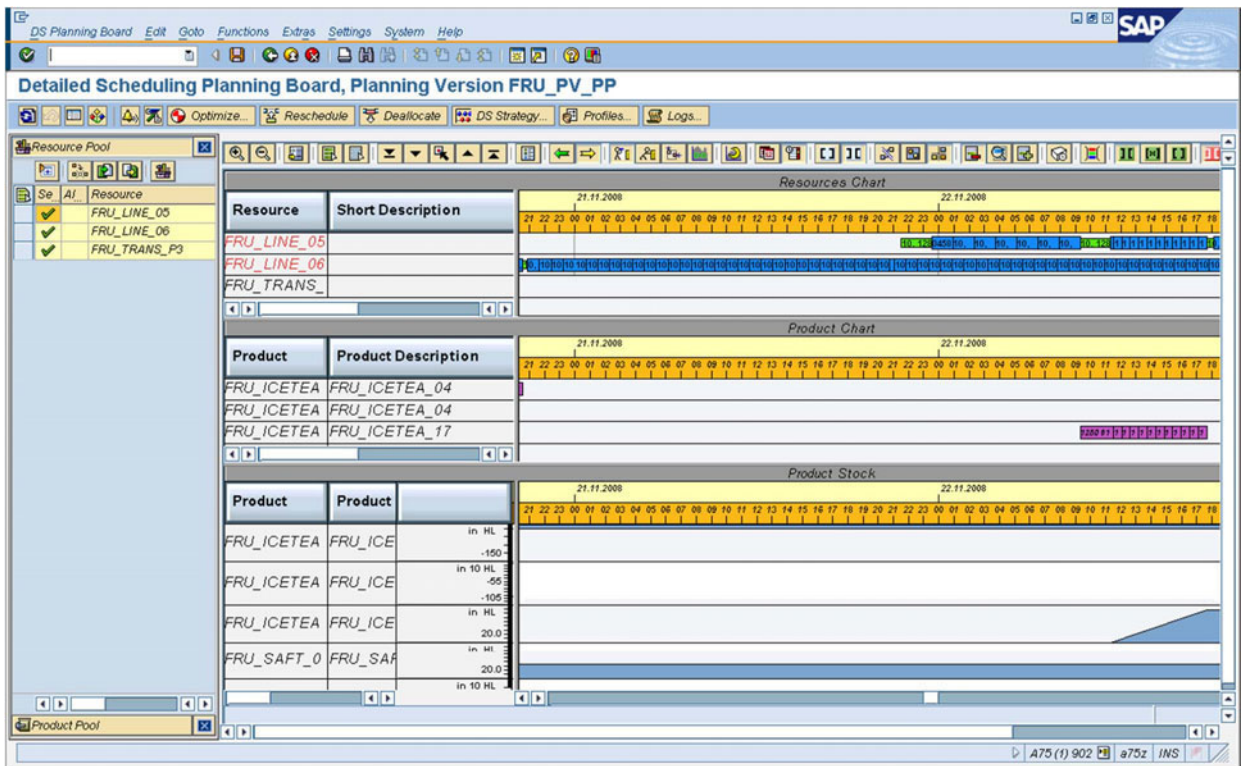


Figure 6.14
Detailed scheduling
planning board
© Copyright 2011. SAP AG.
All rights reserved

Integration of DP Data into PP/DS

As shown in Figure 6.11 and explained above (see Section 6.6.1) PP/DS requires more precise demand data than the preceding planning levels to fulfill its task. These are

- daily demand figures from DP,
- disaggregated stock replenishment orders from SNP (to account for (seasonal) stock built-up), and
- safety stock requirements.

The learning unit “Integration of DP Forecast Data” describes how the first elements are being created, while the second elements have already been created in an SNP learning unit (see Section 5.6). Safety stock requirements

are created using a report in background. Thus, no learning unit is provided for this task. To be able to separate SNP from PP/DS data, different planning versions are used by Frutado. How to create a planning version (for PP/DS) based on an existing one (for SNP) is shown in a distinct learning unit.

PP/DS Production Planning

PP/DS Production Planning contains four learning units. In the beginning of a PP/DS planning run receipt elements (planned orders) need to be created for all requirement elements (demands). Thus, this is the topic of the first learning unit. Planned orders can be created in PP/DS using various options, some of them have already been mentioned, e.g. creating planned orders in the product view or the product planning table. In practice, especially when processing huge amounts of products and locations, planned orders are created mostly in the background automatically using a report. How such a planning run is defined is shown in this learning unit. Nonetheless, the planning run is started immediately in the learning unit and not scheduled regularly (e.g. during the night) which would be the case in practice.

Defining *planning runs* is a powerful tool. While the planning run defined in this learning unit consists only of one step (creation of planned orders), planning runs can generally include several steps building on each other. A typical planning run in a scenario with several production stages could look like this:

- Step 1 - Stage-numbering algorithm
- Step 2 - Product planning (components according to low-level code)
- Step 3 - Fix pegging relationships
- Step 4 - Change order priorities
- Step 5 - Optimization

In the first step, the low-level code for each location product is determined for the selected products. The second step creates secondary demands of items followed by lot-sizing resulting in planned orders for the selected products based on their low-level code. This is important because otherwise components would be planned before their successor products are planned creating an inconsistent situation with lots of unmet requirement elements. After all planned orders have been created for raw materials, semi-finished products, and finished products, these are linked via pegging relationships that have been created dynamically (see Section 3.3.3). As different demands (customer orders) may have different priorities, these priorities shall be inherited to preceding product levels (from finished products to semi-finished products to raw materials). As the assignment of planned orders to preceding orders changes dynamically based on the current planning situation (e.g. using a FIFO-principle) the assignment shall be fixed to reserve planned orders

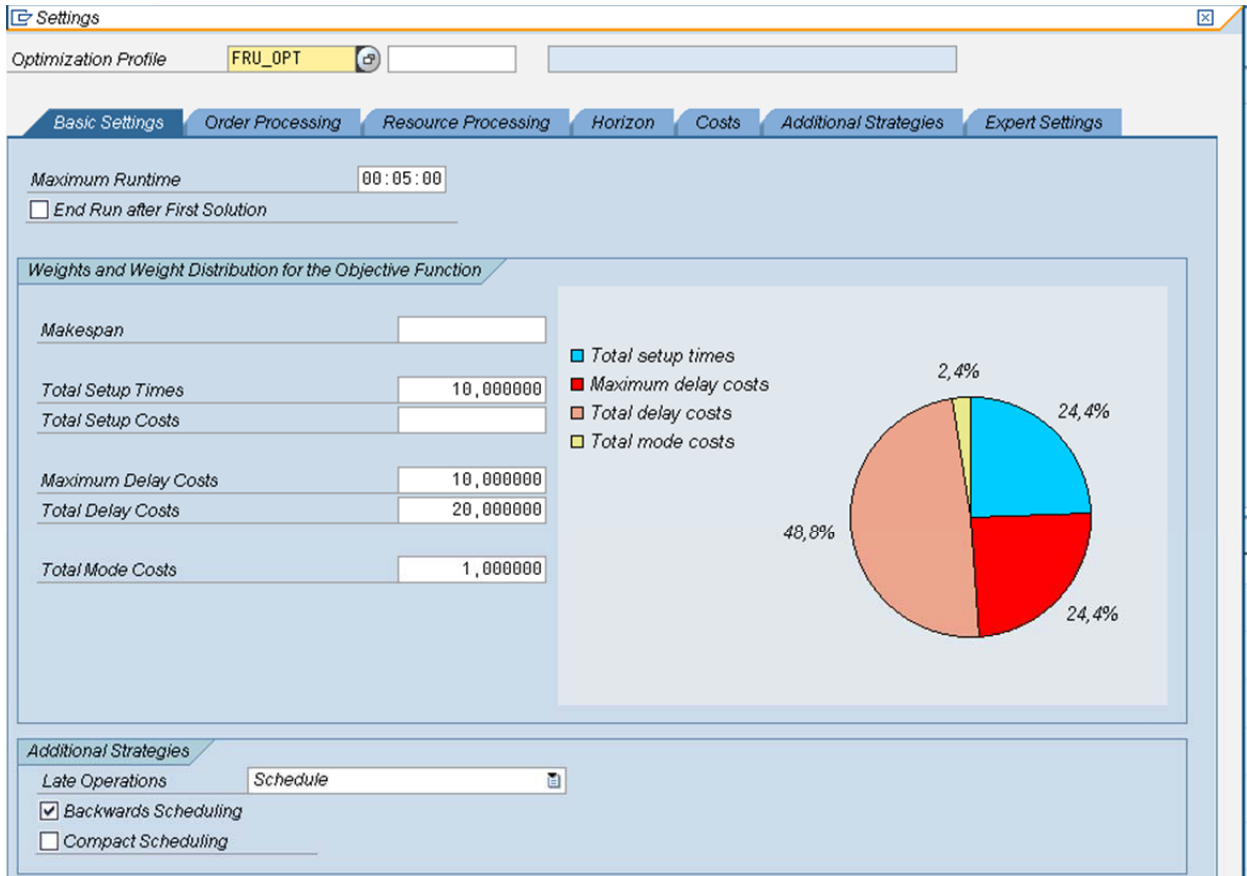


Figure 6.15
PP/DS optimization profile
© Copyright 2011. SAP AG.
All rights reserved

for semi-finished and finished products for the important customer orders and to avoid their consumption by less important tasks. This is achieved in the third step, while the fourth planning step inherits the customer order priorities through the network of the fixed pegging relationships up to the raw material stage. The final step concluding this sample planning run is the optimization of the previously created orders considering order priorities on all production stages.

This example planning run definition has shown that within a planning run production planning heuristics creating planned orders (step 2) can be combined with scheduling heuristics for scheduling and sequencing the planned orders (step 5) as well as so-called “service heuristics” (steps 1, 3, and 4).

At the end of the learning unit, planned orders have been created reflecting the demand situation. The next step is the interactive execution of the PP/DS optimizer from the product planning table, which is the topic of learning unit “Optimization Run”. The objective is to find a feasible plan (considering resource availability) with a favorable setup sequence of the previously planned orders based on a reasonable assignment of planned orders to production lines.

Figure 6.15 shows the basic settings of the PP/DS optimization profile. The PP/DS optimization profile governs which objectives are pursued during optimization, and how these conflicting objectives outweigh each other. The PP/DS optimizer strives to minimize a weighted objective function with the following determining factors:

- Makespan
- Setup times
- Setup costs
- Maximum delay
- Total delay
- Mode costs

The determining factors relevant in the Frutado case are setup times, maximum and total delay, as well as mode costs. Therefore these factors have a positive weighting in the PP/DS optimization profile for Plant 3 which is shown here as an example. However, these objectives are conflicting. Maximum and total delay receive the highest weights, because it is most important for the Frutado company to satisfy all demands on-time. Furthermore, the sequence of planned orders is important, because change-overs between different products cause setup times and setup costs, which should be avoided, as they cause unproductive time on the scarce resources. Finally, some products can be produced on both lines in Plant 3. To express a preference for scheduling certain products on certain production lines, mode costs have been defined and are therefore weighed in the objective function.

The PP/DS optimizer is based on a GA as described exemplary in Section 6.2. During its runtime, the PP/DS optimizer sends status messages that inform the user about the phase the algorithm is currently working in. In the improvement phase of the GA, the currently best result is kept in memory, and its objective function value is plotted for the user as shown in Figure 6.16. Different objectives are shown in different colors for the user to identify which objectives have been sacrificed for which purpose. Only those determining factors with positive weights are shown according to their weights. In the screenshot, a solution has been found with a maximum and total delay of zero which means that all demands can be satisfied in time. Therefore, the remaining objective is to minimize setup time and mode costs. Note, that although the mode costs look rather high, this may be the best possible (optimal) value. Mode costs of zero may not be possible, because even the selection of the most preferable mode, still incurs (substantial) mode costs.

The third learning unit in this area shows the results of the optimization run in the product planning table (see Fig. 6.17). Production line 6 is fully or almost fully utilized for most days in the displayed period, while

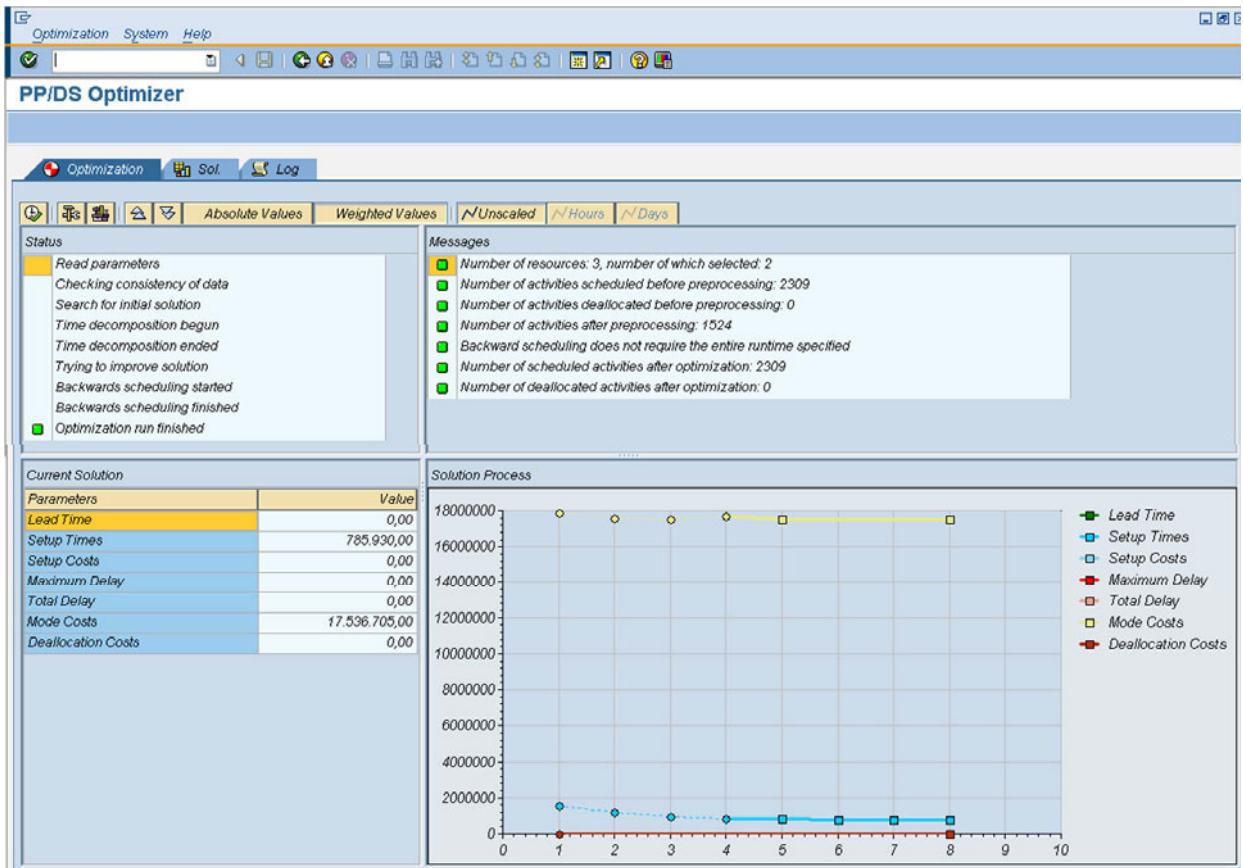


Figure 6.16
PP/DS optimization
run

© Copyright 2011. SAP AG.
All rights reserved

the utilization of Production Line 5 varies between zero and almost full utilization. Most likely this is due to mode costs, because Production Line 6 being the newest production line requiring the least amount of labor is the cheapest production line available in the Frutado company (see Chap. 1). As no resource is overloaded and all demands are fulfilled, no alert is shown.

The last learning unit “Interactive Planning and Evaluation” shows how manual plan modifications are done in the detailed scheduling planning board using drag and drop functionality. This is the most intuitive way a planner can follow to alter automatically created planning results interactively. Planned orders are inserted at their new position, and corresponding setup activities (at the beginning of the inserted orders as well as at the succeeding planned orders) are adjusted according to the entries in the setup matrix.

While automated production planning and the interactive analysis and modification of plans are the main goals in PP/DS, reporting on performance figures is frequently requested. Therefore, this learning unit concludes with some means of showing how performance figures like the resource utilization, work-in-process inventories as well as information on certain orders or operations can be retrieved with different levels of granularity in PP/DS.

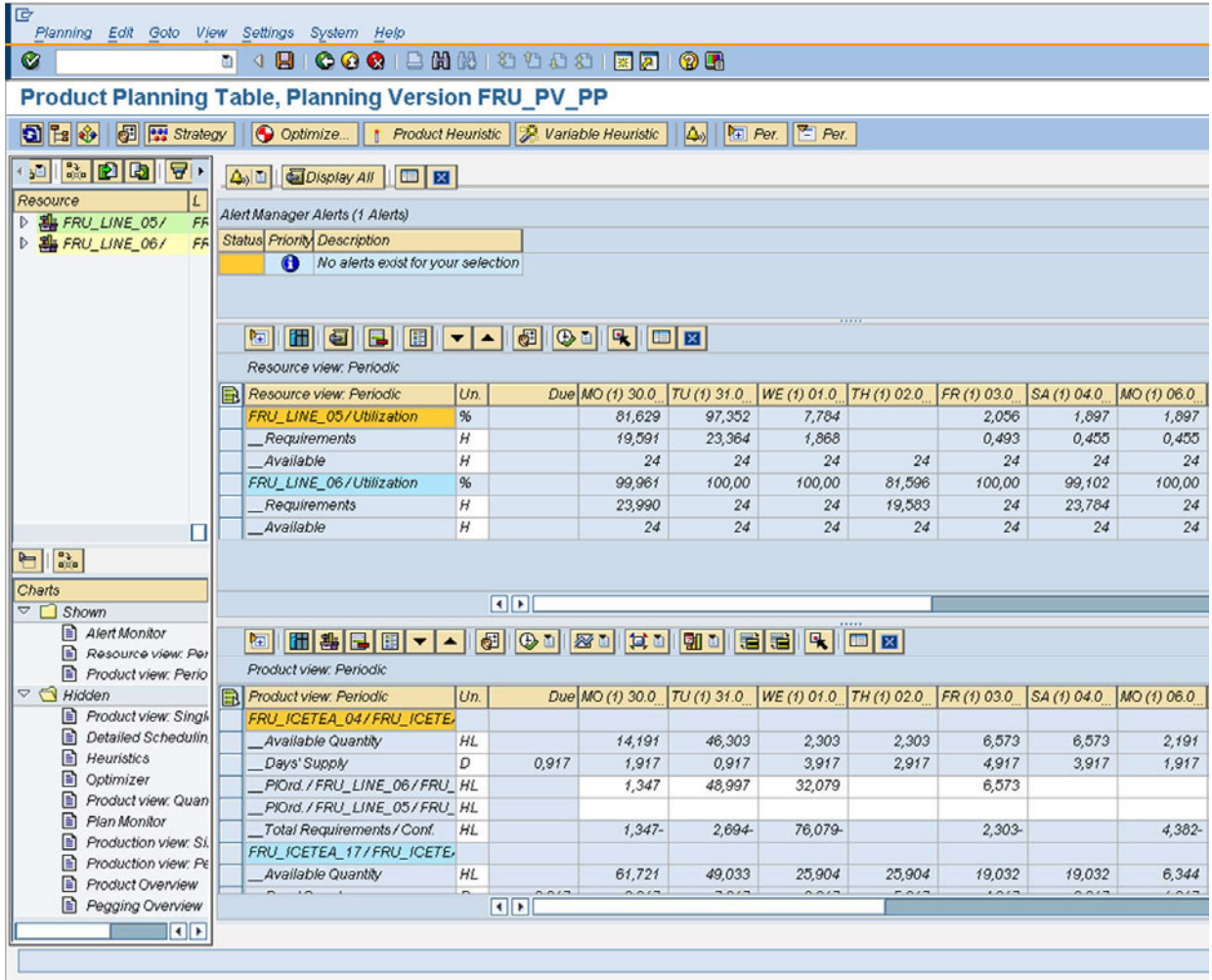


Figure 6.17
 PP/DS optimization
 result
 © Copyright 2011. SAP AG.
 All rights reserved

6.6.3 In-Depth Stream

The learning units of the in-depth stream deal with special planning situations that are characterized by significant short-term deviations between (customer) demand and (capacity) supply. Both situations require (manual) adjustments of the plan. The first case deals with a resource breakdown. Rescheduling is required to allocate already planned orders to the reduced resource availability. The second case is vice versa. Here demand for resource capacity is increased due to an incoming sales order that needs to be met with short notice.

Machine Breakdown

Machine breakdowns are unplanned events. While regular resource maintenance can be planned and anticipated by a reduction of resource availability well in advance, this is not the case here. All of a sudden, a resource becomes unavailable, and the production orders scheduled on the resource need to be

rescheduled to other resources or postponed immediately. Prior to doing this in the planning system, the planner has to find out the expected duration of resource unavailability. Then this information needs to be represented in the planning system. In the learning unit, it is first shown how the resource availability is reduced for the expected resource downtime (1.5 days). As many planned orders are pertained by this downtime, manual rescheduling is not an option, because too many factors (order delay, sequence, and mode selection) would need to be considered by the planner. Therefore, the PP/DS optimizer is started interactively to reschedule the production orders. The resultant plan shows to a great extent that orders are postponed to after the downtime, since the downtime happened on the cheapest Production Line 6 in Plant 3. As this situation occurs in an off-peak season, it is possible to postpone orders rather than shifting them to Production Line 5 which would have been the more costly alternative.

Short-Term Requirement

The storyline of the second learning unit “Short-Term Requirement” is somewhat different. Although the Frutado company keeps safety stocks of all products, the sales representatives are overwhelmed by a large sales order from a newly acquired customer. However, this order needs to be met on short notice. The decision has been made to fulfill this customer order although this may mean postponing other production orders and potentially upsetting other long-term customers. This is a management decision that needs to be taken in advance (or after assessing the implications of changes in plan with the help of PP/DS). In this learning unit the assumption has been made that the decision to fulfill this additional demand has been made. Therefore, the additional production order is created and assigned a high priority compared to the other planned orders present in the system. Rescheduling using the PP/DS optimizer positions the order with high priority such that the additional demand can be met. Consequently other orders are delayed.

Questions and Exercises

1. A setup matrix can be defined by setup groups and setup keys in PP/DS. Why are these two objects used? Prepare a numerical example to show the advantage of this concept.
2. Which interactive planning tools are available in PP/DS? What is their difference? Explain possible use cases.
3. Which objectives are pursued by the PP/DS optimizer? Are these objectives complementary or conflicting? Why? Explain this by means of two examples.
4. Something is wrong with one of your production lines. It breaks down sporadically (several times per week), but can be repaired relatively

quickly (within minutes/hours). What are your options to handle this situation? Think about a short-term and a mid-term solution.

Bibliography

- Baumann, P.; Trautmann, N. (2010) *An MILP approach to short-term scheduling of an industrial make-and-pack production facility with batch splitting and quality release times*, in: *Industrial Engineering and Engineering Management (IEEM), 2010 IEEE International Conference on*, 1230–1234
- Boysen, N.; Fliedner, M.; Scholl, A. (2009) *Production planning of mixed-model assembly lines: overview and extensions*, *Production Planning & Control*, vol. 20, 455–471
- Dickersbach, J. (2009) *Supply Chain Management with APO*, Springer, Heidelberg
- Drexl, A. (1990) *Fließbandaustaktung, Maschinenbelegung und Kapazitätsplanung in Netzwerken, ein integrierender Ansatz*, *Zeitschrift für Betriebswirtschaft*, , no. 60, 53–70
- Drexl, A.; Fleischmann, B.; Günther, H.-O.; Stadtler, H.; Tempelmeier, H. (1994) *Konzeptionelle Grundlagen kapazitätsorientierter PPS-Systeme*, *Zeitschrift für betriebswirtschaftliche Forschung*, vol. 46, 1022–1045
- Hartmann, S.; Kolisch, R. (2000) *Experimental evaluation of state-of-the-art heuristics for the resource-constrained project problem.*, *European Journal of Operational Research*, vol. 127, no. 2, 394 – 407
- Holland, J. H. (1975) *Adaptation in natural and artificial intelligence*, University of Michigan Press, Ann Arbor, Michigan
- Klein, R. (2000) *Scheduling of resource-constrained projects*, *Operations research/computer science interfaces series*, vol. ORCS 10, Kluwer Acad. Publ., Boston
- Klein, R. (2008) *Genetic algorithms*, in: H. Stadtler; C. Kilger (Eds.) *Supply Chain Management and Advanced Planning*, chap. 30, Springer, Berlin et al., 4th ed., 529–536
- Levner, E.; Kats, V.; de Pablo, D. A. L.; Cheng, T. (2010) *Complexity of cyclic scheduling problems: A state-of-the-art survey*, *Computers & Industrial Engineering*, vol. 59, no. 2, 352 – 361
- Mayr, M. (1996) *Hierarchische Produktionsplanung mit zyklischen Auflagemustern: Univ., Diss.–Augsburg, 1996., Theorie und Forschung Wirtschaftswissenschaften*, vol. 441;41, Roderer, Regensburg

- Meyr, H. (1999) *Simultane Losgrößen- und Reihenfolgeplanung für kontinuierliche Produktionslinien: Modelle und Methoden im Rahmen des Supply Chain Management; Produktion und Logistik*, Deutscher Universitäts-Verlag GmbH
- Neumann, K.; Schwindt, C.; Trautmann, N. (2002) *Advanced production scheduling for batch plants in process industries*, OR Spectrum, vol. 24, no. 3, 251–279
- Pritsker, A. A. B.; Waiters, L. J.; Wolfe, P. M. (1969) *Multiproject scheduling with limited resources: A zero-one programming approach*, Management Science, vol. 16, no. 1, 93–108
- Reeves, C. R.; Rowe, J. E. (2003) *Genetic algorithms: Principles and perspectives: A guide to GA*, Kluwer Academic, Boston
- Sahling, F. (2010) *Mehrstufige Losgrößenplanung bei Kapazitätsrestriktionen*, Gabler Research, Gabler Verlag / GWV Fachverlage GmbH Wiesbaden, Wiesbaden
- Scheckenbach, J. B. (2009) *Collaborative Planning in Detailed Scheduling*, Ph.D. thesis, Universität Hamburg, Von-Melle-Park 3, 20146 Hamburg
- Scholl, A. (1999) *Balancing and sequencing of assembly lines: With 75 tables: Techn. Hochsch., Diss.-Darmstadt, 1995.*, Contributions to management science, Physica-Verl., Heidelberg, 2nd ed.
- Silver, E. A.; Pyke, D. F.; Peterson, R. (1998) *Inventory Management and Production Planning and Scheduling*, Wiley, New York, 3rd ed.
- Stadtler, H. (2007) *How important is it to get the lot size right?*, Zeitschrift für Betriebswirtschaft, vol. 77, 407–416
- Stadtler, H. (2008) *Purchasing and material requirements planning*, in: H. Stadtler; C. Kilger (Eds.) *Supply Chain Management and Advanced Planning*, Springer, Berlin et al., 4th ed., 217–229
- Suerie, C. (2005) *Time Continuity in Discrete Time Models: New Approaches for Production Planning in Process Industries*, Springer, Berlin
- Voss, S.; Martello, S.; Osman, I. H.; Roucairol, C. (1999) *Preface*, in: S. Voss; S. Martello; I. H. Osman; C. Roucairol (Eds.) *Meta-heuristics. Advances and Trends in Local Search Paradigms for optimization*, Kluwer Academic, I–X

Global Available-to-Promise (global ATP)

Bernhard Fleischmann¹, Sebastian Geier²

¹ University of Augsburg, Department of Production & Logistics, Universitätsstraße 16, 86135 Augsburg, Germany

² University of Augsburg, Department of Production and Supply Chain Management, Universitätsstraße 16, 86135 Augsburg, Germany

The previously described modules concern planning activities for the supply chain without knowing actual customer orders. The following chapter rather deals with executing the fulfillment of known customer orders.

It is Frutado's policy to fulfill demands whenever possible. For the fulfillment of *customer orders* one should consider a common definition of logistics. Logistics is having the right thing, at the right place, at the right time. But what should one do, if one of the rights cannot be met? The module Global Available-to-Promise (global ATP) is the tool kit of SAP APO for the fulfillment of customer orders. Its purpose is to promise a hopefully reliable delivery date to the customer at the arrival of his order. The module also observes and controls the customer orders after they have entered the company's planning system.

In this section we first present planning models in literature for the short-term order fulfillment. We then give an overview over the SAP APO module *global ATP* and show the implementation for the Frutado company. In conclusion we present a description of the learning unit for this module.

7.1 ATP: Basics and Literature

The term *Available-to-Promise (ATP)* is not new. Schwendinger (1979) already explained the ATP-Quantity of a product for a certain period as

the surplus of the amount on hand plus the projected supply based on the Master production schedule minus the projected forecast in this period.

The incentive to consider the task of order fulfillment goes back to the practitioners rather than to the scientific community. First the software providers of Enterprise-Resource-Planning systems (ERP) or Advanced Planning Systems (APS) implemented algorithms for setting reliable dates and calculating quantities that can be promised to the customers. Scientific literature was rarely involved in the early discussion.

With the rise of the e-commerce in the late 90s and the coherent need for online setting of short and reliable due dates the interest in this topic increased.

Influence of the Decoupling Point on ATP

The process of fulfilling the customer's demand is often simply called demand fulfillment. More precisely, the task of demand fulfillment is to take care of the customer orders downstream of the decoupling point. This concept is explained in Chapter 2.3. Depending on the position of this point, the main problems of demand fulfillment differ. For a detailed discussion of the influence of the decoupling point on the process of demand fulfillment see Fleischmann and Meyr (2004).

In a completely order driven MTO supply chain, the main bottleneck of demand fulfillment is the available production capacity whereas the needed material is either on stock or is purchased according to the customer-specific needs. As we focus on material-constrained order promising methods, MTO is not considered in this chapter.

In case of a MTS decoupling point, the orders are fulfilled from stock so that customers expect short service times. The bottleneck for demand fulfillment therefore is the available quantity of finished goods.

For ATO, beside of the available quantities of components, the assembling capacity may be a potential bottleneck.

Calculation of the Available-to-Promise Quantities

We previously mentioned the *ATP-quantity* or just ATP, which we will specify now.

The quantities that are available for promising of new orders, are derived from the current inventory plus the future supply minus the quantities needed for the already committed orders. For MTS, the ATP concerns the availability of finished goods and for ATO the availability of all components that are used for assembling the finished product.

While the current inventory can be determined from inventory management, future supply must be calculated. Therefore information on released, already in progress, production or purchase orders from the production planning or replenishment modules is necessary. In addition, particularly for a

t	1	2	3	4
I_0	2			
$Supply(t)$	3	5	3	3
$Committed(t)$	2		5	
$I(t)$	3	8	6	9
$ATP(t)$	3	6	6	9

Table 7.1
Calculation of ATP
quantities for $t_0 = 1$

longer horizon, information on planned supply from the Master Planning module is used.

More precisely, $ATP(t)$, the quantity that can be promised for a new order with the future date t , is calculated as follows: With the notations

t_0	= today
T	= planning horizon
$Supply(s)$	= planned supply at day s ($s = t_0, \dots, T$)
$Committed(s)$	= quantity committed for day s ($s = t_0, \dots, T$)
I_0	= today's starting inventory,

the planned inventory at the end of day $t_1 \geq t_0$ is

$$I(t_1) = I_0 + \sum_{s=t_0}^{t_1} (Supply(s) - Committed(s)) \quad (7.1)$$

and the ATP quantity is

$$ATP(t) = \min\{I(t_1) : t \leq t_1 \leq T\}. \quad (7.2)$$

The ATP calculation procedure with a planning horizon of $T=4$ is illustrated in [Table 7.1](#).

Tasks of Demand Fulfillment

The main task *demand fulfillment* can be subdivided into the three tasks Order Promising, Demand Supply Matching and Shortage Planning, according to Fleischmann and Meyr (2004).

The first step in fulfilling a customer order is to confirm the newly arrived order, which is called *Order Promising*. This includes the decision whether to accept or to deny it. In case of acceptance, the confirmation must specify the quantity the customer receives and the delivery date. For MTS, this decision should be based on the ATP. If the customer is free to configure the product he orders, as in an ATO situation, the decision of acceptance depends also on

the technical feasibility. As the decisions in the order promising are mostly based on ATP, the term ATP is often used synonymous with order promising.

Performing the order promising on new customer orders is possible in different modes (see Pibernik 2005 and Ball et al. 2004). Online order promising and batch order promising are extremes. In case of online or *real-time order promising* incoming orders are processed instantly. So the promised date and quantity are calculated by the order entry system for every incoming order. Batch order promising collects incoming orders during a time interval (e.g. one hour or day) and generates promises for these new orders. While online order promising offers the faster response for the customer, the *batch order promising* can calculate dates and quantities with knowledge of a number of new orders. A compromise is a hybrid order promising. Here new orders are promised temporarily in real-time at order entry. Later these vague promises (e.g. week of the year) are refined during a batch order promising run (e.g. day of the week). This two phase procedure can lead to changes between the first promise and later promises.

As an order is not fulfilled by just promising it, the execution of the processes following the order promising (e.g. assembling in case of ATO and shipping) must be controlled. During this time the promised orders (i.e. the demand) have to be matched with the available supply to check if already promised orders can still meet their confirmed dates. This is the reason why this task is named as short-term *Demand Supply Matching*. The scope differs with the type of production. In an ATO situation the assembly orders have to be released and sequenced, whereas in a MTS situation only the shipping of the products has to be initiated. Finding the right source locations for the actual distribution, where sufficient stock is available, is a task of the deployment (see Chap. 8 and Fleischmann 2008).

Often the situation occurs that a company has short ATP quantities for some products or components. This could either happen at the order promising for an incoming order or later at the step of demand-supply-matching (e.g. because some projected supply, from the own production or from an external supplier, is delayed or cancelled). To find appropriate reactions to such a shortage case is the task of *Shortage Planning*. At order promising the new order can be refused or some alternatives (like late delivery or substitution) could be searched. During the demand-supply-matching a short material or product affects all already promised orders requesting the specific material or product. Measures to solve the shortage often lead to a repromising or in the worst case even to a cancellation of already accepted orders. Normally only a very small percentage of the orders needs to be affected, but the decision, which orders are concerned, is of great practical importance.

Real-Time Order Promising for MTS

As explained above, the demand fulfillment differs significantly between MTS and ATO. Since the beverages sold by the Frutado company are all made

to stock, the remainder of this section focuses on the specifics of a MTS situation.

MTS is also the prevalent situation considered in the literature on ATP. Recent publications offer a wide range of models for this task. Selected models will be presented in the following sections. As in the Frutado company incoming orders should be promised on-line, the focus is on the real-time order promising models.

Looking just on the availability of the desired product at the desired date may be not appropriate for the process of order promising. A more sophisticated process for order promising is called rules-based search (see Kilger and Meyr 2008, p. 241). This order promising system searches enough ATP for a new order according to a system of search-rules. This search is executed along dimensions which structure the ATP. These dimensions are for example time, product and source location, but also other dimensions are possible like the customer hierarchy. These dimensions are searched according to a specified sequence. A possible search sequence can be:

1. ATP for the requested product at the requested date in the requested source location is controlled.
2. If ATP in the previous step is not sufficient, earlier ATP of the requested product at the requested location is searched. ATP found before the desired date is held as inventory for the promised order.
3. If still not enough ATP is found, ATP of the requested product at other locations at the requested date and earlier is searched.
4. If still not sufficient, ATP quantities of alternative products (possibly of higher quality) are searched. The previous three steps are executed for these alternative products.
5. If still not sufficient, the previous steps are executed searching for ATP quantities later than the desired date. This means a late fulfillment of the order.

In case the system finds enough ATP for the order, the order is confirmed and the ATP is reduced, as it is no longer available for new incoming orders. If not enough ATP is found to promise the order in full quantity, a quotation for partial delivery is generated, if this is allowed. Otherwise or in case no ATP is found at all, the order must be rejected. Usually, rejected orders are given a very late promised date, at the end of the planning horizon of the ATP check, instead of telling the customer the order is not accepted.

As already mentioned, another search dimension could be the customer hierarchy. The reason for this dimension is that customer orders reach the company's order entry system randomly. If the availability of goods is limited and not all incoming orders can be fulfilled with this scarce resource, it is possible that important customers cannot be served. Important customers

may have a willingness to pay more or they are long-term customers so that they should be privileged in order to maximize the company's profit. This problem is similar to selling aircraft tickets to different passenger classes (e.g. first class and business class), so that ideas of Revenue Management can be applied. Imagine the case that the low priority customers place their orders before the high priority customers and demand exceeds the total supply. Some of the later arriving high priority customers cannot be served if a simple first come first serve (FCFS) strategy for order promising is applied. The company's customers can be clustered into a number of customer segments and a quota of the available goods should be allocated to each segment. New customer orders now can only consume the allocated quantity of the corresponding customer segment. Depending on the company's strategy it is also possible that customers of higher priority classes can access the allocated quantities of lower customer segments.

Finding the right number of customer segments and the right criteria for segmentation is not a trivial task. Methods for this task can be found in Meyr (2008).

An often discussed problem is the allocation planning of the right amount of ATP-quantities to the customer segments. Allocation rules based on the priority of geographical regions and their specific forecasts are presented in Kilger and Meyr (2008). Meyr (2009) suggests segmenting customers according to their importance and profitability and shows how to allocate ATP to these segments in order to maximize profit based on demand forecasts. Pibernik and Yadav (2009) present a model which guarantees a predefined service level for the high priority segment by explicitly considering a stochastic demand.

Once the ATP is allocated to the customer segments one has to decide how to consume these allocated quantities. In the previously presented search rules the dimension customer hierarchy can be used as a search direction as presented in Kilger and Meyr (2008). A consumption policy for allocated ATP based on an optimization model is presented by Meyr (2009). This model for the online-order-promising decides from which customer segment a newly arrived order should be fulfilled.

Questions and Exercises

1. Calculate the ATP quantities for a product for $t_0 = 1$, based on the information in [Table 7.2](#).
2. To which date an arriving order, with the requested day 2 and a requested quantity of 12, can be confirmed? In case of a shortage at the requested date, which measures are possible to solve this situation?

t	1	2	3	4	5
I_0	10				
$Supply(t)$	15	15	20	30	20
$Committed(t)$	20	20	10	30	5

Table 7.2
Supply and committed
quantity for the
product

7.2 Planning Tasks and Data for Frutado

First we show the planning tasks included in the Frutado Global ATP module, then we summarize the necessary data for the short-term planning problem.

7.2.1 Planning Tasks

In the Frutado case the Demand Fulfillment mainly consists in promising new orders. As the offered products are all finished goods, the ATP quantities are calculated on this basis. No availability checks against the raw materials or production capacity are necessary.

Once an order is promised, the promised quantities and the promised dates are not intended to be changed. But unforeseen events might cause a shortage situation, which leads to a repromising of promised orders.

7.2.2 Data

As the order promising is performed on ATP quantities, it is necessary to know about the current inventory and future supply for each final item. This data is provided by the ERP system and various modules of the planning system. The Global ATP gets the current inventory on-stock at the DCs from the ERP system. The short-term supply for the DCs is determined by the forecast-based replenishment performed by the deployment module, as described in Chapter 8. Information on the mid-term supply can be derived from the results of the PP/DS and the SNP module. For the short-term order promising all data must have a granularity of one day.

The Frutado company fulfills the orders of its 60 customers from the 3 distribution centers (DCs). 20 customers are exclusively assigned to each DC. The orders include information about the requested products, the requested quantities and the date the customer requests the order to be fulfilled. It is assumed that they reach the sales department sequentially and only one order at the same time. Customers want to know the confirmed date and quantity immediately, so that the online-mode of order promising has to be applied.

Arriving customer orders reach the company at the order entry system which is part of the ERP system. The ERP system passes the query to the Advanced Planning System (APS) for order promising.

However it should be noted that the Frutado case does not include the ERP system. Therefore it is not possible to show the interactions between the ERP and APO.

For allocation of ATP and checking the new orders against them, information on customer segments is necessary.

7.3 Modeling the Frutado Planning Tasks and Implementation in Global ATP

In this section we describe the Global ATP module of SAP APO and show equivalents between the planning tasks, mentioned in Section 7.1, and the SAP APO. Then the order promising models implemented for the Frutado case and some extensions are presented.

7.3.1 Introduction to SAP® APO Global ATP

As part of the SAP APO the module Global ATP offers the functionality of determining reliable delivery commitments.

First the information flow between different systems has to be examined, see [Figure 7.1](#). As the Global ATP is a very short-term functionality, the order promising is triggered by new customer orders reaching the ERP system underlying the SAP APO. The ERP sends an ATP-request to the SAP APO containing, amongst others, information about the requested date and the requested quantity of all relevant products. After performing the order promising, SAP APO transfers a confirmation to the ERP system with the date and quantities the order can be confirmed to. The figure also shows the surveillance functionality of SAP APO, the backorder processing.

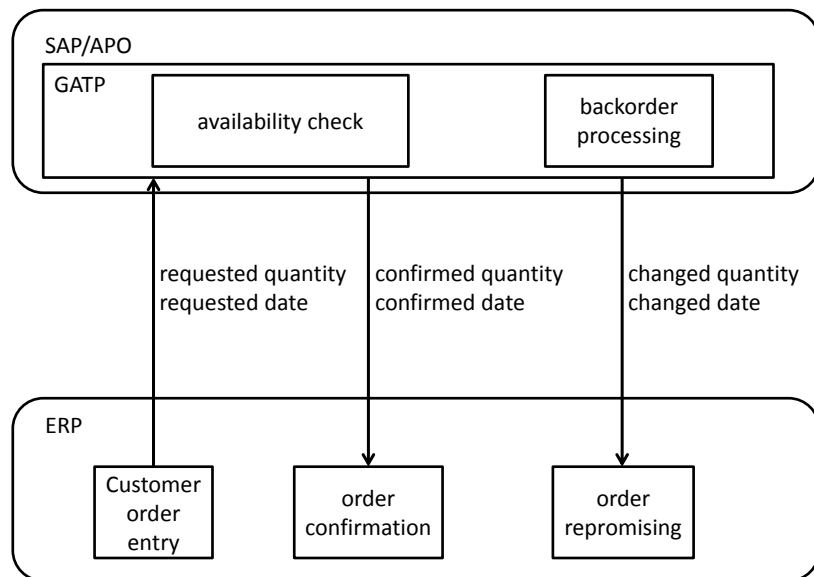


Figure 7.1
Information flows
between SAP APO
and the ERP

basic methods	advanced methods	other functions
product availability check	rules-based ATP	backorder processing
product allocation check	capable-to-promise	simulation
...

Table 7.3
Relevant functions of
global ATP

The functionality of Global ATP extends the ATP check beyond the normal capability of an ERP system. While the ERP checks against inventory of products, the Global ATP also considers future supply derived from the production plans. Moreover it uses more sophisticated search methods like the rules-based search. Functions of Global ATP are divided into basic methods, advanced methods and other functions. The relevant methods for Frutado are shown in [Table 7.3](#).

These Global ATP functions correspond to the planning tasks described in Section 7.1, as explained in the following. Order promising for confirmation of new orders and setting promised dates is represented by various methods like the product availability check, the product allocation check or the rules-based ATP. The surveillance function of demand supply matching is executed by the backorder processing. In case of a shortage situation the backorder processing tries to fix the shortage for the affected orders. Methods for shortage planning at order promising are implemented in basic and advanced methods of Global ATP.

Some of the above mentioned functions are set out in the remainder of this chapter. So the product availability check and the rules-based ATP will be explained in Section 7.3.3 and the simulation function in Section 7.4. Product allocation check, capable-to-promise and backorder processing are considered as extensions in Section 7.3.4.

7.3.2 Customization of Order Promising at the Frutado Company

Before defining specific settings for the availability check the configuration of Global ATP must be made. The Global ATP needs information on ATP of finished goods. The data structures in the SAP APO for the ATP quantities are ATP time series, whose granularities are defined with *bucket parameters*. Setting the bucket parameter to one per day, the system aggregates all receipt and requirement elements of a specific day in the respective bucket.

As SAP APO plans at the product-location level, individual time series are stored in liveCache for each product at each location. Information and data for the ATP check has to be provided for the location-product. An overview of necessary parameters is given in [Figure 7.2](#). This figure also shows relationships between parameters.

Depending on the invoking order type and the requested location-product the strategy of the ATP-check can vary. The parameters *check mode* (de-

pending on the requested product) and *business event* (e.g. a sales order) are transmitted from the ERP with the requesting customer order.

Check instructions define which of the previously mentioned methods for order promising (e.g. availability check, allocation check or rules-based search) are applied. These check instructions could be configured for combinations of values for check modes and business events.

For an order with a specific combination of check mode and business event the so-called *condition technique* can determine the corresponding check instruction. That is assuming this specific combination is defined in the SAP APO.

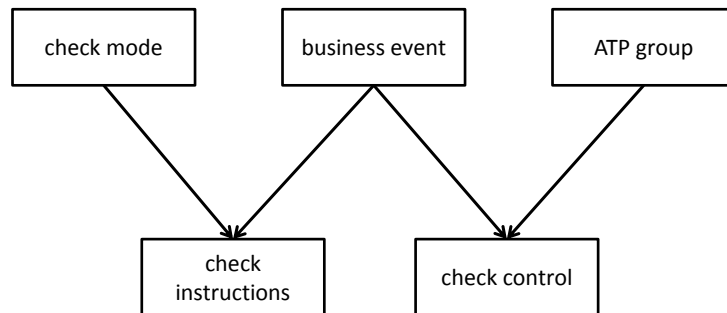


Figure 7.2
Overview of
parameters for global
ATP

For a combination of *business event* and *ATP group* the *check control* for the product availability check is chosen by condition technique. The ATP group is specified by the product requested by the customer order. The elements (physical stock or planned supply) and the horizon of the ATP-quantities depend on the settings of the check control. In other words the check instructions define which method is applied for the availability check (i.e. how ATP is searched), whereas the check control specifies the scope of the check (i.e. which ATP time series are examined).

Given these general settings we now explain the basic Frutado ATP model.

7.3.3 Basic Global ATP Model and Implementation for Frutado

This section contains the model for the product availability check and the rules-based search.

Product Availability Check

In the *product availability check* the order requirements are matched with the supply from the ATP time series. We assume that the only Frutado-relevant combination of ATP group/ business event is 02/A, what means, Frutado has just customer sales orders reaching the SAP APO. For this combination the following check mode for the product availability check is applied.

As the only products sold to customers are finished goods, the check should be based only on these location-products, not on sublocations or versions of them. It is assumed that any quantity of products can be supplied within a certain number of days (see Dickersbach 2009, p. 111). After this *checking horizon* all requirements are confirmed. Within the checking horizon the check methods are applied for order promising. To prevent confirmation of too many orders after the checking horizon, so that the necessary quantity cannot be supplied, we set the checking horizon on a high value of 100 days.

The basic product availability check works according to the following logic:

Based on the customer's requested date the so-called material availability deadline is determined by backward scheduling. As the ATP quantities are based on finished goods, the term material availability deadline might be misleading. To be precise, the quantities of finished goods, expected to arrive at the customer on the requested date, must be available for logistics execution until this deadline. The deadline is calculated by requested date minus the (estimated) transportation lead time and the time needed for picking and packing. If this deadline is in the future, Global ATP is looking from this deadline backwards (that means: direction today) for ATP quantities of the requested location-product. Every time a bucket with ATP is found, the ATP quantity is reduced by the requested quantity. If the total requested quantity can be found, the order can be fully confirmed to the requested date. Otherwise, and also in the case the calculated deadline is before today, the check method is looking forward in the future from the deadline date or today, respective. This will lead to a late delivery of the order.

Figure 7.3 illustrates this logic by three examples. At top, the requested date is far in the future, which leads to a material availability deadline in the future. The product availability check is looking backwards, searching for ATP to confirm the order. As it finds a bucket (1) without enough ATP it looks further backwards (2). Sufficient ATP can be found to confirm the order on time at the requested date. The scenario in the middle calculates a material availability deadline also in the future, but not enough ATP is found (1) to confirm the order on time. So the check method is searching forward (2) for ATP and can confirm the order only with late delivery. In the last case, the customer request is close to today. To fulfill this order on time, the material availability deadline would be prior to today (that means: in the past). As this is not possible, the product availability check is searching for ATP from today looking in the future (1). This also leads to a late delivery for this order, because of the lead time for transportation and picking and packing.

Rules-Based ATP

As a more sophisticated method for checking the availability of products we now explain the *Rules-based ATP* applied for the Frutado company. Here the search is not limited on the requested location-product, but specified

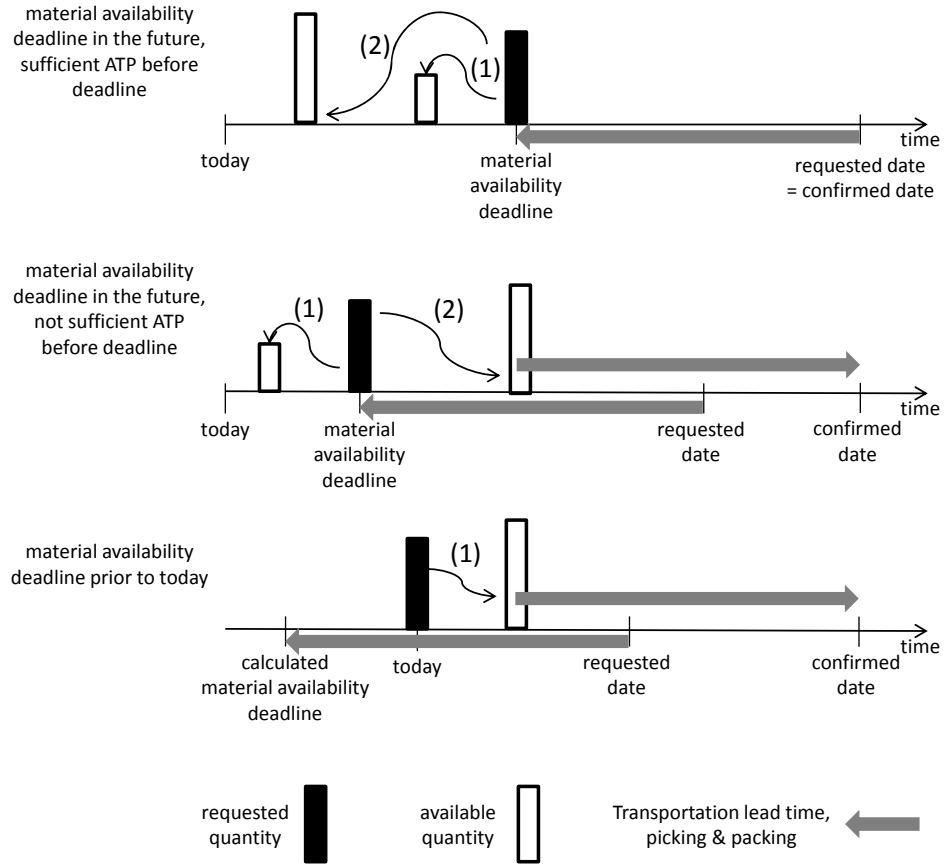


Figure 7.3
 Calculation of the material availability deadline and logic of the product availability check

alternatives to consume ATP-quantities are searched. A possible list of such alternatives in Global ATP is:

- Search for alternative products
- Search for alternative locations
- Search for alternative production-process-models.

In the so-called *rules maintenance* possible alternatives, their sequence and criteria for a valid rules-based ATP check are defined. A possible search sequence is illustrated in [Figure 7.4](#).

As the Frutado company will not substitute requested products by others, product substitution procedures need not be defined. The only accepted alternative is to substitute the delivery source for an order. In other words the order is fulfilled by another DC. This is possible by setting up a new location determination procedure. The Frutado company defines for location FRU_DC_01 the alternative locations FRU_DC_02 and FRU_DC_03, for FRU_DC_02 the alternatives FRU_DC_01 and FRU_DC_03 and so on. There is no prioritization in the sequence of specific alternatives.

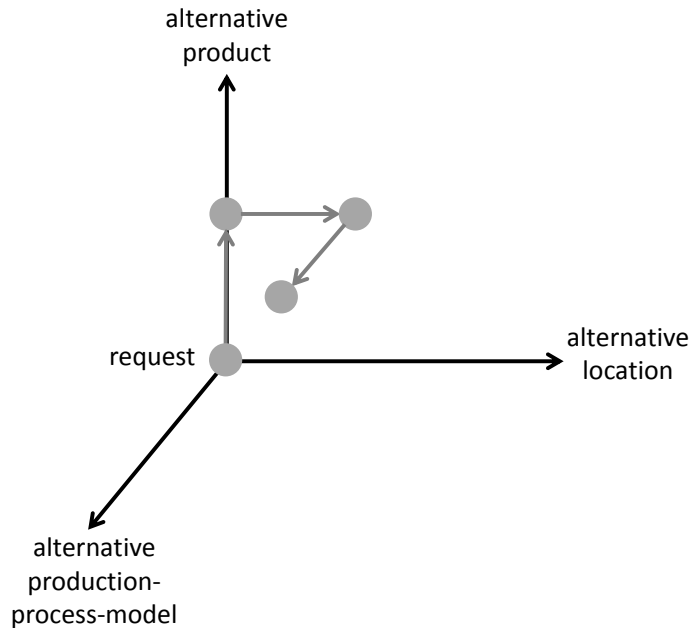


Figure 7.4
Exemplary sequence
for the rules-based
search

In the *rule control* one could define an access strategy for the list of the above mentioned location determination procedure. As there is no prioritization the ATP-check simply starts at the bottom of the list.

The conditions under which Global ATP accepts a confirmation made by the rules-based ATP search are predefined in the *calculation profile* of a rule. Frutado's customers allow a maximum delay of their orders of 5 days but no early delivery. Here one can also define the number of partial shipments.

Summarized the Frutado company applies the following ATP-search rule:

1. Search for ATP of the requested product at the requested DC at the requested date
2. Search for ATP of the requested product at the requested DC prior the requested date
3. Search for ATP of the requested product at the requested DC after the requested date
4. If still not enough ATP is found, the steps 1-3 are repeated for the same product at other DCs

Quantities not yet found cannot be confirmed in the planning horizon of the Global ATP and thus have to be backordered. A simulation of the effects of the different methods is presented in the Learning Units (see Section 7.4).

7.3.4 Extensions

In this section we briefly discuss Global ATP functions not used for the Frutado case.

By using allocations it is possible to achieve a higher service level for high priority customers in shortage situations where demand exceeds available supply. For understanding this concept it is important that allocations are limitations for selling to specific customer segments. Before real-time order promising can use them, the allocated ATP-quantities must be planned.

First the company has to decide on the characteristics of the segments to allocate. These product allocation objects are for example products, product groups, customers or sales regions. These product allocation objects are combined to product allocation groups. If allocation of ATP quantities to customer segments is desired, it must be ensured that already in demand planning, these segments are taken into account. The resulting allocations are then transferred to the product allocation groups. So the allocated quantities of Global ATP are built with respect to the forecasts for the specified segments. As the Frutado company does not segment customers the check against allocation is not applied.

During the time period between order promising and actual fulfillment the situation can occur that the current supply situation for confirmed orders differs from the situation when the first promise (or a subsequent) was made. For example a planned supply for a product may be delayed or cancelled. This means for all affected orders that the fulfillment at their first (or a subsequent) promised dates is no longer possible. So from the current point of view they would not be confirmed to these dates. This situation must be solved, unrealistic dates must be corrected and maybe customers should be informed about delayed delivery. Therefore, the backorder processing is implemented in Global ATP. Its task is to perform a new ATP check for a set of already confirmed orders. Backorder processing can then change confirmed dates, quantities and locations in order to solve the shortage situation. This process can run either in automatic mode or interactively, which allows the planner to redistribute quantities to the orders manually. The backorder processing can also be used for redistributing in case of a shortage situation at the ATP check that would lead to a backlog of a new order, as described above.

If timely production of insufficient ATP is possible, a Capable-to-Promise (CTP) check can be used for avoiding shortage situations and resulting late confirmation of orders. Global ATP could use this method if an integration to the PP/DS is implemented. A CTP check searches for enough production capacity to produce the insufficient quantity of a product.

7.3.5 Processing the Results

We now sum up the results of the Global ATP module and how they are used by other systems in the Frutado company.

The results of a query to the Global ATP are a promised delivery date and quantity, which then, among other information, are transmitted to the ERP, as described in [Figure 7.1](#). Depending on the underlying ERP this information can be edited and sent to the customer.

In the Frutado case, the deployment module (see Chap. 8) uses this data as input for delivery planning.

It should be noted that some tasks of demand fulfillment are overlapping in the Global ATP and deployment modules. The reason is that the latter can not only cover the warehouse replenishment, but also the deliveries to the customers. Both modules can deal with promised customer orders and can handle the tasks demand supply matching and shortage planning, described in Section 7.1. Global ATP provides the backorder processing for these tasks. It can happen that a change of an already promised order cannot be avoided, due to a changed supply situation in the future. Then the backorder processing of Global ATP can send a new, changed confirmation to the ERP-system. It is decided within the ERP-system, whether this change is forwarded to the customer. These tasks can also be performed by the deployment module, which focuses on the transportation costs. For a detailed overview of the planning process “deployment”, please refer to Chapter 8.

While Global ATP can perform on single customer orders, the deployment only treats batches of orders, which is one of the main differences. Thus a real-time order promising needs Global ATP to be applied.

In many practical MTS situations, demand fulfillment occurs, as opposed to the implementation for the Frutado company, in the following way: The deployment module is just used for the replenishment of warehouses and DCs on the basis of demand forecasts. For every incoming order, Global ATP determines the delivery dates and quantities with respect to all scarce resources and decides from which source this order will be shipped. As in the case of MTS the customers want immediate delivery, the promised dates and quantities are transmitted to the Transportation Planning and Vehicle Scheduling module for the delivery planning. Delivery can also be delegated to a logistics service provider. Often, not even an explicit confirmation of the order is sent to the customer but a notification on the shipping of the order.

The Frutado case uses more functionalities of the deployment module in order to show its capabilities to minimize transportation costs and restricts Global ATP to order promising only. Therefore, deployment plans the replenishment of the DCs. For every incoming customer order Global ATP determines a first promised date individually. Deployment then determines the actual delivery dates for a batch of already promised orders with respect to the transportation costs. In doing so, deployment can deviate from the promised dates and quantities calculated by the Global ATP, if a change in the supply situation occurs or a reallocation of orders to DCs might lead to lower delivery costs.

Questions and Exercises

1. Explain the information flows between an ERP system and the module Global ATP.

2. Explain the interconnections of Global ATP and the other SAP APO modules.

7.4 Global ATP Learning Units

The Global ATP learning units are split into the following topics:

1. General settings for Global ATP
2. Settings for the product availability check
3. Simulation of the product availability check
4. Settings for the rules-based availability check
5. Simulation of the rules-based availability check
6. Further functions

The topics 1, 2, and 4 mainly concern aspects of preparing the system. They have been discussed already in Section 7.3.3. We address these only superficially and focus on the topics 3 and 5, where simulation runs for ATP check are shown.

General Settings for Global ATP

To use the Global ATP, the system must be customized for the company. Here master data like business events, check modes and check instructions are configured.

Settings for the Product Availability Check

For the basic product availability check settings regarding the ATP groups are made. We also set up the check control and the scope of check, where we determine the planning horizon and which supply should be considered.

Simulation of the Product Availability Check

In this lesson we show the effects of the basic product availability check.

As there is no underlying ERP-system in the Frutado case, the dynamics of Global ATP cannot be shown. But there is a possibility of simulating the availability situation within SAP APO. One can test how a fictitious order for one product with a requested date and requested quantity would be confirmed. Thus the impact of the general settings can be explained.

We consider two scenarios. In the first one the fictitious order can be fulfilled by the available supply. Then we ask for a higher quantity of the product, keeping the same availability situation. So there remains a quantity that cannot be confirmed.

Scenario I

In the first scenario the fictitious order with the information is sent to the Global ATP:

- Customer: FRU_KUNDE_01 with the appropriate distribution center FRU_DC_01
- Requested product: FRU_SAFT_01
- Desired quantity: 2 HL
- Requested date: 20.04.2009
- Check mode: 030 and business event: A (representing a customer order with basic product availability check only)

The basic availability check can fully confirm this order to the requested date, as can be seen in [Figure 7.5](#).

MAD	MAT	Confirmed Quantity	ONE	FULL	Confirmed Qty	PRO
20.04.2009	12:00	2			2	

Figure 7.5
Confirmation situation
in scenario I
© Copyright 2011. SAP AG.
All rights reserved

In [Figure 7.6](#) we now have a detailed look at the availability situation at the 16.04.2009, the arriving date of this order. No physical stock is detected but projected supplies at 19.03.2009, 09.04.2009, and 12.04.2009 with quantity 1 each are shown. The projected additional ATP quantity is 1 for each day and cumulates to 3 on 20.04.2009. With this information the requested quantity of 2 at 20.04.2009 is available. Logically, this order could be confirmed to this date. As we only show a simulation of the availability check, no real customer order entered the SAP APO. Thus ATP will not be reduced by this simulated demand.

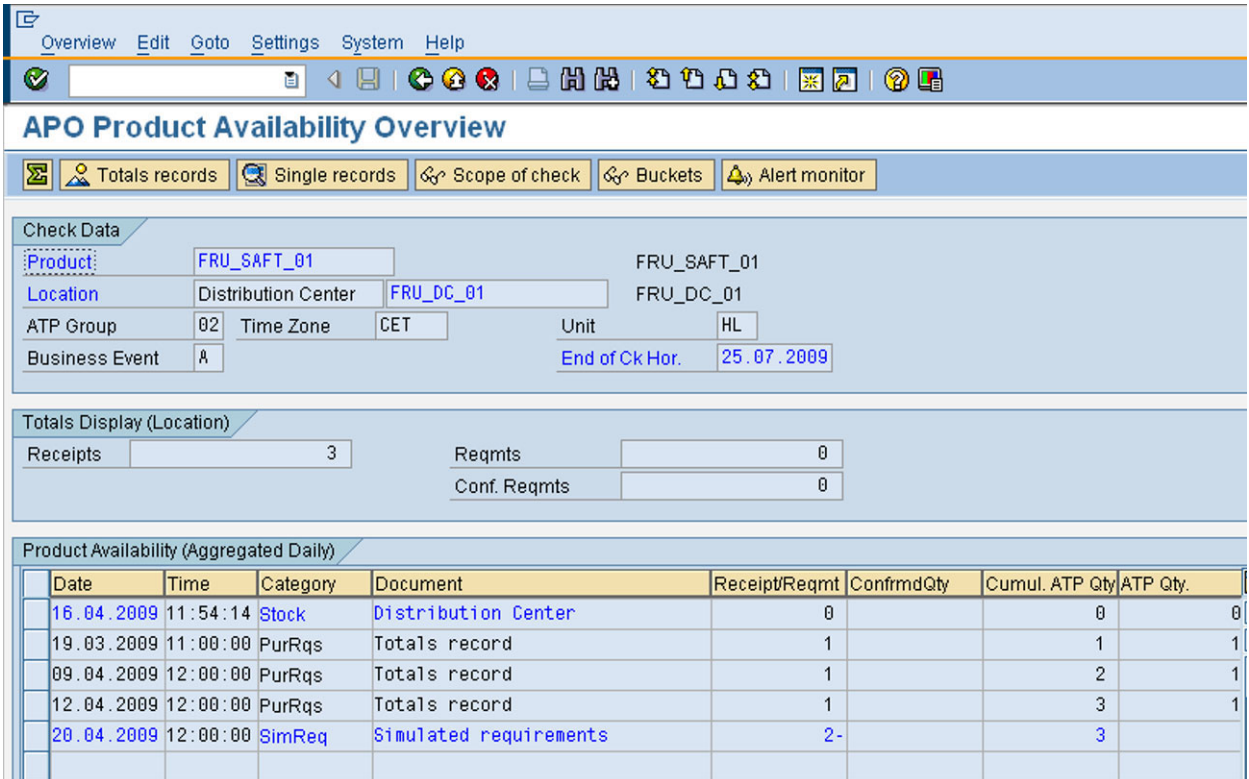


Figure 7.6
 Availability situation in scenario I
 © Copyright 2011. SAP AG.
 All rights reserved

Scenario II

In scenario II we set a fictitious order with the same data but we request a quantity of 5. This leads to the situation depicted in Figure 7.7 where the system cannot fully confirm the order for the requested date. As partial delivery is allowed, a confirmation with quantity 3 is possible for the requested date. The remaining quantity of 2 cannot be fulfilled within the planning horizon of 100 days. Just at the end of the planning horizon the remaining 2 are confirmed.

Settings of Rules-Based Availability Check

For using the rules-based availability check we show how to set up a rule in the integrated rule maintenance. Also settings for Global ATP to determine the relevant rule by using the condition technique are explained.

Simulation of Rules-Based Availability Check (Scenario III)

In scenario II we showed a situation where an order could not be fully confirmed using the basic availability check. Now we show how this situation can be mitigated by a rules-based availability check where substitution of the requested DC is possible.

We therefore change the order from scenario I by the following aspects:

APO Availability Check: Confirmation

Rule
 Check instructions
 Scope of check
 ATP

Product: FRU_SAFT_01
 Distribution Center: FRU_DC_01
 Requested Date: 20.04.2009 12:00
 Open Quantity: 5 HL

Sched. Line Overview

MAD	MAT	Confirmed Quantity	ONE	FULL	Confirmed Qty	PRO
20.04.2009	12:00	3			3	
25.07.2009	00:00	2			5	

Figure 7.7

Confirmation situation in scenario II
 © Copyright 2011. SAP AG.
 All rights reserved

- Check mode: 050 (leading immediately to rules-based availability check)
- Desired quantity: 10 HL

Rules-based search results in the confirmation situation as can be seen in [Figure 7.8](#). Now the DCs 2 and 3 are examined for ATP quantity of the requested product. Each offers an available quantity of 1 at the requested date. The available quantity at DC 1 is 2 at the requested date, which leads in total to the possible confirmation quantity of 4 and the remaining 6 cannot be confirmed at the moment.

Further Functions

The learning unit concludes with an outlook on further functions of Global ATP, like the Capable-to-promise check, the multilevel ATP check and allocation planning. These functions are not used for the Frutado case.

Product/Location	Material Availab...	Rqmt Quantity	Confirmed Qu...	Cumulated Co...	Un...
FRU_SAFT_01 / FRU_DC_01					
Schedule Line 0001	20.04.2009	10 ◊	0	0	HL
Product/Location Substitution					
FRU_SAFT_01 / FRU_DC_02	20.04.2009	10 ▲	1	1	HL
FRU_SAFT_01 / FRU_DC_03	20.04.2009	9 ▲	1	1	HL
FRU_SAFT_01 / FRU_DC_01	20.04.2009	8 ▲	2	2	HL
Remaining Reqmnt (Unchecked)					
FRU_SAFT_01 / FRU_DC_01	20.04.2009	6 ◊			HL

Figure 7.8
Confirmation situation
in scenario III
© Copyright 2011. SAP AG.
All rights reserved

Bibliography

- Ball, M. O.; Chen, C.-Y.; Zhao, Z.-Y. (2004) *Available-to-promise*, in: D. Simchi-Levi; S. D. Wu; Z.-J. Shen (Eds.) *Handbook of Quantitative Supply Chain Analysis – Modeling in the E-Business Era*, chap. 11, Kluwer Academic, 447–483
- Dickersbach, J. (2009) *Supply Chain Management with APO*, Springer, Heidelberg
- Fleischmann, B. (2008) *Distribution and transportation planning*, in: H. Stadtler; C. Kilger (Eds.) *Supply Chain Management and Advanced Planning*, Springer, Berlin et al., 231–246
- Fleischmann, B.; Meyr, H. (2004) *Customer orientation in advanced planning systems*, in: H. Dyckhoff; R. Lackes; et al. (Eds.) *Supply chain management and reverse logistics*, Springer, Heidelberg, 297–321
- Kilger, C.; Meyr, H. (2008) *Demand fulfillment and ATP*, in: H. Stadtler; C. Kilger (Eds.) *Supply Chain Management and Advanced Planning*, Springer, Berlin et al., 181–198
- Meyr, H. (2008) *Clustering methods for rationing limited resources*, in: L. Mönch; G. Pankratz (Eds.) *Intelligente Systeme zur Entscheidungsunterstützung. Konferenzband zur Teilkonferenz der Multikonferenz Wirtschaftsinformatik München, 26.02.2008-28.02.2008*, SCS Publishing House e.V., San Diego et al., S. 19–31

-
- Meyr, H. (2009) *Customer segmentation, allocation planning and order promising in make-to-stock production*, OR Spectrum, vol. 31, 229–256
- Pibernik, R. (2005) *Advanced available-to-promise: Classification, selected methods and requirements for operations and inventory management*, International Journal of Production Economics, vol. 93, 239–252
- Pibernik, R.; Yadav, P. (2009) *Inventory reservation and real-time order promising in a make-to-stock system*, OR Spectrum, vol. 31, 281–307
- Schwendinger, J. (1979) *Master production scheduling's available-to-promise*, in: *APICS Conference Proceedings*, 316–330

Deployment

Martin Grunow¹, Poorya Farahani¹

¹ Technische Universität München, Chair of Production and Supply Chain Management, Arcisstraße 21, 80333 Munich, Germany

This chapter discusses how to develop a detailed distribution plan (deployment plan) to appropriately fill confirmed sales orders and how to replenish inventories at warehouses or distribution centers (DCs) in anticipation of their future demand. In Section 8.1, the deployment problem is discussed. A basic variant of this problem is explained, and a general framework is developed which classifies different extensions. Section 8.2 discusses the planning tasks and data relevant for solving the deployment problem for the Frutado company. The LP formulation of the deployment problem for the Frutado company is presented and explained in Section 8.3. Section 8.4 outlines the implementation procedure of this deployment planning problem in SAP APO and introduces the potential solution methods that can be selected to solve the problem. The learning units for deployment planning are discussed in Section 8.5.

8.1 Introduction to Deployment

Deployment planning determines the distribution operations which are carried out to cover the demand of individual customer locations and to replenish inventories at DCs taking the availability of products at production sites (detailed production planning decisions) into account. This is necessary because supply network planning provides a plan on an aggregate, e.g. weekly level, which determines the production and distribution between plants, warehouses (DCs) and customer groups without taking the detailed operational issues like transportation capacities into account (Chapter 5). Chapter 6 explains how the aggregate production decisions of the supply network plan are detailed in order to generate a short-term production plan. The resulting production

plan provides more accurate information on the availability of each individual product at production plants. Then, in Chapter 7, individual customer orders (as opposed to forecasted demand of customer groups used in supply network planning) are accounted and a delivery date and quantity are promised to each customer. Here, in Chapter 8, a distribution plan is developed which utilizes the more detailed and accurate information on the supply as well as on the demand side. It determines the source, destination, quantity, and date of each shipment on the basis of the available and planned supplies to appropriately fulfill the promised sales orders and to replenish inventories at DCs. For this purpose, the time bucket consideration of the supply network plan is disaggregated into smaller time buckets, an aggregate representation of transportation capacity is taken into account, and distribution operations are disaggregated to cover the demand of each individual customer.

Disaggregation of the previously made aggregate decisions is usually a challenging task due to the so-called aggregation-disaggregation error. Being an intrinsic feature of hierarchical distribution planning (Jonsson et al. 2007), the aggregation-disaggregation error may make it impossible to develop a detailed distribution plan based upon the previously made aggregate decisions. Moreover, in many occasions the realized status of supply and demand deviates from the expected status due to several forms of uncertainty or disruption. For instance, the realized customer orders or detailed forecasts often deviate from the aggregated forecasts. Due to such deviations, the available supply in the system will be either less or more than the demand, and either supply shortage or supply surplus is observed. Thus, deployment decision support models should be equipped with appropriate strategies that are capable of properly handling different supply and demand situations. Safety stocks are essential in this context. However, determination of safety stock levels is not subject of this book. For the sake of simplicity, safety stocks are assumed to be given (cf. Section 2.4) and are therefore omitted from the following discussions.

Following, a basic definition of the deployment optimization problem is presented. The basic problem is formulated as an LP model, and different extensions are discussed using a deployment modeling framework. Before presenting the model, the dimensions and the notation are explained. The notation is further extended in the next sections on an as-needed basis.

Dimensions

QU Quantity Unit

MU Monetary Unit

Indices

j product $j \in J$

i	customer location $i \in I$
l	source location $l \in L$
t	periods of the planning horizon $t \in T = \{1, \dots, T'\}$
m	transportation mode $m \in M$

Data

d_{ijt}	demand size of customer location i for product j in period t [QU]
$\bar{\rho}_{lim}$	actual transportation time from location l to location i with transportation mode m expressed in fraction of periods
ρ_{lim}	transportation lead time from source location l to customer location i with transportation mode m expressed in number of periods: ($\rho_{lim} = \lceil \bar{\rho}_{lim} - 1 \rceil$)
s_{ljt}	planned production of product j at source location l in period t [QU]
c_{limt}	cost of delivering one product unit from source location l to customer location i with transportation mode m in period t [MU/QU]
c'_i	penalty cost of not fulfilling one unit of demand at customer location i [MU/QU]
h_{lj}	cost of storing one unit of product j at source location l for one period [MU/QU]
v_{lm}	aggregate transport capacity of transport mode m at source location l per period [QU]
u_l	inventory capacity at source location l [QU]
b_j	inventory capacity consumption coefficient of product j [QU]
I_{lj0}	initial inventory of product j at source location l [QU]

Variables

I_{ljt}	inventory of product j at source location l at the end of period t [QU]
-----------	---

Z_{lijmt}	amount of delivery from source location l to cover demand of customer location i for product j with transportation mode m in period t [QU]
Z'_{ijt}	amount of unfulfilled demand of customer location i for product j in period t [QU]

Basic Deployment Optimization Problem

Consider a two echelon supply chain in which demand of different products ($j \in J$) of customer locations ($i \in I$) can only be served from inventories of source locations ($l \in L$). Stocks at source locations are replenished according to the detailed production plan s_{ljt} . It is assumed that there are different transportation modes between source and customer locations, and transportation and inventory capacities are limited. Further, it is assumed that the inventory status of customer locations is not known by source locations.

The basic deployment model determines the optimal distribution of the given *inventories* from *source locations* to *customer locations* in each period of the *deployment planning horizon*. Other assumptions used in defining the basic model include:

- Transportation and holding costs are linear in the amount transported/held on stock.
- The consumption of the aggregate transportation capacities is linear in the quantity and transportation time of the shipments.
- Shortage costs are linear in the amount of unfulfilled orders.

The principal challenges in deployment planning are handling inventory *surplus* and inventory *shortage* situations. Following, an LP formulation of the basic deployment problem is presented in which shortage situations lead to lost sales:

Objective Function

$$\begin{aligned}
 (8.1) \quad & \text{Min} \sum_{t \in T} \sum_{j \in J} \sum_{l \in L} h_{lj} \cdot I_{ljt} \\
 & + \sum_{t \in T} \sum_{m \in M} \sum_{j \in J} \sum_{i \in I} \sum_{l \in L} c_{limt} \cdot Z_{lijmt} \\
 & + \sum_{t \in T} \sum_{j \in J} \sum_{i \in I} c'_i \cdot Z'_{ijt}
 \end{aligned}$$

s.t.

Inventory Balance

$$I_{ljt} = I_{lj,t-1} - \sum_{m \in M} \sum_{i \in I} Z_{lijmt} + s_{ljt} \quad \forall l \in L, j \in J, t \in T \quad (8.2)$$

Demand Coverage

$$\sum_{m \in M} \sum_{l \in L} Z_{lijm,t-\rho_{lim}} + Z'_{ijt} = d_{ijt} \quad \forall i \in I, j \in J, t \in T \quad (8.3)$$

Inventory Capacity

$$\sum_{j \in J} b_j \cdot I_{ljt} \leq u_l \quad \forall l \in L, t \in T \quad (8.4)$$

Transportation Capacity

$$\sum_{j \in J} \sum_{i \in I} Z_{lijmt} \cdot \bar{\rho}_{lim} \leq v_{lm} \quad \forall l \in L, m \in M, t \in T \quad (8.5)$$

Non-negativities

$$I_{ljt}, Z'_{ijt}, Z_{lijmt} \geq 0 \quad \forall l \in L, i \in I, j \in J, m \in M, t \in T \quad (8.6)$$

The objective function (8.1) minimizes the total deployment costs including inventory holding and transportation costs between source and customer locations, and shortage costs at customer locations. Transportation costs depend on the volume of products shipped with each transportation mode. In the basic model, the unfulfilled demand of customer locations is treated as lost sales that cannot be filled later.

The inventory status at source locations is tracked in constraints (8.2). The inventory level at source locations changes according to the production plan (s_{ljt}) and the deliveries to customer locations. The inventory status of customer locations is in this model assumed not to be known. Instead, customer locations manage their inventory using their own inventory management system and merely issue some sales orders. Hence, any inventory surplus is held at the source locations. Coverage of customer demands is modeled in constraints (8.3). The demand of each customer location is covered by the receipts from source locations, shipped ρ_{lim} periods ago. The unfulfilled part of the demand is considered as lost sales (Z'_{ijt}) which cannot be fulfilled later and incurs a penalty cost (c'_i). It should be noted that outside the planning horizon ($t < \rho_{lim}$), delivery variables ($Z_{lijm,t-\rho_{lim}}$) in constraints (8.3) are

given as parameters based on the results of previous planning runs. (8.3) assumes that the demand at the beginning of the planning horizon cannot be reduced under the amount already sent.

Constraints (8.4) ensure that required capacities for storing products at the end of each period are smaller than the available inventory capacities at the same locations. Aggregate transportation capacities of each transportation mode at source locations are taken into account through constraints (8.5). For instance, if a source location has two trucks each with a capacity of 20 QU and three trucks with a capacity of 30 QU which all work $1/3$ of a period, then the aggregate truck capacity is approximated by $(130) \cdot (1/3)$ QU per period. The same concept is applied for other available modes of transport. The transport capacity consumption is determined by multiplying the transported quantity with the transportation time $\bar{\rho}_{lim}$. Note that this formulation attributes the transportation capacity consumption to the first period of transport also for transportation operations which take longer than one full period. In the deployment model, the transportation capacity consumption is represented by a single leg of a direct delivery from a source location to a customer. However, it is discussed in Chapter 9 that different deliveries are often combined in a tour. Therefore, the actual capacity consumption is different from the left side of constraints (8.5). To account for different capacity consumption calculations, the aggregate transport capacity used in the deployment model is multiplied by a correction factor. The correction factor is determined on the basis of the outcome of the transportation planning model. The detailed procedure used for updating the aggregate transport capacities is explained in Chapter 9.

The basic deployment problem, as well as the deployment model for the Frutado company, are formulated as LP models and can be solved by standard LP solvers. However, they can also be formulated and solved as discrete optimization problems, in which the delivery quantities can only take discrete values. A discrete deployment model is usually defined by introducing a minimum transportation lot size, and restricting the delivery quantities to multiples of the minimum transportation lot size. Although implementing a discrete solution for the deployment problem is often more convenient, it should be noted that solving a discrete optimization problem is often considerably more complicated than solving its LP counterpart.

There are a number of differences between the basic deployment model and the corresponding supply network planning model discussed in Chapter 5. First, the basic deployment model uses an increased granularity. For example, if the length of time-buckets in supply network planning is one week, the granularity of the deployment model will typically be one day. The finer granularity of deployment planning also allows for modeling individual customers instead of customer groups and individual products instead of product families. Further, the finer granularity allows for considering different modeling options such as back-ordering, or product characteristics such as perishability which will be introduced in the next sections. The second main

difference between the two models lies in the status of the production plan. Supply network planning provides an input for detailed production planning while in deployment planning, detailed production planning decisions are a given. The third difference between the two models is that deployment planning in contrast to supply network planning considers different modes of transport and aggregate transport capacities.

8.1.1 Deployment Modeling Framework

Distribution planning is extensively discussed in literature in terms of strategic distribution network design, tactical distribution planning (cf. Chapter 5), and operational transportation planning (cf. Chapter 9). However, a general modeling framework is still needed for deployment planning. Some of the key attributes in deployment planning are identified in Figure 8.1. Important deployment planning attributes can be classified in a broad sense into physical attributes and model related attributes. Physical attributes relate to the characteristics of products (a) and locations (b) involved. Modeling attributes (c) are comprised of modeling considerations.

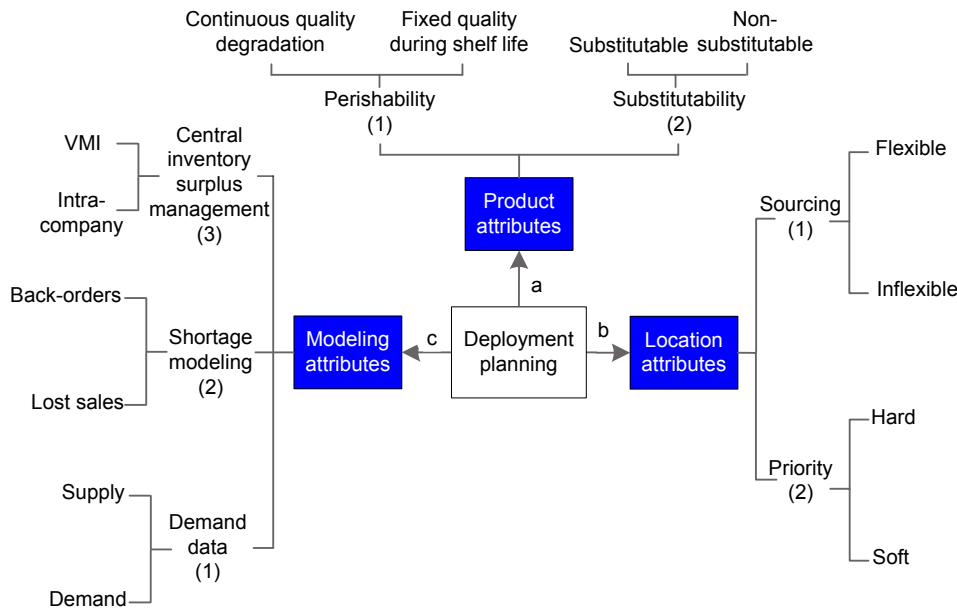


Figure 8.1
Deployment planning attributes

The above definition of planning attributes allows for classifying different deployment problems. In the remainder of this section, we explain how the basic model is influenced by each of the above planning attributes. The basic definition of deployment planning can be extended in several other ways, as well, according to the requirement of a supply chain. For instance, the deployment model may include other supply chain echelons like distribution centers or warehouses.

8.1.2 Deployment Model Classes

This section discusses the impact of the above attributes on deployment planning decisions.

a.1 Product Attributes: Perishability

Perishability limits the possibility of storing products. Two types of product perishability may be observed. The first type involves products that continuously degrade in their quality, and the second type refers to products with fixed shelf life and rather stable quality before the expiration date. Karaesmen et al. (2011) provide a recent survey in distribution planning for perishable products.

Perishability of products can be reflected in the deployment objective function and/or be considered in a set of side constraints. In the first case, a deployment objective could be to provide all customers with a similar product quality. However, including product quality next to other cost related terms in the objective function often leads to a complex multi-objective problem. A simpler approach might be to consider a lower bound on the quality level received by each customer location (by defining a set of constraints on product quality level). In the following, it is explained how the basic deployment model can be extended to account for limited shelf lives of products. For this purpose, an additional time index must be introduced representing the production day of each product. Besides, the required shelf life of each product must also be defined and tracked. As an example, the inventory balance and demand coverage constraints are adapted to account for the required product shelf lives. First, the new notation is introduced.

New Index

\tilde{t} production period

New Data

$I_{\tilde{t}lj0}$ initial inventory of product j at source location l from products produced in period \tilde{t} [QU]

r'_{ij} maximum age of product j to fulfill the shelf life requirements of customer location i expressed in number of periods

New Variables

$I_{\tilde{t}ljt}$ inventory of product j at source location l at the end of period t with products produced in period \tilde{t} [QU]

$Z_{\tilde{t}lijmt}$ amount of product j produced in period \tilde{t} in source location l and shipped in period t with mode m to cover the demand of customer location i [QU]

Inventory Balance

$$I_{tljt} = s_{ljt} - \sum_{m \in M} \sum_{i \in I} Z_{tlijmt} \quad \forall t \in T, l \in L, j \in J \quad (8.7)$$

$$I_{\tilde{t}ljt} = I_{\tilde{t}lj,t-1} - \sum_{m \in M} \sum_{i \in I} Z_{\tilde{t}lijmt} \quad \forall t \in T, l \in L, j \in J, \tilde{t} < t \quad (8.8)$$

Demand Coverage

$$\sum_{m \in M} \sum_{l \in L} \sum_{\tilde{t}=t-r'_{ij}}^{t-\rho_{lim}} Z_{\tilde{t}lijm,t-\rho_{lim}} + Z'_{ijt} = d_{ijt} \quad \forall i \in I, j \in J, t \in T \quad (8.9)$$

Inventory balance constraints at source locations are modeled with constraints (8.7) and (8.8). The products at stock are either from the production in the current period (8.7) or from the inventory at the end of the previous period (8.8). In demand coverage constraints (8.9), only products produced during the last r'_{ij} periods can be used to fulfill customer demand in order to account for the shelf life requirements. This is one of the main complexities of deployment planning for perishable products, because inventory levels at source locations and demand coverage of customers depend not only on the quantity of products at stock but also on their quality (their production periods based on our definition). Note that the values of $Z_{\tilde{t}lijmt}$ in (8.8) are given as parameters for $1 - \rho_{lim} \leq t < 1$.

a.2 Product Attributes: Substitutability

Distribution planning for substitutable products has gained a lot of attention in literature. This issue can be analyzed in a multi-product distribution environment where products can be used interchangeably in times of shortage. Interested readers are referred to Nagarajan and Rajagopalan (2008) for more information. When it is possible to use a group of products interchangeably, and the available inventory is not sufficient to cover the demand for some of the products, the deployment plan should determine whether or not to use substitute products. For this purpose, a penalty cost, smaller than the shortage cost, must be defined for delivering the substitute products.

To represent the concept of product substitution in the basic deployment model, the set of substitute products and the penalty cost of sending substitute products instead of the originally ordered products should be defined. As an example of how the basic deployment model changes, again the inventory balance and demand coverage constraints are reformulated and explained:

New Sets

Ω_{ij} set of products that can be sent to customer location i as substitutes for product j

New Variables

I_{ljt} inventory of product j at source location l at the end of period t with products produced in period \tilde{t} [QU]

$Z''_{lij'jmt}$ amount of product j' transported with mode m from source location l in period t to cover the demand of customer location i for product j [QU]

Inventory Balance

$$(8.10) \quad I_{ljt} = I_{l,j,t-1} - \sum_{m \in M} \sum_{i \in I} Z_{lijmt} - \sum_{m \in M} \sum_{i \in I} \sum_{j': j' \in \Omega_{ij'}} Z''_{lij'jmt} + s_{ljt} \quad \begin{array}{l} \forall l \in L, \\ j \in J, \\ t \in T \end{array}$$

$$(8.11) \quad \sum_{m \in M} \sum_{l \in L} Z_{lijm,t-\rho_{lim}} + \sum_{m \in M} \sum_{l \in L} \sum_{j' \in \Omega_{ij}} Z''_{lij'jmt,t-\rho_{lim}} + Z'_{ijt} = d_{ijt} \quad \begin{array}{l} \forall i \in I, \\ j \in J, \\ t \in T \end{array}$$

Inventory level of product j at source locations is reduced according to the amount of product j delivered to directly fulfill demand of customer locations as well as the amount of product j delivered as substitute to cover the demand of customer locations for other products j' (8.10). Respectively, the demand of each customer for product j is fulfilled with deliveries of product j and deliveries of products j' sent as substitute for product j (8.11).

b.1 Location Attributes: Sourcing

In the basic deployment model, each customer location can be served by each source location. This flexibility in sourcing decisions is specially important when some supply sources face shortage while the others have surplus inventory. However, sourcing flexibility brings up new challenges regarding the way each supplier should prioritize among the customers. Moreover, when the sourcing decisions are determined by the supply network plan, they are often considered fixed and changing them in a short-term plan, like deployment, may cause operational problems. The basic model can be extended to respect the pre-set sourcing decisions. For this purpose, the following sourcing parameters and the corresponding sourcing constraints must be introduced.

New Data

y_{li} =1 if demand of customer location i can be filled by source location l (0, otherwise)

Sourcing Constraints

$$\sum_{t \in T} \sum_{m \in M} \sum_{j \in J} Z_{lijmt} = 0 \quad \forall l \in L, i \in I : y_{li} = 0 \quad (8.12)$$

Constraints (8.12) ensure that each customer location is only supplied by source locations determined in the supply network plan. In the supply network plan, the sourcing decisions are made between the source locations and the aggregate customer groups. In the deployment model, sourcing decisions of the supply network plan are disaggregated to fixed sourcing decisions between each source location and each individual customer location y_{li} .

b.2 Location Attributes: Customer Priority

Customer prioritization, e.g. an ABC classification of customers based on their demand volume, is common practice in supply chain planning. The introduction of customer priorities incorporates new planning measures besides the usual cost terms, thereby often leading to significant changes in the deployment plan. In this regard, two potential approaches might be taken: soft vs. hard customer priorities. Soft customer priorities can be defined through considering different shortage penalties for different combinations of products and customers. In presence of soft priorities, a low priority customer location might be served while the demand of a high priority customer location is not fully covered. In contrast, under a hard priority setting, demand is strictly fulfilled on the basis of priorities.

Since modeling soft customer priorities is quite straight forward, we only discuss a modeling method for introducing hard priorities. A simple modification of the basic deployment model to represent hard priority constraints for customer locations i and i' , which are served by the same source, can be presented as following:

New Data

w sufficiently large number

New Variable

A_{ijt} =1 if customer i observes shortage for product j in period t (0, otherwise)

Customer Priority

$$(8.13) \quad \sum_{l \in L} \sum_{m \in M} \sum_{t' = t - \max\{\rho_{lim}\}}^{t - \min\{\rho_{lim}\}} Z_{li'jmt'} \leq w \cdot (1 - A_{ijt}) \quad \forall i, i' \in I, \\ j \in J, t \in T : i \succ i'$$

$$(8.14) \quad Z'_{ijt} \leq w \cdot A_{ijt} \quad \forall i, i' \in I, \\ j \in J, t \in T$$

Constraints (8.13) ensure that if a high priority customer i observes a shortage, no shipment is made to the lower priority customer i' in periods in which shipments could also have been sent to serve the demand of customer i in period t . Constraints (8.14) set the value of $A_{ijt} = 1$ if customer i observes a shortage of product j in period t . It should be noted that introduction of customer priorities, as in the above formulation, changes the initial LP formulation of the basic deployment model to a Mixed Integer Linear Programming (MILP) model.

c.1 Modeling Attributes: Demand Data

Deployment deals with disruptions in supply and unexpected changes in demand. Supply disruptions may happen due to, e.g. machine breakdowns resulting in supply shortages. A simple way to treat supply disruption in deployment planning is to re-run the deployment model based on the updated supply information. To cope with demand uncertainty, safety stocks are held in supply locations. Different decisions related to safety stock settings in distribution planning are extensively discussed in literature. Here, we explain another aspect of demand uncertainty which is often observed in practice and is relevant for deployment planning. Demand of customer locations usually evolves during the deployment horizon. Therefore, the detailed demand forecasts often exceed the realized sales orders especially for the last days of the horizon (see Figure 8.2).

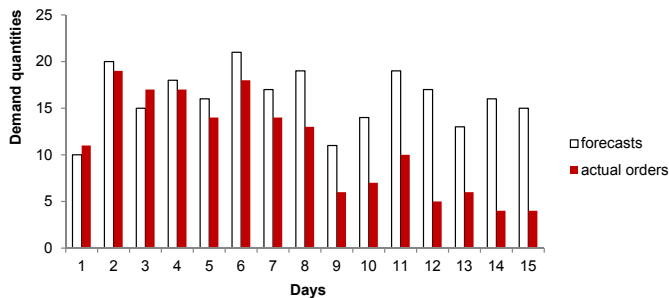


Figure 8.2
Deviations between actual orders and forecast figures

Knowing that actual orders are likely to increase, the deployment model has to be provided with the amount of demand to be covered by an appropriate

combination of both sales data and forecast data. Similarly, the deployment model must be provided with appropriate combination of sales and forecast data when the sales data are higher than the forecast data (days 1 and 3 in Figure 8.2). Revising forecast figures based on sales data is done through defining an appropriate forecast consumption method.

A simple modification of the basic model which copes with this aspect of demand uncertainty is to define the demand as following:

New Data

f_{ijt} demand forecast of customer location i for product j in period t [QU]

O_{ijt} sales order of customer location i for product j in period t [QU]

$0 < \alpha_{ijt} \leq 1$ relative weight of demand forecasts in the deployment plan

with

$$\alpha_{ijt} \geq \alpha_{ij,t-1} \quad \forall l \in L, j \in J, t \in T : t \geq 2 \quad (8.15)$$

and combine sales orders and demand forecasts into the deployable demand data (d_{ijt}) as following:

$$d_{ijt} = O_{ijt} \quad \forall l \in L, j \in J, t \in T : O_{ijt} \geq f_{ijt} \quad (8.16)$$

$$d_{ijt} = \alpha_{ijt} \cdot f_{ijt} + (1 - \alpha_{ijt}) \cdot O_{ijt} \quad \forall l \in L, j \in J, t \in T : O_{ijt} < f_{ijt} \quad (8.17)$$

(8.16) ensure that the deployment model delivers according to sales orders when they exceed demand forecasts. When demand forecasts are larger than order volumes, the demand is determined through a weighted combination of forecasts and sales orders (8.17). To account for the evolutionary characteristic of sales orders, the relative weight of demand forecasts should increase for the later periods of the deployment horizon as represented in (8.15).

c.2 Modeling Attributes: Shortage Modeling

Decision making for deployment planning is strongly influenced by the established agreement between source and customer locations on how to treat product shortages. Shortages might be considered as lost-sales in which case source locations do not have the opportunity to fulfill the demand of customers with late deliveries. Thus, the part of a customer demand which cannot be fulfilled on time is considered as lost sales, which is also reflected in the basic model.

In contrast, when back-ordering is accepted, customers accept late deliveries, mostly until a certain time threshold. Late deliveries incur a backlogging

penalty. If the backlogged orders cannot be fulfilled before the acceptable threshold, they are considered as lost-sales and a higher penalty is incurred.

The basic deployment model can be extended by introducing the following notation to allow for back-ordering.

New Index

t'' late delivery period

New Parameters

E maximum number of periods allowed for late coverage of customer orders

New Variables

$\bar{Z}_{lijtm''}$ amount of late transportation from source location l in period t'' ($t < t'' \leq t + E$) with transportation mode m to cover demand of customer location i for product j in period t [QU]

As an example of how back-ordering changes the basic deployment model, the inventory balance and demand coverage constraints of the basic deployment model are reformulated and explained:

Inventory Balance

$$(8.18) \quad I_{ljt} = I_{lj,t-1} - \sum_{m \in M} \sum_{i \in I} Z_{lijmt} - \sum_{m \in M} \sum_{i \in I} \sum_{t'=t+\rho_{lim}-E}^{t+\rho_{lim}-1} \bar{Z}_{lijtm''} + s_{ljt} \quad \forall l \in L, j \in J, t \in T$$

Demand Coverage

$$(8.19) \quad \sum_{m \in M} \sum_{l \in L} Z_{lijm,t-\rho_{lim}} + \sum_{m \in M} \sum_{l \in L} \sum_{t''=t-\rho_{lim}+1}^{t-\rho_{lim}+E} \bar{Z}_{lijtm''} + Z'_{ij,t} = d_{ij,t} \quad \forall i \in I, j \in J, t \in T$$

The late delivery option in constraints (8.18) and (8.19) allows to cover demand of period t up to E periods later.

c.3 Modeling Attributes: Central Surplus Management

In deployment planning under traditional local inventory control systems, customers do not share their inventory information with suppliers. Therefore, suppliers deploy products to simply cover the demand of customers. Any surplus inventories are held at source locations. This is also the case in the basic model. In contrast, in a central inventory control system, inventories at customer locations are known and managed together with inventories at source locations. Adopting a central inventory control policy significantly increases the flexibility of the deployment planning model. Surplus inventories can now not only be held at source locations but also at customer locations.

Vendor managed inventory

In deployment planning between DCs and customers, the concept of central inventory control is realized in vendor managed inventory (VMI) settings. Deployment decisions under a VMI setting depend on the agreement between DCs and customers upon the ownership of products and their share of inventory holding cost. For instance, when products belong to customers, source locations prefer to push the surplus inventories towards customers. Hence, such VMI agreements typically contain a clause prescribing a maximum inventory level at customer locations. Zhao and Cheng (2009) present an extensive investigation of the effect of VMI settings in distribution planning. As an example of how the formulation of the basic model, which assumes all shortages as lost sales, will be affected by central inventory control, some new sets of constraints are defined. Moreover, the required service levels of customer locations are introduced. The required service level of customer locations can be defined in different ways such as a minimum and maximum inventory level at each customer location, or a certain ratio of the total demand that is allowed to be left unfulfilled. Here, the latter definition is used. Furthermore, the demand coverage constraints (8.3) of the basic model will be replaced by inventory balance constraints at customer locations.

New Data

β_{ij}	required service level at customer i for product j as the maximum share of demand of product j that can be left unfulfilled
u_i	inventory capacity at customer location i [QU]
I'_{ij0}	initial inventory of product j at customer i [QU]

New Variable

$I'_{ijt} \geq 0$ inventory of product j at customer i at the end of period t
[QU]

Inventory Balance at Customer Locations

$$(8.20) \quad I'_{ijt} = I'_{ji,t-1} + \sum_{m \in M} \sum_{l \in L} Z_{lijm,t-\rho_{im}} + Z'_{ijt} - d_{ijt} \quad \forall i \in I, j \in J, t \in T$$

Inventory Capacity

$$(8.21) \quad \sum_{j \in J} b_j \cdot I'_{ijt} \leq u_i \quad \forall i \in I, t \in T$$

VMI Constraints

$$(8.22) \quad \sum_{t \in T} Z'_{ijt} \leq \sum_{t \in T} \beta_{ij} \cdot d_{ijt} \quad \forall i \in I, j \in J$$

$$(8.23) \quad Z'_{ijt} \leq d_{ijt} \quad \forall i \in I, j \in J, t \in T$$

Constraints (8.22) and (8.23) combined with inventory balance constraints at customer locations (8.20) ensure that the inventory level of each customer is sufficient to provide the required service level. The inventory level at each customer is restricted to the inventory capacity of the customer through (8.21).

Intra-company surplus management

Central surplus management is also possible when source locations are production sites and customer locations distribution centers, which are all part of the same company. The model above will lead to a cost-optimal allocation of surplus inventories between source and customer locations. Yet, in many supply chains, products are pushed from production sites to DCs immediately after production, even if the DC-demand in the deployment horizon is less than the amount produced. In such situation, the model can be changed by considering an inventory capacity (u_l) of zero at production sites (source locations). The resulting model will then find a solution, in which inventories are distributed among DCs (customer locations) in a cost-optimal way. However, such a solution may not be desirable, if the pushed inventories will later (i.e. after the end of the deployment horizon) have to be cross-docked

to other DCs. This would double the handling effort in DCs, add to the shipment costs and increase the reaction time on customer orders. In such cases, it is reasonable to distribute the surplus among DCs according to their share of the total demand. This can be achieved in the deployment model by multiplying the demand of each DC by a factor which represents the relative surplus, i.e the total supply divided by the total demand of all DC locations. Thereby, the surplus is completely divided between DCs. An example for a model which pushes inventories from plants to DCs according to such considerations is presented for the Frutado company in Section 8.3.

Questions and Exercises

1. Why is only the aggregate transport capacity taken into account in the deployment model?
2. How can constraints (8.5) be changed such that the capacity consumption of transport operations which take longer than one period is evenly distributed over the periods in which transportation takes place?
3. Is the introduced perishability modeling technique appropriate when products are highly perishable, e.g. , fresh meat and vegetables? If not, what would be the necessary considerations to model deployment planning for such items?
4. How can strict customer priority be modeled when the two customers have both common and separate supply sources?

8.2 Planning Tasks and Data for Frutado

8.2.1 Planning Tasks and Level of Detail

For the Frutado company, the deployment plan determines the detailed distribution of produced beverages from production sites to DCs, and from DCs to customers. The production date and quantity of products at production sites serve as input data (refer to Chapter 6). Products must then be shipped to DCs to fulfill the demand forecasts. Further, it is assumed that customer orders have been confirmed through promising delivery dates and quantities (refer to Chapter 7). The task of deployment planning is to appropriately use the produced products at production sites to fulfill daily demand forecasts at DCs, and to use the inventories stored at DCs to fulfill confirmed customer orders. Deployment planning for the Frutado company is carried out for a horizon of two weeks and on a granularity level of days.

8.2.2 Data

The Fruatdo company classifies customers into groups according to an ABC classification based on their ordering volumes. Products can only be stored

in DCs and inventory capacities of DCs are assumed to be always sufficient. Storing different products at DCs incurs different storage costs. Transportation lanes are defined between production sites and DCs, among DCs, and between DCs and customers (see [Figure 1.1](#)). The defined transportation lanes also determine the cost of transportation between different locations.

In the Frutado company, customers are clustered over DCs such that each customer is only served through one DC. Also, an aggregate representation of transportation capacities is defined and is used for deployment planning (refer to the explanation of aggregate transportation capacity in [Section 8.1](#)). Other deployment related data include storage costs at DCs, late delivery and no-delivery (lost sale) penalties, and maximum accepted delay in delivery of products.

8.3 Modeling Deployment for Frutado

The deployment model for Frutado can be classified and formulated based on the developed framework in [Section 8.1.2](#). Although products are fruit juice with limited shelf lives, they can be stored for a considerably long time. Thus, we do not include perishability in the short-term deployment model. In the Frutado company, sourcing decisions for customer locations are fixed by defining transportation lanes between DCs and customer locations. However, each DC can be supplied by any of the production sites. A central inventory control system manages the inventory status of DCs, and customer locations merely issue sales orders. Moreover, an ABC customer classification is applied, and back-ordering is possible. For a better understanding of the deployment model for Frutado, it is therefore recommended to study the basic deployment model, the model and the extensions on sourcing, customer priority, shortage management through back-ordering and inventory control.

Furthermore, the Frutado company has adopted a make-to-stock (MTS) production strategy. Therefore, products are produced and stocked in DCs in anticipation of customer orders. Customer orders arrive at DCs on a weekly basis. To make the problem more tractable, we decompose the deployment model for the Frutado company into two sub-models: a deployment planning model between production sites and DCs, and a separate deployment planning model between DCs and customer locations. Thus, at the beginning of each week, a deployment model is solved to fulfill demands of customers from inventories in DCs, and another deployment model is solved to fulfill demand forecasts of the second week at DCs from products at production sites. Therefore, DCs are always replenished based on their demand forecasts one week prior to customer order realizations. To cover customer orders of the first week, a certain amount of initial stock is assumed to be available at DCs. Besides, since road transport with trucks is the only mode of transport in the Frutado company, we drop the transport mode index from all related data and decision variables.

In the Frutado company, production sites do not have inventory capacity

(the amount of inventory at the end of each day must be zero at each production site). Therefore, the deployment model between production sites and DCs ships the produced products of each production site to its dedicated DC, as well as other DCs. Shipment of inventories between a production site and its non-dedicated DCs is inevitable as some products may be only produced in some of the production sites. Furthermore, products are shipped to non-dedicated DCs when the production is larger than the demand (pushing the surplus inventories). Following, the deployment model between production sites and DCs for the Frutado company is presented as an LP model on a granularity level of days. The following new notation is used:

New Variable

D_{ijt} demand of product j at DC i in period t which is not lost or backlogged

Objective Function

$$\begin{aligned}
 &Min \sum_{t \in T} \sum_{j \in J} \sum_{i \in I} h_{ij} \cdot I'_{ijt} \\
 &+ \sum_{t \in T} \sum_{j \in J} \sum_{i \in I} \sum_{l \in L} c_{lit} \cdot \left(Z_{lijt} + \sum_{t' = t + \rho_{li} - E}^{t + \rho_{li} - 1} \bar{Z}_{lijt't} \right) \\
 &+ \sum_{t \in T} \sum_{j \in J} \sum_{i \in I} c'_i \cdot Z'_{ijt} \\
 &+ \sum_{t \in T} \sum_{j \in J} \sum_{i \in I} \sum_{l \in L} \sum_{t'' = t - \rho_{li} + 1}^{t - \rho_{li} + E} \bar{c}_i \cdot (t'' + \rho_{li} - t) \cdot \bar{Z}_{lijtt''}
 \end{aligned} \tag{8.24}$$

s.t.

Shipment Constraints at Production Sites

$$\sum_{i \in I} Z_{lijt} + \sum_{i \in I} \sum_{t' = t + \rho_{li} - E}^{t + \rho_{li} - 1} \bar{Z}_{lijt't} = s_{ljt} \quad \forall l \in L, j \in J, t \in T \tag{8.25}$$

Push Constraints

$$\sum_{t \in T} \sum_{l \in L} \left(Z_{lijt} + \sum_{t' \in T} \bar{Z}_{lijt't} \right) = \sum_{t \in T} d_{ijt} \frac{\sum_{t \in T} \sum_{l \in L} s_{ljt}}{\sum_{i \in I} \sum_{t \in T} d_{ijt}}$$

$$\forall i \in I, j \in J : \sum_{t \in T} \sum_{l \in L} s_{ljt} > \sum_{i \in I} \sum_{t \in T} d_{ijt} \tag{8.26}$$

Inventory Balance

$$(8.27) \quad I'_{ijt} = I'_{ij,t-1} + \sum_{l \in L} Z_{lij,t-\rho_{li}} - D_{ijt} \quad \forall i \in I, j \in J, t \in T$$

$$(8.28) \quad D_{ijt} = d_{ijt} - \sum_{t''=t-\rho_{li}+1}^{t-\rho_{li}+E} \sum_{l \in L} \bar{Z}_{lijtt''} - Z'_{ijt} \quad \forall i \in I, j \in J, t \in T$$

Variable Domains

$$(8.29) \quad I'_{ijt}, Z_{lij,t}, \bar{Z}_{lijtt'}, Z'_{ijt}, D_{ijt} \geq 0 \quad \forall l \in L, i \in I, j \in J, t, t' \in T$$

The objective function (8.24) aims to minimize the costs for inventory, distribution, non-delivery (lost-sales) and late delivery of products. Constraints (8.25) determine that all products must be shipped right after production. Constraints (8.26) control the distribution, in case surplus exists ($\sum_{t \in T} \sum_{l \in L} s_{ljt} > \sum_{i \in I} \sum_{t \in T} d_{ijt}$). Inventories are pushed to DCs according to their share of the total demand volume. For this purpose, the shipments to a DC over the deployment horizon have to amount to its total demand multiplied with a factor which represents the relative total surplus. Note that a simple formulation can be used because there are no capacity limits for transport between plants and DCs and for storage at DCs which need to be considered for the Frutado company. The inventory balances for DCs are given in (8.27). The inventory consumption in these balances depends on the amount of the demand which is neither backlogged nor lost (8.28). Constraints (8.29) describe variable domains.

To model deployment planning between DCs and customers for the Frutado company, it should be noted that the Frutado company has no VMI agreements with customers. Hence, customers place orders which are the input for the deployment model. Thus, inventory modeling is only necessary at the DC level. Further, customer priorities are defined in terms of considering different shortage costs (soft priorities). Accordingly, the shortage costs of a high priority customer i (both back-ordering and lost sales costs) are higher than the shortage costs of a lower priority customer i' ($c'_i > c'_{i'}$, $\bar{c}_i > \bar{c}_{i'}$ if $i \succ i'$). In the Frutado company, sourcing decisions for customer locations are fixed. Moreover, to avoid product shortage, products can be transshipped between all DCs. The transshipment of products between DCs are carried out through a dedicated set of trucks. Following, the deployment problem for Frutado is formulated between DCs (as source locations) and customer locations as an LP model on a granularity level of days. The following new notation is used in developing this deployment model:

New Data

$\tilde{c}_{ll'}$	cost of product transshipment from DC l to l' [MU/QU]
$\tilde{\rho}_{ll'}$	actual transportation time from DC l to l' expressed in fraction of periods
$\tilde{\rho}_{ll'}$	transportation lead time from DC l to l' expressed in number of periods ($\tilde{\rho}_{ll'} = \lceil \tilde{\rho}_{ll'-1} \rceil$)
\tilde{v}_l	aggregate transport capacity for transshipment operations at source location l per period [QU]
δ_{li}	=1 if customer location i is served by DC l (0, otherwise)

New Variables

$\tilde{S}_{ll'jt}$	transshipment volume of product j from DC l to l' in period t [QU]
---------------------	--

Objective Function

$$\begin{aligned}
 & \text{Min} \sum_{t \in T} \sum_{j \in J} \sum_{l \in L} h_{lj} \cdot I_{ljt} \\
 & + \sum_{t \in T} \sum_{j \in J} \sum_{i \in I} \sum_{l \in L} c_{lit} \cdot \left(Z_{lijt} + \sum_{t'=t+\rho_{li}-E}^{t+\rho_{li}-1} \bar{Z}_{lijt't} \right) \\
 & + \sum_{t \in T} \sum_{j \in J} \sum_{i \in I} c'_i \cdot Z'_{ijt} \\
 & + \sum_{t \in T} \sum_{j \in J} \sum_{i \in I} \sum_{l \in L} \sum_{t''=t-\rho_{li}+1}^{t-\rho_{li}+E} \bar{c}_i \cdot (t'' + \rho_{li} - t) \cdot \bar{Z}_{lijtt''} \\
 & + \sum_{t \in T} \sum_{j \in J} \sum_{l, l' \in L} \tilde{c}_{ll'} \cdot \tilde{S}_{ll'jt}
 \end{aligned} \tag{8.30}$$

s.t.

Transportation Lane Constraints

$$\sum_{t \in T} \sum_{j \in J} Z_{lijt} + \sum_{t \in T} \sum_{j \in J} \sum_{t''=t-\rho_{li}+1}^{t-\rho_{li}+E} \bar{Z}_{lijtt''} = 0 \quad \forall i \in I, l \in L : \delta_{li} = 0 \tag{8.31}$$

Inventory Balance

$$(8.32) \quad I_{ljt} = I_{lj,t-1} - \sum_{i \in I} Z_{lijt} - \sum_{i \in I} \sum_{t'=t+\rho_{li}-E}^{t+\rho_{li}-1} \bar{Z}_{lijt't} + s_{ljt} + \sum_{l' \in L: l' \neq l} \left(\tilde{S}_{l'lj,t-\bar{\rho}_{l'l}} - \tilde{S}_{l'jt} \right) \quad \forall l \in L, j \in J, t \in T$$

Demand Coverage

$$(8.33) \quad \sum_{l \in L} Z_{lij,t-\rho_{li}} + \sum_{l \in L} \sum_{t''=t-\rho_{li}+E}^{t''=t-\rho_{li}+E} \bar{Z}_{lijtt''} + Z'_{ij,t} = d_{ij,t} \quad \forall i \in I, j \in J, t \in T$$

Transportation Capacities

$$(8.34) \quad \sum_{j \in J} \sum_{i \in I} \left(\bar{\rho}_{li} \cdot \left(Z_{lij,t} + \sum_{t'=t+\rho_{li}-E}^{t+\rho_{li}-1} \bar{Z}_{lijt't} \right) \right) \leq v_l \quad \forall l \in L, t \in T$$

$$(8.35) \quad \sum_{j \in J} \sum_{l' \in L: l' \neq l} \tilde{\rho}_{l'l} \cdot \tilde{S}_{l'jt} \leq \tilde{v}_l \quad \forall l \in L, t \in T$$

Variable Domains

$$(8.36) \quad I_{lj,t}, Z_{lij,t}, \bar{Z}_{lijt't}, Z'_{ij,t}, \tilde{S}_{l'jt} \geq 0 \quad \forall l, l' \in L, i \in I, j \in J, t, t' \in T$$

The objective function (8.30) of the deployment planning model between Frutado's DCs and customers aims to minimize the costs for inventory, distribution, late delivery, non-delivery (lost sales), and transshipment of products between the DCs. As explained in Section 8.2.2, in the Frutado company, each customer location can only be served by its own DC. Constraints (8.31) ensure that these fixed sourcing decisions are respected. Inventory balances of DCs are monitored through constraints (8.32). The following operations determine the inventory level of a DC:

1. Inventory level in the previous period ($I_{lj,t-1}$)

2. Delivery to customers $\left(\sum_{i \in I} Z_{lij,t} - \sum_{i \in I} \sum_{t'=t+\rho_{li}-E}^{t+\rho_{li}-1} \bar{Z}_{lijt't} \right)$

3. Planned supplies during that period ($s_{lj,t}$). In the deployment model for Frutado, the values of $s_{lj,t}$ is decided by the deployment model between plants and DCs which was run in the previous planning week.

4. Transshipments received from other DCs. $\left(\sum_{l' \in L: l' \neq l} \tilde{S}_{l'lj, t - \tilde{\rho}_{l'l}} \right)$
5. Transshipments sent to other DCs. $\left(\sum_{l' \in L: l' \neq l} \tilde{S}_{ll'jt} \right)$

Due to constraints (8.33), demand of each customer location must be covered either on-time, or during an acceptable lateness period of E days. Otherwise, it will be considered as lost sales. Aggregate transport capacity from DCs to customers is modeled through (8.34). Transshipment capacity between DCs is modeled in (8.35). Constraints (8.36) determine variable domains.

Questions and Exercises

1. Is it possible to track product shelf-lives in the Frutado model based on constraints (5.7) in Chapter 5?
2. The deployment problem for the Frutado company is formulated as an LP model implying that shipments can take any fractional values. To model this problem as a discrete optimization problem, a minimum transportation lot size must be defined between each supply and demand location. How does the introduction of a minimum transportation lot size change the Frutado model?

8.4 Implementation

Deployment planning of the Frutado company is implemented in SAP APO. In SAP APO, deployment planning is one of the sub-planning modules of supply network planning (SNP). Interactions between the deployment planning module and the other planning modules of SAP APO are illustrated in Figure 8.3. The dotted lines in this figure represent the scope of the different planning levels. The detailed sourcing decisions determine the supply source(s) for each individual location as opposed to the aggregate sourcing decisions of SNP which are made for customer groups, and for aggregate time buckets. Figure 8.3 also illustrates the feedback mechanisms which deal with the aggregation-disaggregation errors. Transportation capacities are only considered in an aggregated form in the deployment plan as opposed to the detailed modeling of the transportation related items in transportation planning and vehicle scheduling (TP/Vs). Therefore, a feedback mechanism is considered, explained in detail in Chapter 9.

As displayed in Figure 8.3, four sets of data need to be available for developing the deployment plan: the aggregate sourcing decisions made by the SNP module, the detailed forecasts at DCs provided by the Demand Planning module, the available and planned inventories at production sites

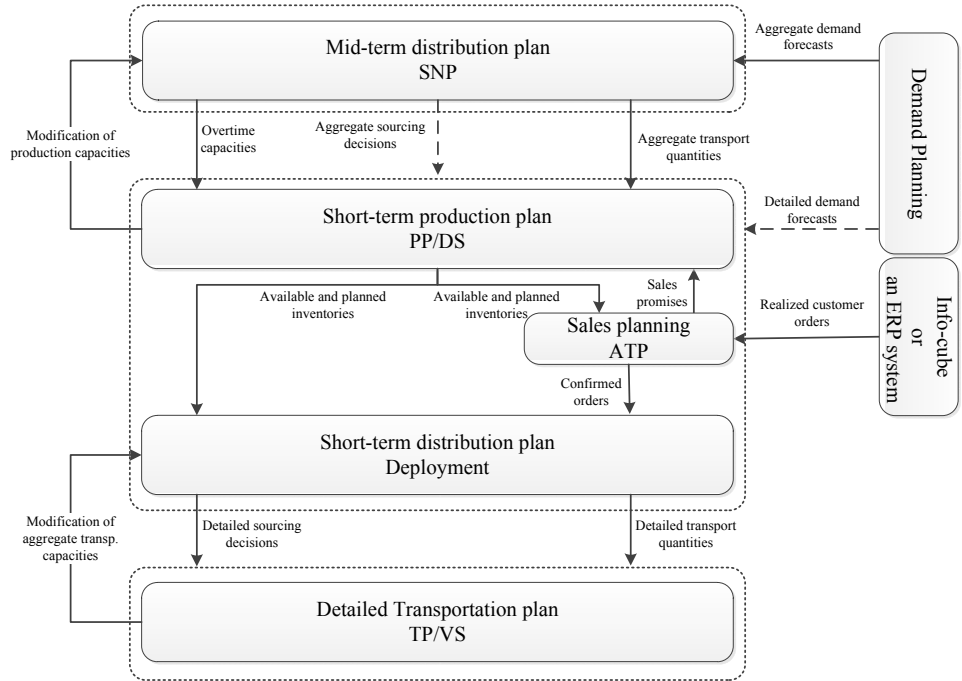


Figure 8.3
The interaction between deployment planning and other SAP APO modules

- The two arrows in dashed line provide information for both PP/DS and Deployment planning modules.

and DCs determined by the production planning and detailed scheduling (PP/DS) module, and the confirmed customer orders provided by the global available-to-promise (GATP) module.

In order to define the scope of deployment planning, three time horizons must be specified: deployment pull horizon, deployment push horizon and SNP Checking horizon. The deployment pull horizon defines the length of time during which orders must be fulfilled. The deployment push horizon is the length of time during which the received supplies can be used to fulfill the orders. The SNP Checking horizon further restricts the available supply through limiting the deployment push horizon. For the Frutado company, all deployment horizons are considered the same and set to two weeks. As the conclusion of deployment planning, detailed daily transportation quantities are determined and passed on to the TP/VS module for generating detailed vehicle routes and schedules.

Deployment planning is carried out in the same planning book as SNP. In order to keep the results of different planning modules separately, the planning version of SNP can be copied to create new planning versions for deployment planning. Moreover, a data-view must be defined presenting the important key figures and the granularity of time buckets for deployment planning. The defined data-view should also consider the appropriate planning calendar. Since, the deployment model is solved for two weeks, a two-week planning calendar with continuous time definition during each planning day is created and applied to the deployment data-view. In order to determine

the deployable quantities during the deployment horizon, the following set of data must be defined.

Deployment Related Quantities

Deployment related quantities determine the supply and demand quantities that must be considered during the deployment planning horizon. In SAP APO, deployment related quantities are called “Available-to-Deploy” (ATD), and can either be defined on the location level in which case they will be identically applied to all products, or be customized for different products at each location. For the Frutado company, ATD quantities are defined on the location level.

Two groups of ATD quantities must be defined for deployment planning: ATD issues and ATD receipts. ATD issues (determining the demand figures to be included in deployment planning) are defined at production sites as demand forecasts of DCs and at DCs as customer orders. ATD receipts (determining the supply figures to be included the deployment planning) are defined at production sites as confirmed production quantities, and at DCs as available inventories. Thus, the deployment plan tries to fulfill demand forecasts at DCs with products produced in the production sites, and customer orders with products stored in DCs.

8.4.1 Deployment Planning Initialization

In order to develop a deployment plan for the Frutado company, the following procedure must be taken as the initialization stage of deployment optimization:

1. Delete SNP Transport Orders

As explained before, SNP generates a weekly distribution plan between production sites and DCs, which forms a basis for the PP/DS module. However, these transport orders are eliminated and new detailed transport orders are generated by the deployment model to allow better utilization of inventories.

2. Import Customer Orders

ATP is the first place where customer orders must be imported in the system. However, running the ATP module is not necessary for developing a deployment plan. If the ATP module is not carried out, customer orders or their demand forecasts can be used instead of confirmed customer orders for running the deployment plan. For this purpose, customer orders must be imported. In practice, customer orders are entered via an underlying transaction system like ERP. However, handling an ERP system is not among the didactical aims of this book. Thus, daily demand forecasts of customers, instead of sales orders, are manually uploaded in APO using InfoCubes.

3. Delete Demand Forecasts of DCs

This step deletes demand forecasts at DCs, identified by the forecast category FA, during the first-week of the deployment plan. This is because the first-week deployment plan only covers customer orders, which are now imported in the system.

After running the above initialization tasks, the deployment optimization can be run via either the deployment data-view or the background transaction codes. To run the deployment optimization of the Frutado, all relevant locations and products must be selected. Further, it should be noted that the option “delete no deployment orders” must be selected. Otherwise, the result of the second planning week of each run will not remain in the system for the next planning run. The solution of an optimization operation can be viewed either in the deployment data view, or through the SNP optimization log file.

It should be noted that by selecting all relevant locations and products the deployment optimization function generates an optimal plan for the integrated supply chain of the Frutado company. However, in some cases solving the optimization model for the entire supply chain might be computationally intractable. To reduce the problem complexity the deployment problem can be decomposed to smaller non-integrated sub-problems. This can be done by, e.g. , splitting the integrated complex supply chain into smaller/simpler supply chains and running the deployment optimization separately for the decomposed supply chains. As explained in Section 8.3, for modeling and solving the deployment planning problem for Frutado, we decomposed the deployment model to two sub-models: deployment planning model between DCs and customers based on customer orders in the current planning week, and deployment planning model between production locations and DCs based on the demand forecasts of DCs in the next planning week. This choice has been made to both reduce the computational complexity of the problem, and to avoid the challenge of simultaneously serving DCs and customer locations in the same period. Moreover, application of heuristic methods, explained in In-depth stream learning unit of the supplementary DVD, can also be considered as alternative methods that most often require lower computational time. Despite the computational advantages of using decomposition and other deployment heuristic methods, their impact on the optimality of the final solution must be carefully examined before using them.

8.4.2 Solution Methods in SAP® APO

The deployment planning problem of the Frutado company is modeled and solved based on the related costs. In the Frutado company, products are distributed to customers in continuous quantities. Thus, it is sufficient to solve the deployment problem using the linear optimizer of SAP APO. As explained in Section 8.1.1, the deployment optimization problem can also be solved as a discrete optimization problem in which case products are

distributed to customers in discrete quantities. However, before deciding on which optimizer to use, one should know that the required computational time for solving the deployment problem with the discrete optimizer is remarkably longer.

As explained in Section 8.1, the principal challenges of deployment planning are planning under supply shortage and supply surplus. To better account for these two situations SAP APO has embedded its cost optimization algorithm within two special solution frameworks of fair-share and push. Following, the two solution frameworks of fair-share and push are briefly introduced.

Fair-Share Rules in Deployment Optimization

Fair-share rules are applied to split the effect of shortage among all or some of the customers in a fair way. In SAP APO, two fair-share measures are considered:

1. Fair-share distribution by demand: distribution of available products evenly among all customer orders.
2. Fair-share distribution by demand and earliest delivery: fulfilling customer orders by giving priority to orders with earlier requested delivery dates. For customer orders with the same requested delivery date, fair-share rule 1 is applied.

To further explain the fair-share rules, a small deployment problem is illustrated in [Table 8.1](#) for two customers served by one supply source. The deployment plans under the above fair-share rules are presented in [Tables 8.2](#) and [8.3](#).

Period	ATD quantity (supply quantity)	Demand of customer 1	Demand of customer 2
1	900	300	0
2		0	200
3		200	500

Table 8.1
Demand and supply
status for a sample
deployment problem

In [Table 8.2](#), the deployment plan is developed based on the fair-share rule 1. Thus the total shortage during the deployment horizon (Total demand - Total supply = Total shortage: $500 + 700 - 900 = 300$) is divided between both customers such that each receive the same percentage of shortage during each period. In this table, the observed shortage for each customer is calculated as $(300/1200) \cdot 100\% = 25\%$.

In [Table 8.3](#), the deployment plan is developed based on fair-share rule 2. Thus, the earliest customer orders are first fulfilled through the available supply (periods 1 and 2) until the available supply is not sufficient to fully cover

Table 8.2
Deployment plan
based on fair-share
rule 1

Period	ATD quantity (supply quantity)	Distribution to customer 1	Distribution to customer 2
1	900	225	0
2	675	0	150
3	525	150	375

orders during a period. In period 3, the available supply is 400 while the total demand is 700 (200 + 500). Thus, the system splits the 300 units of shortage among customers such that each receives the same percentage of shortage. The observed percentage of shortage in this case is $(300/700) \cdot 100\% = 42.85\%$.

Table 8.3
Deployment plan
based on fair-share
rule 2

Period	ATD quantity (supply quantity)	Distribution to customer 1	Distribution to customer 2
1	900	300	0
2	600	0	200
3	400	114.3	285.75

Push Rules in Deployment Optimization

When the total supply is more than the total demand during the deployment horizon, the inventory surplus can be stored either in the supply locations, or in the cost optimal locations (when appropriate cost terms can be defined), or it can be pushed towards the demand locations. Pushing the inventory surplus towards the demand locations might be selected due to several reasons like agreement between suppliers and customers, insufficient storage capacity at supply locations, limitations in storing products due to quality or safety issues, and complexity of defining appropriate storage costs. Similar to the fair-share method, two variants of this strategy, called push rules, are built in SAP APO as following:

1. Push distribution by demand: distributing the supply surplus evenly among all customers during the entire planning horizon.
2. Push distribution by demand and earliest delivery: applying the push rule 1 but through pushing the supply surplus to earliest possible demand day.

Questions and Exercises

1. How can the fair-share rule 1 policy be included in the deployment model of the Frutado company?

2. Devise a deployment plan for table 1 using push rule 2, the same demand quantities, and ATD quantity of 1500.
3. What are the advantages and disadvantages of using a discrete vs. a linear optimization method for deployment planning?

8.5 Deployment Learning Units

8.5.1 Overview

The deployment learning units are split into four broad themes:

1. Master data
2. Deployment planning initialization
3. Deployment optimization
4. In-depth stream: deployment heuristics

The first three units include essential steps for implementing the Deployment module of the Frutado supply chain. The first theme is dedicated to explaining the maintenance of relevant master data for deployment. Details on the initialization stage to deployment optimization are presented in the second theme. In the third theme, the learning units aim at showing how to create cost and optimization profiles, how to run the planning, and how to analyze the results. In order to show the other potential heuristics that can be applied in deployment planning, an in-depth stream section is presented in the fourth part. [Figure 8.4](#) shows the arrangement of deployment learning units.

8.5.2 Basic Stream

Deployment specific master data is already discussed in Section 8.4, and the initialization stage is discussed in Section 8.4.1. In this section the setting, execution and result analysis of deployment optimization are explained. First, a short introduction to the deployment cost profile maintenance is presented showing how to set the weights for the different cost or revenue-related objective terms in the deployment optimization problem.

Before running the deployment optimization, the appropriate settings for the optimization have to be defined. Deployment optimization settings include a set of choices on deployment solution methods, deployment strategies, priority vs. cost-based deployment methods, constraints consideration, and demand priority settings. Further, since deployment is a sub-module of SNP, some of the defined optimization settings of SNP also influence the result of the deployment planning function. These include SNP lot size profile, SNP optimization bound profile, SNP planning profile, and Parallel processing profile. However, defining and setting these profiles is not within the scope of deployment planning for the Frutado company.

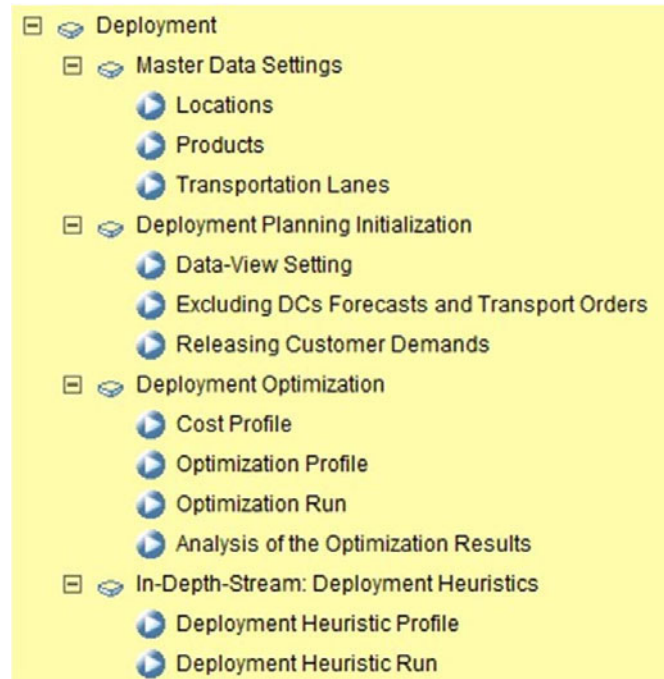


Figure 8.4
Deployment planning
learning units
© Copyright 2011. SAP
AG. All rights reserved

Deployment optimization settings can be adjusted and saved for future use in the deployment optimization profiles. The defined settings of deployment optimization profiles are equally applied to all selected locations and products of the network. However, some of the deployment settings can be customized for product and location combinations in the master data settings. In such cases, the settings of the optimization profile are replaced by the defined settings of the master data. In the second lesson, an optimization profile is maintained which explains the potential planning capabilities of deployment optimization in SAP APO. For example, [Figure 8.5](#) shows that the deployment optimization for the Frutado company takes the transportation and storage capacity into account, but the handling capacity and maximum product quantity stored are not considered.

After defining the cost and optimization profiles, we can run the deployment optimization. For the Frutado company, we choose “Linear Optimization” for deployment planning. However, other optimization methods such as “Discrete Optimization” and “Automatic Cost Generation” are also explained in this learning unit. The results of deployment optimization are analyzed in detail in the SNP optimizer log data and also in the interactive deployment data-view.

8.5.3 In-depth Stream

In this learning unit we explain how to set and run different deployment heuristic strategies. Deployment heuristics are different variants of the fair-

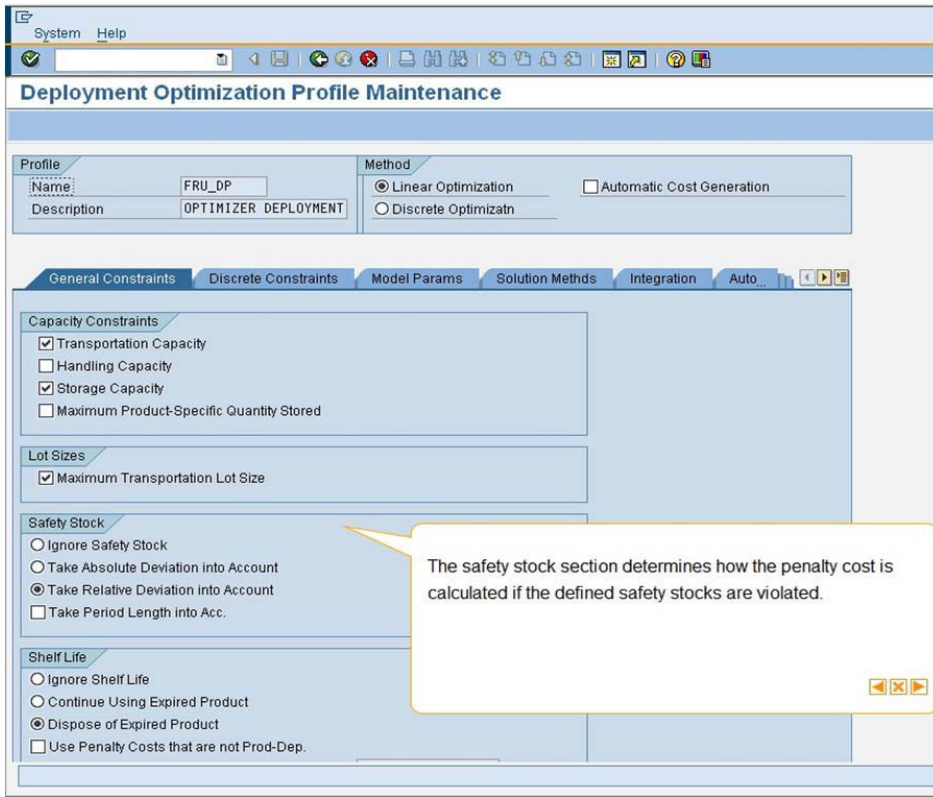


Figure 8.5
Optimization profile
© Copyright 2011. SAP
AG. All rights reserved

share and push rules explained in Section 8.4.2. These include the following methods:

- Fair-share Rule A: Proportional Distribution Based on Demands
- Fair-share Rule B: Proportional Distribution Based on Target Stock
- Fair-share Rule C: Percentage Distribution Based on Quota Arrangements
- Fair-share Rule D: Distribution Based on Distribution Priority
- Fair-share Rule X: User defined fair-share distribution

Also the following variants of push rules can be applied in the deployment heuristics:

- Pull Distribution
- Pull/Push Distribution
- Push Distribution by Demand
- Push Distribution by Quota Arrangement

- Push Distribution Taking the Safety Stock Horizon into Account
- User defined push rule

The detailed explanation of these fair-share and push rules is provided during this learning unit. In contrast to the deployment optimizer, the deployment heuristic methods consider each product and each supply source separately. Further, they try to align themselves with the available SNP plan. Moreover, when using the deployment optimizer, the user can decide whether to keep the previously planned deployment orders or to delete them and create a new deployment plan. With the deployment heuristic function, though, it is not possible to delete the previously planned deployment orders. In the first lesson of this learning unit, we explain how to set a deployment heuristic profile. Performing the deployment heuristic and analyzing the results are explained in the second lesson for one sample product.

Questions and Exercises

1. What are the differences between Fair-share rule A and Fair-share rule B?
2. Deployment heuristics plan for each combination of supply and products separately. How does such planning approach influence the result of deployment planning for a supply chain?
3. Between the deployment optimization and heuristic methods, which one is easier to implement in practice?

Bibliography

- Jonsson, P.; Kjellsdotter, L.; Rudberg, M. (2007) *Applying advanced planning systems for supply chain planning: three case studies.*, International Journal of Physical Distribution & Logistics Management, vol. 37, no. 10, 816 – 834
- Karaesmen, I. Z.; Scheller-Wolf, A.; Deniz, B. (2011) *Managing perishable and aging inventories: Review and future research directions*, in: K. G. Kempf; P. Keskinocak; R. Uzsoy (Eds.) *Planning Production and Inventories in the Extended Enterprise, International Series in Operations Research & Management Science*, vol. 151, Springer
- Nagarajan, M.; Rajagopalan, S. (2008) *Inventory models for substitutable products: Optimal policies and heuristics*, Management Science, vol. 54, no. 8, 1453–1466
- Zhao, Q.-H.; Cheng, T. (2009) *An analytical study of the modification ability of distribution centers*, European Journal of Operational Research, vol. 194, no. 3, 901–910

Part III

Transportation Planning/Vehicle Scheduling (TP/VS)

Martin Grunow¹, Bryndís Stefánsdóttir¹

¹ Technische Universität München, Chair of Production and Supply Chain Management, Arcisstraße 21, 80333 Munich, Germany

There are large potentials for cost savings in optimizing transportation operations in supply chains. Besides the financial importance, transportation processes have significant impact on service quality especially in industries where short delivery times and high service reliability are important features. In this regard, using optimization packages, which rely on optimization methods for solving planning problems, are instrumental. In the food industry for example there are typically few plants supplying large amount of geographically dispersed customers and consequently the expenses related to distribution activities are very large. An efficient distribution system is therefore vital in food distribution, not only to minimize the costs but also to make sure that the customers receive high-quality products.

In this chapter we treat Transportation Planning and Vehicle Scheduling (TP/VS), i.e. the short-term planning problems of transportation processes. In these problems, transportation routes are determined and vehicle resources are scheduled with respect to different types of constraints such as requested delivery dates and transportation capacities. Detailed decisions are made, with a planning horizon of typically a few days. The results are planned shipments.

In the following (Section 9.1) we will give an introduction to transportation planning as discussed in the literature. Firstly packing problems are described, which occur when products are combined into transportation loads, e.g. packed on pallets. Then the basic Vehicle Routing Problem (VRP) is formulated and typical problem classes of VRPs are shortly reviewed. Section 9.2 describes different solution approaches for VRPs. Section 9.3

introduces the short-term transportation planning tasks which have to be modeled for the Frutado company. The data needed in order to model the Frutado transportation problem is also described. In Section 9.4 we present the modeling of Frutado’s planning tasks. Firstly, the Transport Load Builder (TLB) is described, then a model for Frutado’s vehicle routing is introduced, and finally some possible extensions to the model are identified. Implementation and hierarchical integration with Deployment is described in Section 9.5. In the last part (Section 9.6) the structure of the learning units for TP/VS is shown, with explanations on the various sub-learning units which are provided for this module.

9.1 Transportation Planning and Vehicle Scheduling in the Literature

9.1.1 Transportation Load Building

Distributed goods have to be packed into transportation loads, both for protection and for ease of handling. The general objective of *transportation load building* is to reduce transport, storage and handling costs by making more efficient use of stowage volume and loading space. A transportation load can for example be a pallet with various products. The first step is to decide which products should be assigned to the pallet, and then it is determined how the products are stacked on the pallet. There are many software packages available on the market specializing in transportation load building, e.g. PALLETMANAGER (Pallet Manager 2011), CAPE Systems (CAPE Systems 2011) and Quick Pallet Maker (Quick Pallet Maker 2011).

The problem of assigning products to pallets is known in the literature as the *bin packing* problem. The objective is to minimize the number of bins (pallets) needed, where the capacity of the bins is fixed. Two variations of the problem are that either the full list of products to be packed is known in advance (off-line), or that initially the products to be packed are not known (on-line). Here we will concentrate on the off-line version and develop a mathematical model where products of different weights are assigned to identical pallets in order to minimize the number of pallets needed. Note that the model below only considers the weights of the products and pallets, but not their spatial dimensions or other criteria.

Symbols

Indices

B pallet $b \in B$

L product $l \in L$

Data

w_l weight of product l

q weight capacity of pallet

Variables

y_b =1 if pallet b is used, 0 otherwise
 $b \in B$

x_{lb} =1 if product l is put on pallet b , 0 otherwise
 $l \in L, b \in B$

Objective Function

$$\text{Min} \sum_{b \in B} y_b \quad (9.1)$$

s.t.

Pallet Capacity

$$\sum_{l \in L} w_l \cdot x_{lb} \leq q \cdot y_b \quad \forall b \in B \quad (9.2)$$

Product Assignment

$$\sum_{b \in B} x_{lb} = 1 \quad \forall l \in L \quad (9.3)$$

Binary Conditions

$$y_b \in \{0, 1\} \quad \forall b \in B \quad (9.4)$$

$$x_{lb} \in \{0, 1\} \quad \forall l \in L, b \in B \quad (9.5)$$

The objective function (9.1) minimizes the number of pallets used because of handling effort and stowage space. Constraints (9.2) ensure that the weight of the products assigned to the pallets does not exceed the pallet capacity. Each product should be assigned to exactly one pallet (9.3). Finally,

constraints (9.4) and (9.5) define the binary variables. Note that more restrictions could be incorporated into the model, like restrictions on the composition of the load. As an example consider the case where only similar products, or products from the same loading group, should be loaded together on the same pallet. The bin packing problem is an NP-hard problem, and therefore it is often not possible to provide exact solutions within a realistic time limit. For diverse methods of solving different types of bin packing problems, we refer to a survey on approximation algorithms for bin packing in Coffman et al. (1997).

The model above only assigns the products to specific pallets, but it does not consider how the products should be ordered on the pallet. Finding the best way of efficiently loading products (e.g. rectangular boxes) on a pallet is known in the literature as the *pallet loading* problem. The primary issue in pallet loading is to consider the geometric composition of the pallet loads, like the length, width and height limits. However, secondary issues often have to be taken into account like ensuring load stability. For example if all layers of the pallet are ordered in the same way, columns of boxes are formed which could separate from the rest of the pallet. Also if the products are of different weight, it might for example be necessary to assign heavy items to the bottom layer or to distribute the products evenly. Another specific issue is that the products on the pallets should be arranged in the order of unloading, with the aim of minimizing handling time at the destination.

The rectangular pallet loading problem is known to be NP-complete and therefore heuristic approaches often have to be employed, Alvarez-Valdes et al. (2005) developed for example a tabu search algorithm to solve the problem. For a survey on research related to pallet loading we refer to Ram (1992). A typology with a wider scope including also all types of cutting and packing problems can be found in Wäscher et al. (2007), which is partially based on a typology introduced by Dyckhoff (1990).

Questions and Exercises

1. Describe the difference between off-line and on-line bin packing.
2. The problem formulation above results in a lot of symmetries, which make problem solving inefficient. Give suggestion on how to break the symmetries.
3. Describe the interdependency between bin packing and pallet loading.
4. Create an example that shows the interdependency between bin packing and pallet loading.

9.1.2 Formulation of the Basic Vehicle Routing Problem

Most approaches in transportation planning are based on the well-known *Vehicle Routing Problem* (VRP). The VRP is a problem concerning the

distribution of goods between depots and customers. Routes are determined for each vehicle, which starts and ends at its depot, such that all operational constraints are fulfilled and the cost of transportation is minimized. The cost is typically approximated by distance or travel time. Other objectives can also be considered, e.g. minimizing the number of vehicles, balancing the duration of routes or minimizing penalties for lack of service.

The VRP has been studied widely, and there exists a vast body of literature. Dantzig and Ramser (1959) introduced the problem, then Clarke and Wright (1964) improved the approach taken by Dantzig and Ramser. Since then, there has been a rapid progress in solving the problem and its different variants, particularly due to development of computer systems. The VRP occurs frequently in practice and many variants of the VRP are motivated from real life applications. Besides the typical application involving delivery of goods, other applications are for example waste collection, order-picking in warehouses and school bus routing.

A special case of the VRP arises when only one vehicle is available, which is known as the *Traveling Salesman Problem* (TSP). A traveling salesman should visit several customers exactly once, starting and ending at a home location (depot). It is assumed that the distances between the locations are given and that the vehicle has no capacity limit. The shortest route during which each customer is visited only once should be determined.

For the VRP there is a set of vehicles available, located in one or more depots. The vehicles have a limited capacity, and each customer has a known demand. The scheduling problem in this case is therefore to assign customers to vehicles and routes subject to capacity constraints. A basic VRP is given below, formulated as a three-index vehicle flow model.

Symbols

Indices

$G = (V, A)$	directed graph on which the problem is defined
N	customers $i \in N = \{1, \dots, n\}$
S	non-empty subset of customers $S \subseteq N$, $\bar{S} = V \setminus S$
V	nodes $V = N \cup \{0, n + 1\}$ nodes 0 and $n + 1$ are the origin and destination depot, respectively
A	arcs $(i, j) \in A$
K	vehicles $k \in K$

Data d_i demand of customer i q^k capacity of vehicle k c_{ij} cost of arc (i, j) **Variables** x_{ij}^k =1 if arc (i, j) is used by vehicle k , 0 otherwise
 $(i, j) \in A, k \in K$

Objective Function

$$(9.6) \quad \text{Min} \quad \sum_{(i,j) \in A} \sum_{k \in K} c_{ij} \cdot x_{ij}^k$$

s.t.

Customer Assignment

$$(9.7) \quad \sum_{j \in N \cup \{n+1\}} \sum_{k \in K} x_{ij}^k = 1 \quad \forall i \in N$$

Flow Constraints

$$(9.8) \quad \sum_{j \in N \cup \{n+1\}} x_{0j}^k = 1 \quad \forall k \in K$$

$$(9.9) \quad \sum_{i \in N \cup \{0\}} x_{i,n+1}^k = 1 \quad \forall k \in K$$

$$(9.10) \quad \sum_{i \in N \cup \{0\}} x_{ij}^k - \sum_{i \in N \cup \{n+1\}} x_{ji}^k = 0 \quad \forall j \in N, k \in K$$

Subtour Elimination

$$(9.11) \quad \sum_{i \in \bar{S}} \sum_{j \in S} \sum_{k \in K} x_{ij}^k \geq 1 \quad \forall S \subseteq N, |S| \geq 2$$

Vehicle Capacity

$$\sum_{i \in N} \sum_{j \in N \cup \{n+1\}} d_i \cdot x_{ij}^k \leq q^k \quad \forall k \in K \quad (9.12)$$

Binary Conditions

$$x_{ij}^k \in \{0, 1\} \quad \forall (i, j) \in A, k \in K \quad (9.13)$$

The model formulation starts with the objective function (9.6), minimizing the cost of the arcs used by each vehicle. Constraints (9.7) restrict the assignment of each customer to exactly one vehicle route. The next three constraints restrict the flow of the vehicles; all vehicles should start at the origin depot (9.8), end at the destination depot (9.9) and flow conservation for all customers and vehicles should be respected (9.10). Constraints (9.11) are subtour elimination constraints, where the left hand side represents the flow into subset S with \bar{S} defined as $\bar{S} = V \setminus S$. Each subset of customers should have at least one arc going into it. The total demand for a particular vehicle route cannot exceed the capacity of the vehicle (9.12). Finally, constraints (9.13) define the variables as binaries.

Questions and Exercises

1. Describe the difference between TSP and VRP.
2. In order to get an idea of the computational efforts of the problem, please estimate the size of the problem. How many constraints and variables are needed of the problem instances in a general form?
3. Formulate a model that minimizes the number of vehicles, instead of the cost of transportation as presented in the model above.
4. Given the data exhibited in [Tables 9.1-9.3](#), solve the planning model presented above by use of an MIP solver. Solve the problem first without subtour elimination constraints. Which subtours occur? Add subtour elimination constraints for those subtours. Repeat the process until no subtours occur. How many iterations are needed?
5. In order to minimize the number of customer locations visited in a route, it is possible to define a penalty cost for each stop-off. It can either be defined as a soft constraint where the optimizer attempts to schedule as few stop-offs as possible, or as a hard constraint where the number of stop-offs cannot exceed a predefined value. This can for example be of relevance in food distribution where a temperature-controlled truck is used. Too many stops, where the truck is opened, can negatively affect the product, and therefore maximum of vehicle stops needs to be

defined (Ambrosino and Sciomachen 2007). Add a hard constraint for maximum stop-offs to the model presented above. Define the maximum number of stop-offs in a vehicle route as ω .

c_{ij}	Node j						
	0	1	2	3	4	5	6
Node i							
0	0	60	76	74	78	141	117
1	60	0	26	24	75	122	81
2	76	26	0	11	82	120	74
3	74	24	11	0	82	121	75
4	78	75	82	82	0	75	69
5	141	122	120	121	75	0	62
6	117	81	74	75	69	62	0

Table 9.1
Cost c_{ij}

	Customer i					
	1	2	3	4	5	6
d_i	8	2	3	10	2	2

Table 9.2
Customer demand d_i

	Vehicle k	
	1	2
q^k	9	18

Table 9.3
Vehicle capacity q^k

9.1.3 Typical Problem Classes of Vehicle Routing Problems

Many variants of the basic VRP exist. In this section we will review the main ones shortly. For more details on the various problems and solution methods, we refer to a book on the subject edited by Toth and Vigo (2002b). Several survey papers are also available (e.g. Laporte and Osman 1995). Several classification schemes for vehicle routing and scheduling problems have been proposed in the literature (Desrochers et al. 1990, Bodin and Golden 1981). The variants which will shortly be explained here are:

- Capacitated VRP (CVRP)

- Distance-Constrained VRP (DVRP)
- VRP with Time Windows (VRPTW)
- VRP with Backhauls (VRPB)
- VRP with Pickup and Delivery (VRPPD)
- Time Dependent VRP (TDVRP)

Capacitated VRP (CVRP)

The basic version of the VRP, introduced in Section 9.1.2, is the Capacitated VRP. In this version only the restriction with regard to the vehicle capacity is imposed. The vehicles may be identical or have different capacities such as in the model of Section 9.1.2. They are located at a central depot. Each customer has a demand, and the sum of demands on a route has to be less than the vehicle capacity. This capacity can for example be expressed in terms of weight or volume.

Distance-Constrained VRP (DVRP)

In this problem, the capacity constraint is replaced by a constraint for maximum route duration (or distance). For each arc there is an associated travel time, and for each customer there may be associated service time. The sum of travel times and service times on a route must be less than a predefined maximum route length. If both constraints for capacity and maximum route duration are imposed, the problem is called Distance-Constrained CVRP (DCVRP).

VRP with Time Windows (VRPTW)

This problem consists of designing a set of routes for a fleet of vehicles, for which the capacity needs to be respected. In addition all customers must be served within their time windows. Time windows can in general be hard or soft. For hard time windows, vehicles arriving too early at a customer must wait until the customer is ready to begin service. Arriving too late is not allowed. In the case of soft time windows, they can be violated at a cost. Driver considerations can also be integrated into the VRPTW model, e.g. by defining a minimum or a maximum number of hours per route, or guaranteeing break periods of minimal duration within long routes (Cordeau et al. 2002). VRPTW can arise in a number of applications; consider for example deliveries to customers located in pedestrian zones.

VRP with Backhauls (VRPB)

This problem is an extension of the CVRP. The customers are divided into two types, linehaul and backhaul customers. The linehaul customers need

a delivery of products, whereas the backhaul customers request a pickup. A vehicle can serve both linehaul and backhaul customers, but precedence constraints require that all deliveries are carried out before any pickup. These constraints are associated with the loading/unloading activities and the difficulty in reordering the vehicle load along the route (Toth and Vigo 2002a). As an example of VRPB consider a retail supply chain, where a number of shops (linehaul customers) are served from a warehouse but also the warehouse needs to be supplied by various wholesalers (backhaul customers).

VRP with Pickup and Delivery (VRPPD)

In this problem, the customers are associated with both pickup and delivery. It is assumed that the delivery is performed before the pickup, such that the load of a vehicle when arriving at a location is defined as the total load minus demand for deliveries plus the demand for pickups (Toth and Vigo 2002a). A special case of the VRPPD is the VRP with Simultaneous Pickup and Delivery (VRPSPD). In this problem the customers are associated with both delivery (e.g. goods) and pickup (e.g. waste), and each vehicle leaves the depot carrying the total amount of goods it must deliver and returns the waste back to the depot.

Time Dependent VRP (TDVRP)

In this type of VRP, the travel time between two locations depends both on the distance between them and also on the time of day. In real life applications the delivery time may differ significantly during the day particularly for urban areas, with for example congestion due to rush hours in the morning and afternoon. In the literature there have been some contributions to this problem, see for example articles by Fleischmann et al. (2004), Jabali et al. (2009) and Van Woensel et al. (2008). One should keep in mind that there is always some uncertainty in real world applications which are hard to model for example changes in weather conditions or a traffic jam due to an accident.

Questions and Exercises

1. Formulate a model that extends the model given in Section 9.1.2, such that it also considers time windows (i.e. the VRPTW). Formulate one model for hard time windows and another one for soft time windows.
2. Describe the main difference between VRPB and VRPPD.

9.2 Solution Approaches for Vehicle Routing Problems

Problem instances for VRPs are typically large, and therefore the time needed for computation is long. Consequently, heuristic approaches are normally employed in order to solve the problem. In this section we will shortly review different heuristics and solution methods for the problem. Firstly we look at the case when only one uncapacitated vehicle is available, i.e. the TSP. Solving practical TSP instances to optimality are difficult, because the size of the solution space explodes as more customer locations are added. For n customer locations, the number of feasible routes equals $n!$

Example:

5 customer locations: $5! = 120$
25 customer locations: $25! > 1.55 \cdot 10^{25}$

Different heuristics have been developed for the TSP (Jünger et al. 1995). Like the TSP, it is difficult to solve the VRP. In this section, we are focusing on solution algorithms for a variant of the VRP, namely the VRPTW, because it resembles Frutado's routing problem the most, as will be described later (Section 9.4.2). Lower bounds for the VRPTW can be obtained by using optimization approaches, e.g. Lagrangian relaxation and column generation (Cordeau et al. 2002). The basic idea of *Lagrangian relaxation* is to move a constraint which significantly adds to the problem complexity into the objective function. A penalty is incurred if the constraint is not satisfied. *Column generation* is based on a decomposition scheme, where the problem is simplified by splitting it into a master problem and a subproblem. The basic idea is to use the subproblem to indicate variables which can improve the objective function of the master problem. The column generation approach has sometimes been referred to as Dantzig-Wolfe decomposition, since it represents a generalization of the decomposition presented by Dantzig and Wolfe (1960).

The VRPTW is an NP-hard problem, and even finding a feasible solution to the VRPTW when the number of vehicles is fixed is itself an NP-complete problem (Desrosiers et al. 1995). As a consequence of the complexity of the VRPTW, heuristics which are able to find a good solution within reasonable time are important. A variety of heuristics have been reported in the literature for the VRPTW. Bräysy and Gendreau (2005a) provide a survey on traditional heuristic approaches, like route construction and route improvement heuristics (local search methods). *Route construction heuristics* iteratively insert unrouted customers into partial routes until a feasible solution has been created. *Route improvement heuristics* iteratively improve the solution to a problem by performing searches for neighboring solutions. The disadvantage of these heuristics is that they generally require a considerable adaptation effort to account for new problem situations.

The heart of recent work for solving the VRPTW has been devoted to new heuristic approaches, namely meta-heuristics. *Meta-heuristics* are

solution procedures that guide subordinate heuristic methods in a higher level framework to efficiently search the solution space to find high-quality solutions. The subordinate heuristics can for example be the simple route construction and improvement heuristics described above. An iterative master process is used which allows the search to escape from local optimum by allowing deteriorating solutions. Bräysy and Gendreau (2005b) provide a survey on meta-heuristics for the VRPTW, including for example tabu search algorithms, genetic algorithms and evolution strategies, simulated annealing, guided local search, and variable neighborhood search.

The algorithm offered in SAP APO to solve the TP/VS problem is based on *evolutionary algorithms*. In the literature evolutionary algorithms are typically divided into three main types, those are genetic algorithms, evolution strategies, and evolutionary programming. We refer to a survey on evolutionary algorithms for the VRPTW in Bräysy et al. (2004) for more details. In evolutionary algorithms a population of individuals, representing different solutions to the problem, go through a selection procedure and are manipulated by genetic operators. An iterative calculation of a sequence of populations favors the generation of better solutions through an integrated selection mechanism (Homberger and Gehring 1999).

The design of the solution algorithm employed by the TP/VS module in SAP APO is presented below based on the description in Gottlieb and Eckert (2005). The algorithm is called *evolutionary local search*, because it combines ideas from several meta-heuristics with local search. It is a population based algorithm which employs selection principles known from evolutionary computation. A general scheme of the algorithm is given in [Figure 9.1](#).

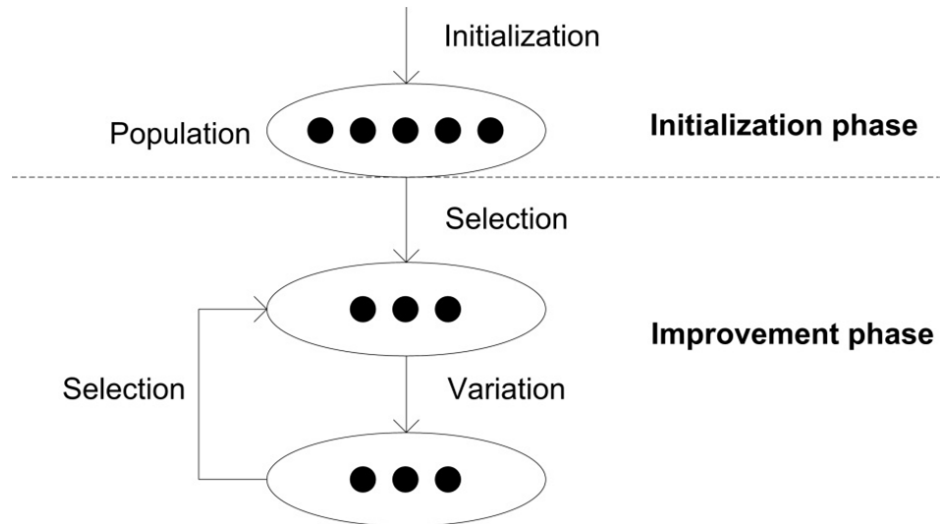


Figure 9.1
 Algorithm for the solution approach in the TP/VS module (based on Gottlieb 2006)

In the initialization phase, the orders are assigned to vehicles, and the vehicles are routed and scheduled using different strategies. This creates

several initial solutions, i.e. a set of individuals forming a population. A selection process eliminates the worst individuals of the population, and the evolutionary search process works further on the remaining individuals. In the improvement phase, there is a variation step followed by another selection process, where each loop iteration is called a generation. Independent variation of all individuals is performed by using more than 20 specialized move operators. A move operator changes an existing solution to a new solution, for example by swapping customers between different vehicle's routes. Subsequently moves are applied with certain probabilities and different search methods like local search, iterated local search, variable neighborhood search, and tabu search are used (Gottlieb 2006). The move operators can be grouped, and each group focuses on certain aspects of the solution, such as:

- Assignment of orders to vehicles
- Routing of vehicles
- Scheduling of loading, unloading, and transport activities

In each generation an iterated local search is used. Once a local optimum is determined, perturbation is applied, and a new local search process starts. If an individual in the population has not improved during several generations it is replaced by a perturbed solution. The algorithm described above has many parameters. How these parameters are set influences the algorithm. SAP delivers a default set of parameters by using test runs on benchmark cases, derived from real world customer scenarios.

9.3 Planning Tasks and Data for the Frutado Company

9.3.1 Planning Tasks

The first step in the planning process is to load the orders on euro pallets on which a maximum of 1400 kg can be loaded. This is called *transportation load building*. The task is to combine confirmed distribution orders that are generated for individual products into multi-product shipments, to ensure that the capacity of the vehicles is properly utilized. This also protects the products and allows handling efficiencies. The next step is to plan the transportation processes. The Frutado company has decided to carry out the transportation by itself. Frutado's supply chain is depicted in [Figure 9.2](#).

The products are first transported to the distribution center (DC) at the plant site. The resources used for transportation between plants and DCs do not represent any bottleneck in the planning process. These transportation processes are therefore not modeled. Transportation between the DCs is needed because the plants do not produce all variants of products, but each DC has to have the whole assortment of products available. Transportation between DCs can also be necessary in order to avoid production bottlenecks

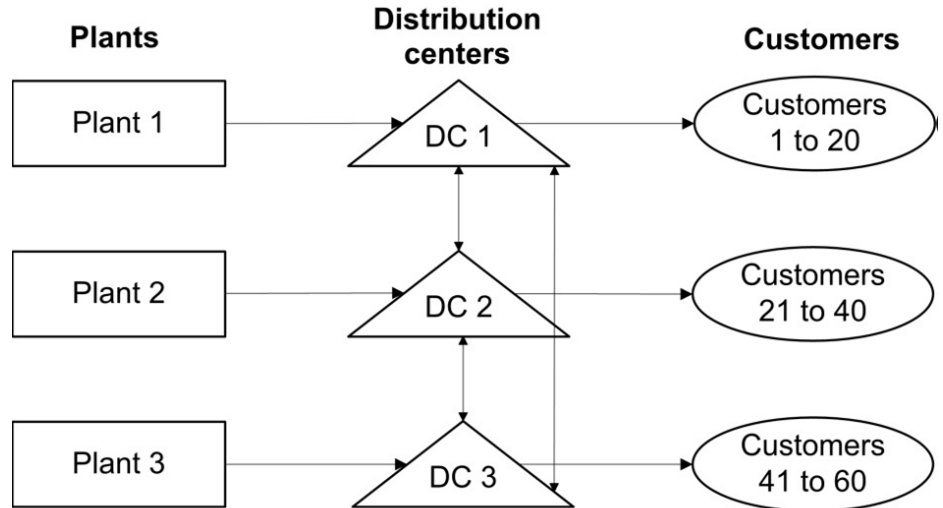


Figure 9.2
Supply chain network
for the Frutado
company

at the production plants. Finally, transportation is needed between each of the DCs and its corresponding customers. These transportation processes are carried out in three delivery areas which can be planned separately. Transportation planning is carried out for a period of one week. Since transportation is not planned for the weekends, this leads to a five day planning horizon.

The Frutado company owns their fleet of vehicles for transportation. The vehicle resources are partitioned into vehicle type groups which share the same characteristics and cost structure, such as duration and distance between each pair of locations in the transportation network. Every vehicle is assigned to a DC and is available during the planning process according to the opening hours of the DC. A maximum duration for a delivery route is defined in order to account for the working time for the drivers. A fixed cost is incurred each time a vehicle is used. Variable costs are incurred for the duration (e.g. per hour) of the trip and also for the distance (e.g. per km) traveled by the vehicle, dependent on the type of vehicle.

An order must be loaded at its source location (e.g. DC) and unloaded at the destination location (e.g. customer). These loading/unloading activities require a local *loading resource*. A loading resource can serve only one vehicle at a time, which means that other vehicles must wait until a loading resource is free. Both fixed and variable *loading/unloading times* need to be modeled. The fixed time occurs when arriving at a location, due to time needed for the trucks to maneuver into a suitable unloading position or the time it takes to untie tied down goods. The variable time depends on the weight of the orders, since it takes longer time to handle the heavy orders.

The customers are divided according to their geographical location evenly between the three DCs, such that each DC serves a total of 20 customers. Clustering of the customers can also be of relevance if for example the drivers have a special knowledge of specific areas. The customers are classified

into three customer categories according to an ABC classification based on their share of sales. The customers do not have significant storage space and can therefore order on a daily basis from their corresponding DC. The customers accept late delivery, however only for a maximum of four days after the required delivery day. Even though an order can be delivered late, a delivery may only happen within specific opening hours of the customers. Consequently there are short time windows (opening hours) within larger time windows (late delivery).

Lateness and non-delivery *penalty costs* for a customer order specify the amount charged as a penalty for late or non-delivered orders. Penalty costs do not represent actual costs, but rather a theoretical value used to determine the most suitable delivery date. It can be hard to quantify such service related values. For the Frutado company, there is a cost differentiation according to customer priority, following the ABC classification enabling preferential delivery to the customers with a higher priority. Figure 9.3 depicts the time windows within which an early or late delivery is allowed.

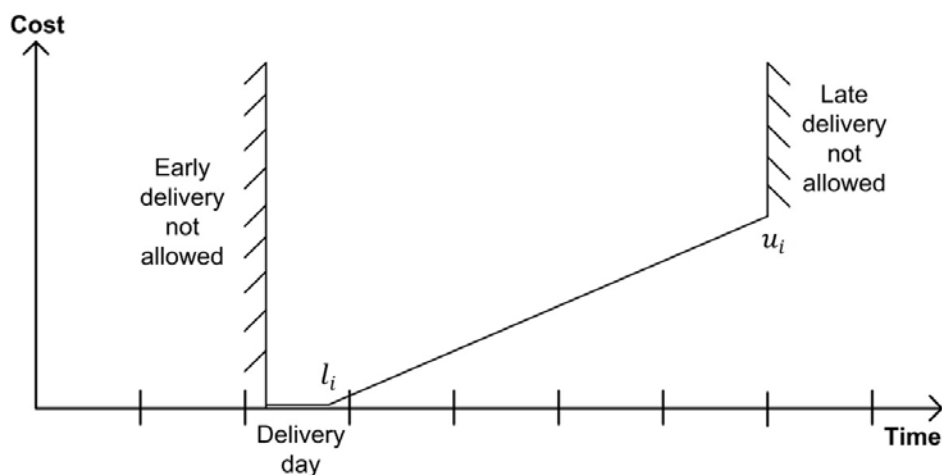


Figure 9.3
Cost for delayed shipment

The figure shows that early delivery is not allowed (a hard constraint; cf. Section 9.1.3). No penalty cost is incurred during the opening hours on the requested delivery day, after that (point in time l_i) linearly increasing penalty costs are incurred. Note that point in time l_i is exactly the closing time of the customer, for example in the evening as in Figure 9.3. A soft constraint needs to be defined to model the delay. After four days (point in time u_i) of delay the order is canceled, and a non-delivery penalty cost is incurred, to be modeled as a hard constraint.

9.3.2 Data

The data which is needed in order to model Frutado's transportation problem is described in this section. Note that units of the different parameters and variables are given; however the model can easily be adapted to other units of measurements.

Symbols**Indices**

$G = (V, A)$	directed graph on which the problem is defined
N	customers $i \in N = \{1, \dots, n\}$
V	nodes $V = N \cup \{0, n + 1\}$ nodes 0 and $n + 1$ are the origin and destination depot, respectively
A	arcs $(i, j) \in A$
K	vehicles $k \in K$
T	days, demand days: $t \in T$, shipping days: $t' \in T$

Data

d_i^t	demand of customer i for demand day t [tons]
t_{ij}^k	travel time for arc (i, j) on vehicle k [minutes]
e_{ij}^k	distance for arc (i, j) on vehicle k [km]
α	loading factor for variable loading time at depot [minutes/tons]
β	unloading factor for variable unloading time at customers [minutes/tons]
$(a_i^{t'}, b_i^{t'})$	time windows at node i on shipping day t' [minutes]
q^k	capacity of vehicle k [tons]
f	maximum length of trip [minutes]
l_i	latest on time delivery for customer i (from this point in time delay is calculated, see Fig. 9.3) [minutes]
u_i	limit for non-delivery of customer i (the point in time when the order is canceled, see Fig. 9.3) [minutes]
m	minutes in a calendar day ($24 \cdot 60 = 1440$) [minutes]
$M_{1,2,3,4}$	constants for Big-M method

cf^k	fixed cost for vehicle k [cost]
cr^k	distance dependent cost for vehicle k [cost/km]
ch^k	duration dependent cost for vehicle k [cost/minute]
cd_i	cost of delayed delivery (lateness) at customer i [cost/minute]
cn_i	cost of non-delivery at customer i [cost]

Variables

$x_{ij}^{kt'}$	=1 if arc (i, j) is used by vehicle k on shipping day t' , 0 otherwise $(i, j) \in A, k \in K, t' \in T$
$v_i^{ktt'}$	=1 if demand day t of customer i is served by vehicle k on shipping day t' (with $t' \geq t$), 0 otherwise $i \in N, k \in K, t \in T, t' \in T$
$w_i^{kt'}$	starting time of service at node i by vehicle k on shipping day t' (if vehicle k does not visit node i on day t' then the variable is undefined) [minutes] $i \in V, k \in K, t' \in T$
$s_i^{kt'}$	service time at node i for vehicle k on shipping day t' [minutes] $i \in V, k \in K, t' \in T$
y_i^t	=1 if demand of customer i for demand day t is served, 0 otherwise $i \in N, t \in T$
z_i^t	delay for customer i for demand day t [minutes] $i \in N, t \in T$

9.4 Modeling the Frutado Planning Tasks

9.4.1 Transportation Load Building for the Frutado Company

Before modeling the transportation processes, this section describes how transportation loads are built up. The TLB, which is a short-term planning tool, can be used for this planning process. The transportation load building

for the Frutado company is analogous to the bin packing model described in Section 9.1.1, where products are combined into pallets. The ordering on the pallet itself is therefore not considered as this is not part of SAP APO. Note that no optimization methods are used in the TLB in SAP APO, and therefore the optimality of its solution is not guaranteed.

As described in Section 9.3.1, orders are to be loaded on euro pallets on which a maximum of 1400 kg can be loaded. Note that only the capacity constraints of the pallet are considered in the TLB, not the capacity restrictions of the vehicle. The TLB is executed for each customer. For each demand day transport loads are built within the weight limits (1400 kg). Therefore each transport load only includes deliveries to a specific customer, not to multiple customers.

Questions and Exercises

1. Given the data in [Table 9.4](#). Write down the bin packing model presented in Section 9.1.1 by stating each constraint explicitly. It is assumed that the capacity of a pallet (q) is 1400 kg.
2. Solve the model by use of an MIP solver.

Product	Type	w_l
4	Ice Tea	84
5	Juice	1057
6	Ice Tea	101
7	Juice	757
8	Ice Tea	292
9	Ice Tea	98
12	Juice	360

Table 9.4
Product weight w_l [kg]

9.4.2 Frutado's Vehicle Routing Problem

Real life VRPs often do not fit completely into one of the standard models described in Section 9.1.3. The problems can vary a lot, each imposing their own special constraints and objective, like for the Frutado case study. In this section a model is developed for the Frutado company's VRP. The basis of the model is the known VRPTW described by Cordeau et al. (2002), where the VRPTW is formulated as a multi-commodity network flow model with time window and capacity constraints. However, many adjustments have

to be made for the model to account for the special settings of the Frutado company.

For the Frutado company, the customers have been partitioned among the DCs (depots), and the vehicles have to return to their home depot at the end of each route. In this case, the overall problem can be decomposed into several independent problems, each associated with a single depot (Toth and Vigo 2002a). This approach is taken when developing the model; that is optimizing the transportation between one DC and its customers. In principle there is also transportation between the DCs, however this problem is rather simple, and the problems are not interconnected as a different fleet is used for this purpose. Therefore the focus of the following model is on the transportation between one DC and its customers.

The objective is a minimization of the global *transportation cost*; including fixed cost per vehicle (cf^k), cost for traveled distance (cr^k), duration cost (ch^k), lateness cost (cd_i) and non-delivery cost (cn_i). The road network is described through a graph, where arcs (i, j) represent the roads, and vertices i symbolize DCs (depots) and customers. Each arc is associated with travel time duration (t_{ij}^k) and distance (e_{ij}^k); the arcs are moreover dependent on the vehicle type.

Different constraints are taken into account when assigning orders to the vehicles. Generally the constraints can be categorized into hard and soft constraints. The *hard constraints* must be fulfilled for the planning to be feasible, therefore hard constraints will always be adhered to. Hard constraints are for example loading capacities of vehicles (q^k), opening hours of locations (a_i^t, b_i^t) and the requirement that vehicles return back to their DC. The *soft constraints* on the other hand are less critical; they are desirable constraints and should be fulfilled to the highest possible level. These constraints are modeled by using penalty costs (like lateness and non-delivery costs) which are then part of the total cost. Soft constraints can therefore be violated subject to penalty costs and their priority can be influenced, e.g. by an according weighting. The total cost is then minimized in the optimization run whilst fulfilling the hard constraints and weighting the penalty costs of the soft constraints.

Before presenting the model, some assumptions are described. For ease of model presentation it is assumed that no backlogs from the previous week are taken into account. It is assumed that the products a customer is demanding per day have been aggregated into a single delivery (d_i^t). Therefore the model does not count for different types of products. Note that following the TLB planning, some customers have many TLB shipments in one day. However, in the model below, the customer demands for different products from a particular demand day are aggregated into one single delivery.

In the model, loading resources are not modeled because it is assumed that there are enough loading resources at the locations and that they are not constraining for the problem. The fixed loading/unloading times are accounted for in the traveling times (t_{ij}^k) for the arcs and the variable

loading/unloading times are modeled as service times ($s_i^{kt'}$) at the locations, i.e. loading at DCs and unloading at customers. The service times are variables which depend on the weight of the deliveries that are loaded or unloaded on the particular shipping day. It is assumed that the time it takes to load/unload a delivery, linearly increases depending on the weight of the delivery.

Objective Function

$$\begin{aligned}
 (9.14) \quad \text{Min} \quad & \sum_{(i,j) \in A} \sum_{k \in K} \sum_{t' \in T} cr^k \cdot e_{ij}^k \cdot x_{ij}^{kt'} \\
 & + \sum_{j \in N} \sum_{k \in K} \sum_{t' \in T} cf^k \cdot x_{0j}^{kt'} \\
 & + \sum_{k \in K} \sum_{t' \in T} ch^k \cdot (w_{n+1}^{kt'} - w_0^{kt'}) \\
 & + \sum_{i \in N} \sum_{t \in T} cn_i \cdot (1 - y_i^t) \\
 & + \sum_{i \in N} \sum_{t \in T} cd_i \cdot z_i^t
 \end{aligned}$$

s.t.

Customer Assignment

$$(9.15) \quad \sum_{t \in T} v_i^{ktt'} \leq \sum_{j \in N \cup \{n+1\}} x_{ij}^{kt'} \cdot M_1 \quad \forall i \in N, k \in K, t' \in T$$

$$(9.16) \quad \sum_{k \in K} \sum_{t' \in T} v_i^{ktt'} = y_i^t \quad \forall i \in N, t \in T$$

Flow Constraints

$$(9.17) \quad \sum_{j \in N \cup \{n+1\}} x_{0j}^{kt'} = 1 \quad \forall k \in K, t' \in T$$

$$(9.18) \quad \sum_{i \in N \cup \{0\}} x_{i,n+1}^{kt'} = 1 \quad \forall k \in K, t' \in T$$

$$(9.19) \quad \sum_{i \in N \cup \{0\}} x_{ij}^{kt'} - \sum_{i \in N \cup \{n+1\}} x_{ji}^{kt'} = 0 \quad \forall j \in N, k \in K, t' \in T$$

Vehicle Capacity

$$\sum_{i \in N} \sum_{t \in T} d_i^t \cdot v_i^{ktt'} \leq q^k \quad \forall k \in K, t' \in T \quad (9.20)$$

Loading and Traveling Times

$$\sum_{i \in N} \sum_{t \in T} d_i^t \cdot v_i^{ktt'} \cdot \alpha \leq s_0^{kt'} \quad \forall k \in K, t' \in T \quad (9.21)$$

$$\sum_{t \in T} d_i^t \cdot v_i^{ktt'} \cdot \beta \leq s_i^{kt'} \quad \forall i \in N, k \in K, t' \in T \quad (9.22)$$

$$w_i^{kt'} + (t_{ij}^k + s_i^{kt'}) \leq w_j^{kt'} + (1 - x_{ij}^{kt'}) \cdot M_2 \quad \forall (i, j) \in A, k \in K, t' \in T \quad (9.23)$$

Time Window

$$a_i^{t'} \leq w_i^{kt'} \leq b_i^{t'} \quad \forall i \in V, k \in K, t' \in T \quad (9.24)$$

Route Duration

$$w_{n+1}^{kt'} - w_0^{kt'} \leq f \quad \forall k \in K, t' \in T \quad (9.25)$$

Delay

$$(w_i^{kt'} + m \cdot (t' - 1)) - (l_i + m \cdot (t - 1)) \leq z_i^t + (1 - v_i^{ktt'}) \cdot M_3 \quad \forall i \in N, k \in K, t \in T, t' \in T \quad (9.26)$$

Non-delivery

$$(w_i^{kt'} + m \cdot (t' - 1)) \leq (u_i + m \cdot (t - 1)) + (1 - v_i^{ktt'}) \cdot M_4 \quad \forall i \in N, k \in K, t \in T, t' \in T \quad (9.27)$$

Non-negativity and Binary Conditions

$$x_{ij}^{kt'} \in \{0, 1\} \quad \forall (i, j) \in A, k \in K, t' \in T \quad (9.28)$$

$$(9.29) \quad v_i^{ktt'} \in \{0, 1\} \quad \forall i \in N, k \in K, t \in T, t' \in T$$

$$(9.30) \quad w_i^{kt'} \geq 0 \quad \forall i \in V, k \in K, t' \in T$$

$$(9.31) \quad y_i^t \in \{0, 1\} \quad \forall i \in N, t \in T$$

$$(9.32) \quad z_i^t \geq 0 \quad \forall i \in N, t \in T$$

$$(9.33) \quad s_i^{kt'} \geq 0 \quad \forall i \in V, k \in K, t' \in T$$

The model will now be explained in detail starting with the objective function (9.14). The transportation cost which is composed of five different cost terms is minimized:

- The first part is the distance dependent cost factor multiplied by the distance travelled
- The second part is the fixed cost for all vehicles leaving the start depot and going to a customer
- The third part is the duration-dependent cost factor multiplied by the duration of the trip of all vehicles
- The fourth part is the penalty cost of non-delivery if the customer is not served
- The fifth part is the penalty cost for delay

Constraints (9.15) connect variables $v_i^{ktt'}$ and $x_{ij}^{kt'}$, by checking on which shipping day the demand is served. Since it is assumed that the shipping day t' for an order is always larger than the demand day t ($t' \geq t$), it is not possible to deliver an order too early which is a requirement according to [Figure 9.3](#). Consequently a constraint is not necessary to present early delivery.

Constraints (9.16) sum over all vehicles and shipping days to check whether the customer is served. The variable y_i^t assumes a value of 1 if the demand of customer i from period t is served. The next three constraints are the flow constraints. All vehicles should start at the origin depot and go either to a customer or the destination depot (9.17). Note that all vehicles are used on each shipping day, but a dummy tour is created for the vehicles not required as they can go straight to the destination depot without any cost. All vehicles should end at the destination depot on all shipping days (9.18) and flow conservation should also be respected for all customers in all shipping days (9.19).

The capacity of the vehicles should be respected (9.20), note that customer demands of more than one day can be served by one vehicle. Constraints (9.21)

set the variable loading times (service times) at the depot for each vehicle and shipping day, by summing over all deliveries loaded on a vehicle and multiplying with the loading factor. Similarly, constraints (9.22) set the variable unloading times (service times) at the customers where demands of more than one day can be served in one shipping day. Schedule feasibility is ensured with regard to traveling and service time (9.23). Time windows have to be respected for all nodes (9.24), i.e. the service at node i has to start within the time window, where $a_i^{t'}$ indicates the start of the time window and $b_i^{t'}$ the end of the time window. Maximum route duration is set for all vehicles in all shipping days (9.25).

Constraints (9.26) calculate the delay in minutes and set a value on z_i^t , where the constraints become ineffective if $v_i^{ktt'} = 0$. Note that $w_i^{kt'}$ and l_i are changed such that they are not dependent on the day but rather the minute, where m represents the minutes in a calendar day ($24 \cdot 60 = 1440$). This is done in order to calculate the linearly increasing delay, as represented in Figure 9.3. An example is provided to demonstrate the calculation.

Example: Customer i requests an order on day two ($t = 2$). It is assumed that the limit for delay (l_i) is at the end of the requested delivery day when the customer has closed. In this example the customer closes at 6 p.m. (18 hours after midnight), therefore $l_i = 18 \cdot 60 = 1080$. After this point in time, linearly increasing penalty cost is incurred. Assume that the order is finally delivered on day four ($t' = 4$) at 11 a.m., therefore: $w_i^{kt'} = 11 \cdot 60 = 660$. Then the delay is:

$$z_i^t = (660 + 1440 \cdot (4 - 1)) - (1080 + 1440 \cdot (2 - 1)) = 2460 \text{ minutes}$$

Constraints (9.27) set limits for non-delivery. It is not possible to serve customers when the upper end of the time window is reached. Finally there are non-negativity constraints and binary conditions (9.28)-(9.33). Note that the mathematical model is not solvable to optimality for the Frutado company in an acceptable time frame; therefore a heuristic solution approach like described in Section 9.2 has to be used.

Questions and Exercises

1. Some of the constraints in the model above are formulated by using the Big-M formulation, when using this formulation the constant M should have the lowest possible value in order to increase the computational efficiency. Find values for constants M_1 , M_2 , M_3 , and M_4 .
2. Why is it not necessary to use a subtour elimination constraint in the model above, like for the model presented in Section 9.1.2?

9.4.3 Extensions

In this section, two possible extensions to the model presented above will be described. Note that these extensions are also part of SAP APO and can therefore be incorporated into planning problems modeled with the software. The first extension is to allow for *multiple routes* per vehicle, as long as the maximum duration of the route is not violated. A vehicle can therefore serve some customers, return back to its DC and load new orders for other customers. Allowing for multiple routes per vehicle is very relevant for real-life problems. Methods for solving multi-trip VRP have been proposed in the literature, e.g. Zeng et al. (2008) solve the problem for a soft drink transportation company in Singapore. Following is a description of how the model above can be extended to include multiple routes. The idea is to introduce dummy depots, which the vehicles can return to and pickup more orders to deliver.

The following set is therefore added:

G dummy depots for multiple routes $g \in G$

The dummy depots need to be added to the set for all nodes:

V nodes $V = N \cup \{0, n + 1, \dots, n + g + 1\}$
 nodes 0 and $n + 1$ are the origin and destination depot as before
 nodes $n + 2$ to $n + g + 1$ are the dummy depots, depending on how many routes are allowed

The following variable needs to be added:

$p_i^{kt'}$ loading variable specifying the amount of load, just after servicing node i by vehicle k on shipping day t' [tons]
 $i \in V, k \in K, t' \in T$

The following changes need to be done. Non-negativity constraints need to be added for the new variable. Constraints (9.15), (9.17), and (9.18) need to be changed to account also for flow to/from dummy depots. Constraints (9.19) need to hold also for all dummy depots, such that all flow into a dummy depot equals the flow out of it. Constraints (9.21) need to be adjusted to account for loading times at the dummy depot. The constraints for capacity (9.20) need to be replaced by the following constraints:

Loading

$$p_j^{kt'} + \sum_{t \in T} d_j^t \cdot v_j^{ktt'} \leq p_i^{kt'} + (1 - x_{ij}^{kt'}) \cdot M \quad \forall i \in V, j \in N, k \in K, t' \in T \quad (9.34)$$

$$p_{n+g+1}^{kt'} = q^k \quad \forall g \in G, k \in K, t' \in T \quad (9.35)$$

$$p_i^{kt'} \leq q^k \quad \forall i \in V, k \in K, t' \in T \quad (9.36)$$

In constraints (9.34), a value is set on the variable $p_i^{kt'}$ after each demand is served, only if the arc is used. Note that the demand can be from any of the demand days, which is the reason for the summation. Note also that the end arc is only for the customers ($j \in N$), because the capacity at a dummy depot is set to full (9.35), and therefore more orders can be picked up. The capacity at the destination depot does not need to be restricted. Finally, constraints (9.36) ensure that the loading variable never exceeds the capacity of the vehicle.

The second extension that is described here concerns *returns planning*. In a supply chain it is of importance to consider the recycling and disposal of the products. Attempts are progressively being made to avoid throwing away products, thus extending the transportation network by a recycling and disposal network (Knolmayer et al. 2002). There have been some contributions to this problem in the literature, e.g. Privé et al. (2006) study this case for a company which is engaged in distributing soft drinks and in picking up empty recyclable cans and bottles. This is also very relevant to look into for the Frutado scenario, where the customers can send their empty bottles and packaging back to the DC for recycling. This problem is known in the literature as the VRP with simultaneous pickup and delivery (VRPSPD), as described in Section 9.1.3. Customers are associated with both demand and pickup, the demand is supplied from a single depot and the pickups are transported to the same depot.

Questions and Exercises

1. Formulate a model that extends the model given in Section 9.1.2, such that it also considers returns planning where orders for different customers are delivered, and empty bottles are also shipped back to the DC. Define the demand for pickup of empty bottles as λ_i . Hint: Use a loading variable specifying the amount of load just after servicing a customer, similar to the extension outlined above.
2. Given the data exhibited in [Tables 9.1-9.3](#) as well as in [Table 9.5](#), solve the planning model formulated in Question 1 by use of an MIP solver.

Table 9.5
Customer demand for
pickup of empty
bottles λ_i

	customer <i>i</i>					
	1	2	3	4	5	6
λ_i	3	0	1	2	1	1

9.5 Implementation and Integration with Deployment

For the implementation in SAP APO, TLB is first run, followed by TP/VS. In TLB planning the confirmed transport quantities from Deployment are utilized and TLB shipments created. TP/VS then utilizes the TLB shipments. In TP/VS a transportation plan for one week is generated. Since transportation is not planned for weekends, this leads to a five day planning horizon. Transportation plans are generated for each DC, orders are assigned to trucks, and daily routes are generated. Detailed decisions are made, including timing of activities in a continuous time representation and the sequence of customer deliveries. The results are planned shipments.

For the Frutado company it is relevant to look at the *hierarchical integration* between Deployment and TP/VS. Transportation planning is a detailed planning method, whereas deployment planning is made on a more aggregated level. Vehicle resources are used for detailed TP/VS to map the capacity and availability of vehicles. In Deployment, abstract transportation capacities are used for aggregated planning to ensure feasibility of the Deployment plan. Transportation resources used in Deployment are bucket resources, meaning that each day a certain amount of an abstract capacity is available.

In addition, there is a difference of the capacity usage in Deployment and TP/VS. In Deployment, the planning is on a higher level, and the vehicle routes are not known. Consequently it is assumed that a delivery run includes only one customer. In contrast to TP/VS; where several customers are served in one route. This can lead to various capacity related problems. Some research has been done on these aggregation-disaggregation issues, see for example the article on distributed decision making by Schneeweiss (2003). He explains the interactive process of decision making on different planning levels, in which a higher level can anticipate the lower levels possible reaction. In order to align both modules in this respect, it is possible to provide feedback in form of *correction factor* from TP/VS, to improve the anticipation of the required transportation capacity in the Deployment module.

Explaining the integration in greater detail is best done by looking at an example. [Figure 9.4](#) shows an example of the difference between the routes in Deployment and TP/VS.

In Deployment the vehicle routes are not known, and therefore single legs are used. The capacity usage in Deployment is calculated by multiplying the duration of each leg by the corresponding customer demand, that is:

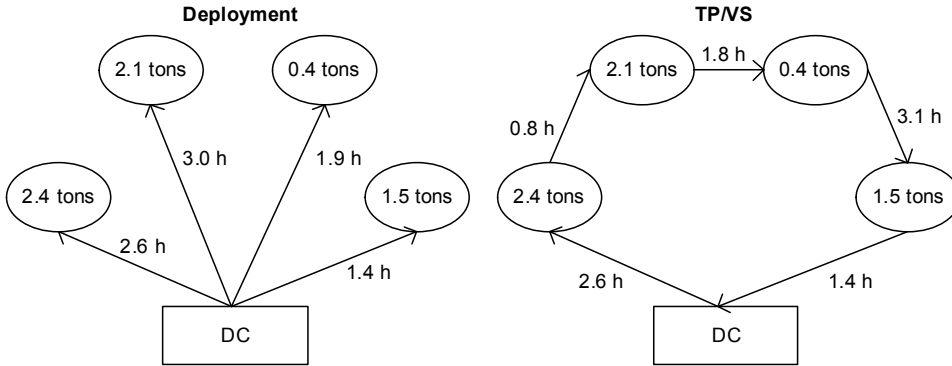


Figure 9.4
Routes used for calculation of capacity usage in Deployment and TP/VS

$$2.6 \cdot 2.4 + 3.0 \cdot 2.1 + 1.9 \cdot 0.4 + 1.4 \cdot 1.5 = 15.4 \text{ tons} \cdot \text{hours}$$

Note that only the time to the customer is considered, not the time of the trip back. In TP/VS the capacity usage is calculated by multiplying the capacity of the truck by the total duration of the route. In this example a 9 ton truck is used to deliver to the same four customers, and the trip takes 9.7 hour, the capacity usage is therefore:

$$9.7 \cdot 9 = 87.3 \text{ tons} \cdot \text{hours}$$

A correction factor can now be calculated which can be used for integrating the planning modules. In this example the correction factor is calculated as:

$$15.4/87.3 = 0.18$$

The correction factor is multiplied by the available TP/VS capacity, resulting in the capacity that should be used for Deployment. The bucket capacity of the transportation resource is adjusted. After each week the capacity is adjusted. It is important to check the capacity and update it regularly since the customer orders vary in each week. Choosing too small correction coefficient might cause problems with deriving a feasible solution in Deployment, in TP/VS on the other hand a higher degree of punctuality is obtained. A larger coefficient leads to smaller shortages observed through Deployment, and the vehicle usage in TP/VS will be better but some non-deliveries might occur in TP/VS. Note that SAP APO does not automatically perform this feedback mechanism from TP/VS to Deployment. Therefore the calculations of the correction factor have to be carried out manually.

9.6 TP/VS Learning Units

9.6.1 Overview

The TP/VS learning units are split into four broad themes

1. Master data
2. Model building
3. Planning run
4. In-depth stream

The first three units include essential steps for implementing the TP/VS module of the Frutado supply chain. The first theme is dedicated to explain the maintenance of relevant master data for TP/VS. Details on building the model are presented in the second theme. In the third theme, the learning units aim at showing how to run the planning, how to analyze the results, as well as how to hierarchically integrate TP/VS with Deployment. In order to cover other important features of the TP/VS module, an in-depth stream section is presented in the fourth theme as a group of supplementary TP/VS learning units. [Figure 9.5](#) shows how the learning units are further split into sub-learning units.

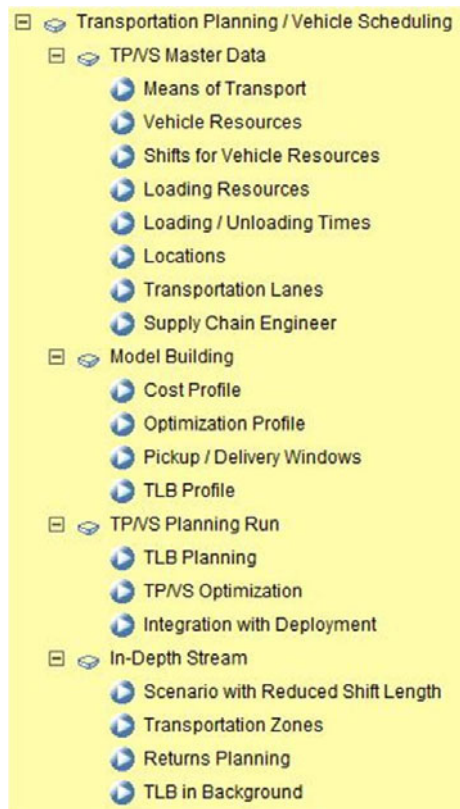


Figure 9.5
TP/VS outline
© Copyright 2011. SAP
AG. All rights reserved

The position of the TP/VS module and its data exchange with the other planning modules for the Frutado company is illustrated in Figure 9.6.

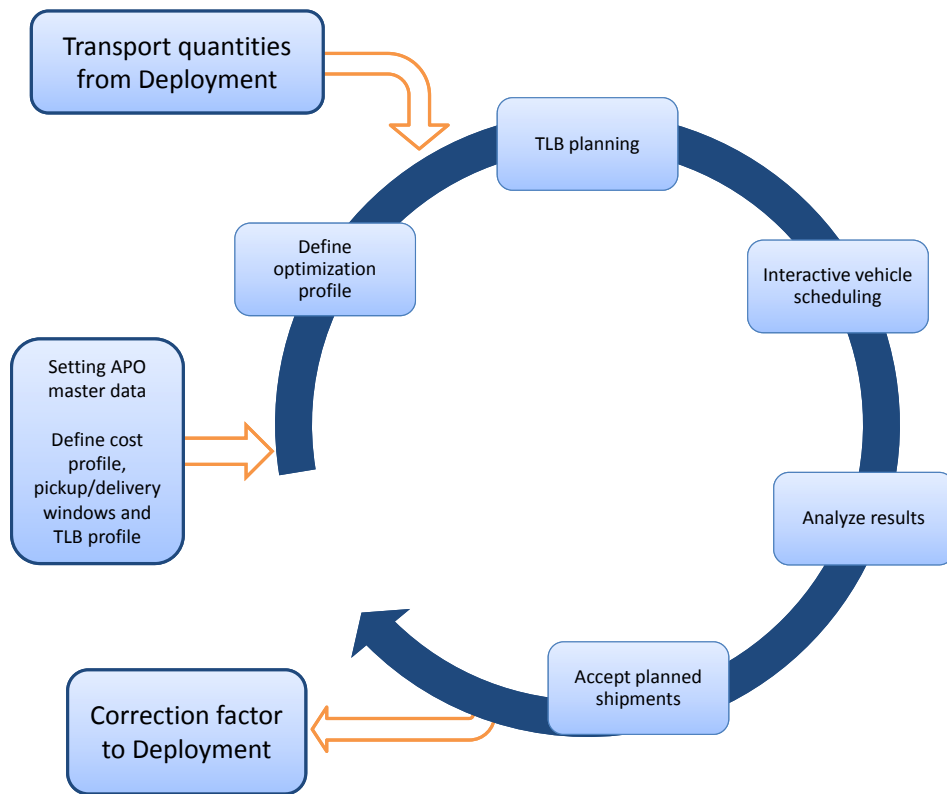


Figure 9.6
Position of the TP/VS module and its data exchange (Planning cycle)

The first step is to set the relevant master data for TP/VS which remain unchanged during the planning. Then the model is built up; including the definition of the cost profile, pickup/delivery windows and TLB profile. Note that the optimization profile is in the loop because it is adjusted for different planning runs. The transport quantities from Deployment are utilized in the TLB planning. Then interactive scheduling is carried out, its results are analyzed and the planned shipments accepted. Finally, a feedback in form of a correction factor is provided back to Deployment. The next two sections briefly describe what can be observed when working through the different learning units.

9.6.2 Basic Stream

TP/VS Master Data

This section describes the TP/VS relevant master data, which is summarized in Table 9.6. In the first learning unit *means of transport* are described, which refers to the types of vehicles used. Each type represents a family of vehicles with equivalent cost structures, travel characteristics and geographical availability. For the Frutado company, three different types of trucks are used,

Master Data	Number
means of transport	3
vehicle resources	20
loading resources	3 at each DC (9) 1 at each customer (60)
locations	3 DCs 60 customers
transportation lanes	1266

Table 9.6
Required master data

indicating different means of transport. In order to model road distances as realistically as possible, a multiplication factor for the Euclidean distance is set for the means of transport. It establishes a proportion between the linear distance of two locations and the actual distance covered by the vehicle. This factor together with the average speed of the means of transport is used to calculate the transportation duration. A smaller factor is chosen for the smaller means of transport. This shortens the distance for smaller means of transport, which is reasonable because smaller vehicles can drive on a wider variety of roads (like bridges with weight restrictions). Note that this multiplication factor and average speed are not incorporated as parameters into the model presented in Section 9.4.2, because the transportation duration is used directly as an input.

The next learning unit describes *vehicle resources*, which are concrete trucks by which the transportation between locations is executed. Any number of vehicle resources can be assigned to a means of transport; however a particular vehicle resource can only be assigned to one means of transport. For the Frutado company vehicle resources are assigned to different DCs, both for transportation between the DCs, and also for transportation to the customers. The route should start and end at the corresponding DC. The availability of the vehicle resources is modeled in SAP APO by assigning the vehicle resources to shifts, as shown in the third learning unit.

The next learning units are provided on loading resources and loading/unloading times. Loading resources are assigned to locations, where the opening hours of the locations are modeled by the settings of their loading resources. The loading resources are acting as constraints, since their time windows have to be respected and loading/unloading can only take place during the opening hours. In the optimization model in Section 9.4.2, constraints (9.24) describe these time windows.

For the TP/VS module it is relevant to look at two types of *locations*; DCs and customers. The precise address of all customers and DCs are needed in order to determine their geographical latitude and longitude, which is important for calculation of transportation distances and durations. A

priority for the customer locations is also maintained according to the *ABC* classification.

Transportation lanes, shown in the next learning unit, represent the connections between different locations of the supply chain and can be thought of as the road between two locations. An example of the settings for a transportation lane is shown in [Figure 9.7](#).

MTr	Means of Transport	Start date	End Date	All Prods	Aggr. Plng	DetId Plng	Trsp. Cal	Fix Duratn	Trsp. D...	Ret.Period	Fix Dist.	Trsp. Dist.	Unit	Precisi...	Transpntn Cos
Z_f	Z_FRU_BIG4	01.01.197	31.12.210	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	FRU_CAL	<input type="checkbox"/>	1:30	:10	<input type="checkbox"/>	82,534	KM		0,01
Z_FR	Z_FRU_SM4	01.01.197	31.12.210	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	FRU_CAL	<input type="checkbox"/>	1:08	:10	<input type="checkbox"/>	79,360	KM		0,01

Figure 9.7
Settings for a transportation lane
© Copyright 2011. SAP AG.
All rights reserved

Each transportation lane is defined by its source and destination locations, as well as the means of transport that can use the lane. For the Frutado company, the transportation processes have been divided up into four delivery areas. That is one area for transportation between DCs and three areas for transportation between DCs and its corresponding customers. These areas are planned separately, and therefore only lanes between locations in the delivery area should be created, e.g. a lane between DC 1 and customer 60 is not relevant. Note that lanes have to be created in both directions depending on the direction of the traffic. For example a lane is created from DC 1 to customer 1, and also from customer 1 to DC 1. This leads to a total of 1266 lanes. Transportation zones can be used to reduce the data maintenance effort of using transportation lanes; this is shown in a learning unit in the in-depth stream section.

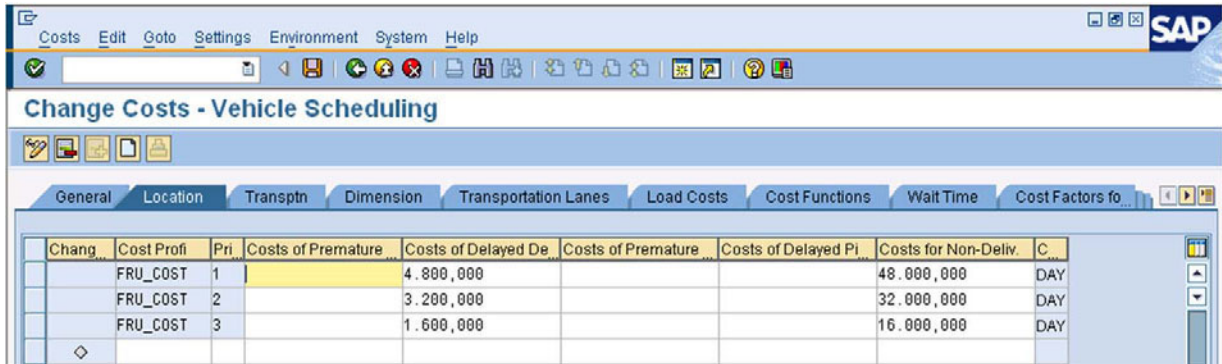
SAP APO can automatically calculate the transportation duration and distance for each means of transport assigned to the lane. Those are calculated based on geographical coordinates, the multiplication factor and the speed of the means of transport. Note that by connecting a geographic information system through an internet graphic server, it is possible to calculate the duration and distance with exact road detail. Using the distance calculated by geo-coding will provide better results compared to using the straight distance (Dickersbach 2006). However, even though the distances are known exactly the duration can vary depending on traffic and weather conditions (Knolmayer et al. 2002).

Finally, the last learning unit shows how the Supply Chain Engineer feature of SAP APO works, which is a very convenient method for accessing

and visualizing data of different components of the supply chain and the TP/VS module accordingly.

Model Building

For building up the model, different profiles have to be maintained, this is shown in the learning units of this theme. Firstly, a cost profile is defined to maintain the common cost terms as well as penalty costs, see an example in [Figure 9.8](#).



The screenshot shows the SAP APO 'Change Costs - Vehicle Scheduling' window. The window title is 'Change Costs - Vehicle Scheduling'. The menu bar includes 'Costs', 'Edit', 'Goto', 'Settings', 'Environment', 'System', and 'Help'. The toolbar contains various icons for file operations and navigation. Below the toolbar, there are several tabs: 'General', 'Location', 'Transpnt', 'Dimension', 'Transportation Lanes', 'Load Costs', 'Cost Functions', 'Wait Time', and 'Cost Factors fo'. The 'Location' tab is active, and it displays a table with the following data:

Chang...	Cost Profi	Pri	Costs of Premature	Costs of Delayed De	Costs of Premature	Costs of Delayed Pi	Costs for Non-Deliv. C.	C.
	FRU_COST	1		4.800,000			48.000,000	DAY
	FRU_COST	2		3.200,000			32.000,000	DAY
	FRU_COST	3		1.600,000			16.000,000	DAY

Figure 9.8
Settings in a cost profile
for TP/VS

© Copyright 2011. SAP AG.
All rights reserved

The figure shows how the location-relevant cost is defined, which includes lateness and non-delivery costs of an order based on the priority of the location. Pickup/delivery windows are defined in the next learning unit. This determines whether the lateness and non-delivery costs are incurred or not. Note that these pickup/delivery windows, as they are defined in SAP APO, are not related to the opening hours of the customer. As described earlier, the opening hours of the customers are modeled by the settings of their loading resources.

In order to define the scope of TP/VS for different planning scenarios, different *optimization profiles* can be maintained as shown in the next learning unit. The optimization profile contains restrictions regarding planning version, cost profile, optimizer runtime, horizons, order types, locations, and vehicle resources, included in a certain TP/VS planning scenario. Finally, a weight-dependent TLB profile is created which contains the weight limits for the pallets.

Planning Run

The first learning unit in this theme shows how to execute TLB planning for one of the customers on a specific demand day. Then a learning unit shows how the interactive vehicle scheduling is run. Shipments are planned on the basis of costs. The routes for the vehicles as well as transportation dates/times are established. During a planning run SAP APO shows the current best solution and the solution process on a graph. When the runtime limit is reached, the best solution is presented. This is shown in [Figure 9.9](#).

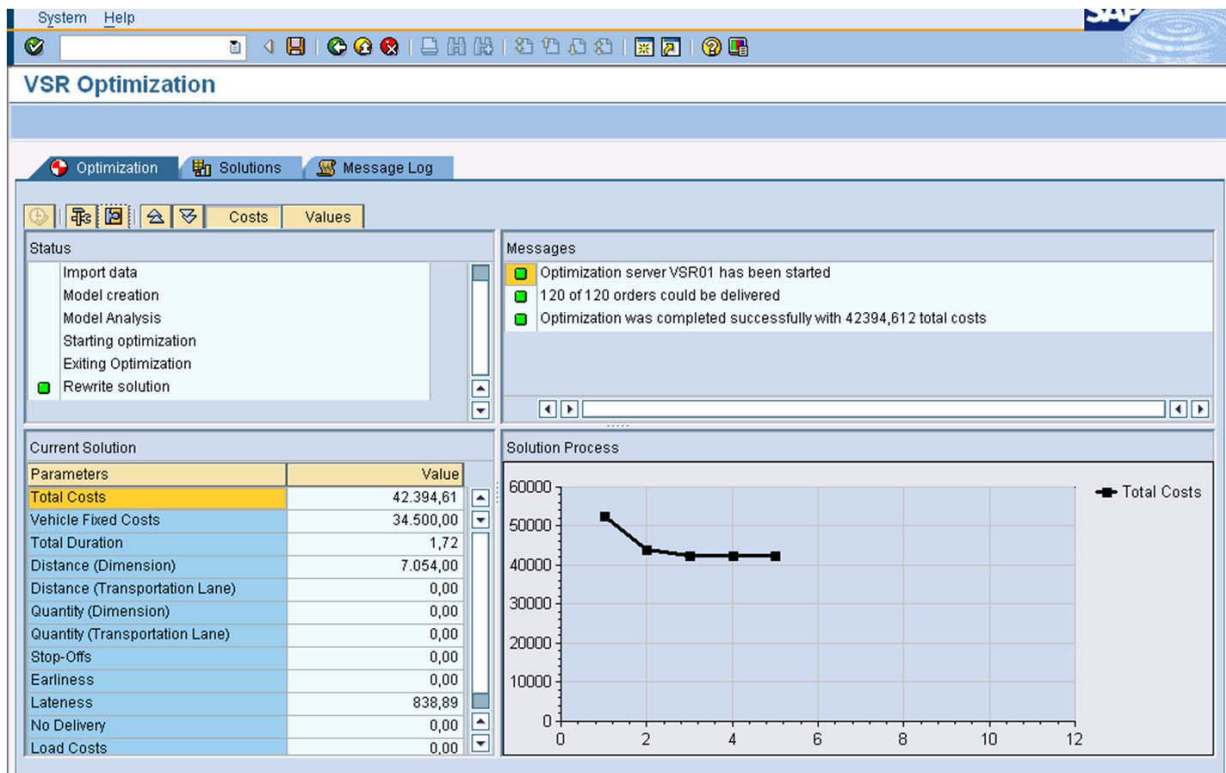


Figure 9.9
Interactive vehicle scheduling
© Copyright 2011. SAP AG.
All rights reserved

Analyzing the results is of great importance to see if the scheduled shipments meet the expectation of the planner and to check if the input data has been clean. After analyzing the results, the planned shipments can be accepted in SAP APO, and the results are processed further. Detailed travel times and distances of a route can also be displayed based on a road network. This requires an external geographic information server connected to the SAP APO system. Finally, the last learning unit in this theme describes the hierarchical integration between Deployment and TP/VS, as described in Section 9.5.

Questions and Exercises

The following questions should be answered while working through the learning units.

1. Explain how shifts are used for vehicle resources for the Frutado company.
2. What options are chosen for the TP/VS optimization profile?
3. Which master data object is the TLB profile assigned to?

9.6.3 In-Depth Stream

Theme four includes in-depth stream learning units. These in-depth stream learning units are carefully chosen such that the students can learn alternative modeling methods in SAP APO and also learn about practical issues, such as returns planning. Firstly, a learning unit is provided on a planning run with reduced shift length. In this learning unit a test is made where the maximum duration of a trip is set to eight hours, instead of ten hours as before. The reason for this may for example be some new working time legislation.

Then a learning unit on *transportation zones* is provided which shows an alternative method of modeling master data. Instead of creating transportation lanes between all locations in a delivery area transportation zones can be used. The main advantage of using transportation zones is that the amount of master data which needs to be processed can be reduced considerably, as the system only dynamically generates the necessary lanes for planning.

The third learning unit in the in-depth stream section shows how returns planning can be integrated into the transportation planning of the Frutado scenario. Please refer to a discussion on returns planning in Section 9.4.3, as a possible extension to Frutado's VRP. An example is created in the learning unit, to demonstrate the returns planning function in SAP APO. A route is planned for each vehicle, where orders for different customers are delivered, and the empty bottles are shipped back to the DC. In this way forward and return logistics are integrated. Finally a learning unit is provided on how to implement TLB in the background, rather than implementing it for each customer separately.

Questions and Exercises

The following questions should be answered while working through the in-depth stream section.

1. Compare the results for planning with 10 hours shift length and 8 hours shift length.
2. Explain the steps needed in order to model transportation zones.
3. Compare the settings of the TP/VS optimization profile for the basic planning problem and for returns planning.

Bibliography

- Alvarez-Valdes, R.; Parreno, F.; Tamarit, J. M. (2005) *A tabu search algorithm for the pallet loading problem*, Or Spectrum, vol. 27, 43–61
- Ambrosino, D.; Sciomachen, A. (2007) *A food distribution network problem: a case study*, IMA Journal of Management Mathematics, vol. 18, no. 1, 33–53

- Bodin, L.; Golden, B. (1981) *Classification in vehicle routing and scheduling*, Networks, vol. 11, no. 2, 97–108
- Bräysy, O.; Dullaert, W.; Gendreau, M. (2004) *Evolutionary algorithms for the vehicle routing problem with time windows*, Journal of Heuristics, vol. 10, no. 6, 587–611
- Bräysy, O.; Gendreau, M. (2005a) *Vehicle routing problem with time windows, Part I: Route construction and local search algorithms*, Transportation Science, vol. 39, no. 1, 104–118
- Bräysy, O.; Gendreau, M. (2005b) *Vehicle routing problem with time windows, Part II: Metaheuristics*, Transportation Science, vol. 39, no. 1, 119–139
- CAPE Systems (2011) *Homepage*, URL <http://www.capesystems.com/>, date: April, 2011
- Clarke, G.; Wright, J. V. (1964) *Scheduling of vehicles from a central depot to a number of delivery points*, Operations Research, vol. 12, no. 4, 568–581
- Coffman, E. G.; Garey, M. R.; Johnson, D. S. (1997) *Approximation algorithms for bin packing: a survey*, in: D. Hochbaum (Ed.) *Approximation Algorithms for NP-hard Problems*, chap. 2, PWS Publishing, 46–93
- Cordeau, J.-F.; Desaulniers, G.; Desrosiers, J.; Solomon, M. M.; Soumis, F. (2002) *VRP with time windows*, in: P. Toth; D. Vigo (Eds.) *The Vehicle Routing Problem*, chap. 7, SIAM Monographs on Discrete Mathematics and Applications, Philadelphia, 157–193
- Dantzig, G. B.; Ramser, J. H. (1959) *The truck dispatching problem*, Management Science, vol. 6, no. 1, 80–91
- Dantzig, G. B.; Wolfe, P. (1960) *Decomposition principle for linear programs*, Operations Research, vol. 8, no. 1, 101–111
- Desrochers, M.; Lenstra, J. K.; Savelsbergh, M. W. P. (1990) *A classification scheme for vehicle routing and scheduling problems*, European Journal of Operational Research, vol. 46, no. 3, 322–332
- Desrosiers, J.; Dumas, Y.; Solomon, M. M.; Soumis, F. (1995) *Time constrained routing and scheduling*, in: M. O. Ball; T. L. Magnanti; C. L. Monma; G. L. Nemhauser (Eds.) *Handbooks in Operations Research and Management Science, 8: Network Routing*, chap. 2, Elsevier Science, Amsterdam, 35–139
- Dickersbach, J. T. (2006) *Supply Chain Management with APO*, Springer, Berlin, 2nd ed.
- Dyckhoff, H. (1990) *A typology of cutting and packing problems*, European Journal of Operational Research, vol. 44, no. 2, 145–159

- Fleischmann, B.; Gietz, M.; Gnutzmann, S. (2004) *Time-varying travel times in vehicle routing*, Transportation Science, vol. 38, no. 2, 160–173
- Gottlieb, J. (2006) *Solving real-world vehicle scheduling and routing problems*, URL <http://www.winfo.tu-bs.de/projekte/snlm07/gottlieb.pdf>
- Gottlieb, J.; Eckert, C. (2005) *Solving real-world vehicle scheduling and routing problems*, International Scientific Annual Conference Operations Research, Bremen, Germany
- Hombberger, J.; Gehring, H. (1999) *Two evolutionary metaheuristics for the vehicle routing problem with time windows*, Information Systems and Operational Research, vol. 37, 297–318
- Jabali, O.; Van Woensel, T.; de Kok, A. G.; Lecluyse, C.; Peremans, H. (2009) *Time-dependent vehicle routing subject to time delay perturbations*, IIE Transactions, accepted April 2009
- Jünger, M.; Reinelt, G.; Rinaldi, G. (1995) *The traveling salesman problem*, in: M. O. Ball; T. L. Magnanti; C. L. Monma; G. L. Nemhauser (Eds.) *Handbooks in Operations Research and Management Science, 7: Network models*, chap. 4, Elsevier Science, Amsterdam, 225–330
- Knolmayer, G.; Mertens, P.; Zeier, A. (2002) *Supply Chain Management Based on SAP Systems – Order Management in Manufacturing Companies*, Springer, Berlin, 1st ed.
- Laporte, G.; Osman, I. H. (1995) *Routing problems: A bibliography*, Annals of Operations Research, vol. 61, no. 1, 227–262
- Pallet Manager (2011) *Homepage*, URL <http://www.packyourpallet.com/>, date: April, 2011
- Privé, J.; Renaud, J.; Boctor, F.; Laporte, G. (2006) *Solving a vehicle-routing problem arising in soft-drink distribution*, Journal of the Operational Research Society, vol. 57, no. 9, 1045–1052
- Quick Pallet Maker (2011) *Homepage*, URL <http://www.koona.com/>, date: April, 2011
- Ram, B. (1992) *The pallet loading problem: A survey*, International Journal of Production Economics, vol. 28, no. 2, 217–225
- Schneeweiss, C. (2003) *Distributed decision making - a unified approach*, European Journal of Operational Research, vol. 150, no. 2, 237–252
- Toth, P.; Vigo, D. (2002a) *An overview of vehicle routing problems*, in: P. Toth; D. Vigo (Eds.) *The Vehicle Routing Problem*, chap. 1, SIAM Monographs on Discrete Mathematics and Applications, Philadelphia, 1–26

-
- Toth, P.; Vigo, D. (Eds.) (2002b) *The Vehicle Routing Problem*, SIAM Monographs on Discrete Mathematics and Applications, Philadelphia
- Van Woensel, T.; Kerbache, L.; Peremans, H.; Vandaele, N. (2008) *Vehicle routing with dynamic travel times: A queueing approach*, European Journal of Operational Research, vol. 186, no. 3, 990–1007
- Wäscher, G.; Hausner, H.; Schumann, H. (2007) *An improved typology of cutting and packing problems*, European Journal of Operational Research, vol. 183, no. 3, 1109–1130
- Zeng, L.; Ong, H. L.; Ng, K. M.; Liu, S. B. (2008) *Two composite methods for soft drink distribution problem*, Advances in Engineering Software, vol. 39, no. 5, 438–443

Final Remarks

Hartmut Stadtler¹

¹ University of Hamburg, Institute for Logistics and Transport, Von-Melle-Park 5, 20146 Hamburg, Germany

This chapter comprises two sections. First, we will discuss issues related to the implementation (strategies) of an APS. Second, we will point out the strengths and limitations of (current) APS.

10.1 Implementation of an APS

Having worked through the previous chapters and the associated learning units you now probably will agree that making use of an APS is nothing you can do as an “extra” to your daily work – it needs (nearly) full-time attention of (at least) one employee within a company. And even then this employee should be in charge of just one or two modules of an APS in order to be able to unfold its full potential by checking and maintaining the input data, adapting the module to changes in the (business) environment, creating alternative solutions and analyzing and discussing results of the planning run(s).

Using an APS is really similar to playing a Stradivari (as stated in the Introduction): it needs permanent training and exercises – and to start with a good teacher (in the case of an APS it usually is an expert consultant who gives advice regarding the overall software concept and supports its implementation).

The good news is that you do not need to implement all the modules of an APS at once (like in the Frutado case) in order to achieve substantial benefits. Often companies deliberately concentrate on only a few modules which support their business best and will start with those modules in which the relation of benefits compared to efforts is the largest. Due to the close links of some modules regarding the data flows (directives etc.) some clusters of modules seem to be a “natural” choice. Note that the choice of clusters

may depend on the type of industry and the complexity of decision problems, too. Although the following discussion assumes the implementation of the SAP APO it should hold for most other APS (but probably with different names for the modules):

A good choice to start with (for any cluster) is demand planning (DP). Reliable demand forecasts are the cornerstone of any planning activity and are needed for make-to-stock processes (see the decoupling point). Starting with the DP module incurs a further advantage: its logic and basic functionality is similar to the one known from an ERP system. This also applies to the global Available-to-Promise (ATP) module. This characteristic should allow users to get acquainted with its use and to grasp the additional functionality offered by an APS quickly.

Next, one might either add supply network planning (SNP) or production planning / detailed scheduling (PP/DS) to form an initial cluster. In a supply chain (SC) where there is much flexibility regarding the allocation of production quantities among alternative production plants (in general: locations) and where there are reliable medium-term demand forecasts, the greatest benefits should be gained by implementing SNP. On the other hand, if the allocation of products to plants is fixed and the resource utilization becomes crucial with many potential bottlenecks within a plant, together with significant sequence dependent setup times and costs, the choice would be to add PP/DS first.

Having implemented SNP or PP/DS one might consider the introduction of global ATP. Without medium- or short-term production planning global ATP can only refer to the inventory position of products – and thus will not be different from an ATP that comes along with an ERP system.

Deployment is based on the results of SNP. Implementing the deployment module makes sense for companies processing and controlling its own distribution system (at least some warehouses).

TP/VS allows a SC to create detailed plans for the transport of goods (both from suppliers and to customers) executed either by a SC's own fleet or by the services provided by a third party logistics provider (3PL).

Some companies may primarily look for an enhancement of the rather weak planning functions of an existing ERP system. Here, the planning capabilities and graphical user interfaces of PP/DS and TP/VS may be very attractive in combination with an ERP system.

Having installed a cluster of APS modules – how are the remaining planning tasks of the SCP matrix taken into account? There are two reasons for not making use of an APS module: either the planning tasks are of minor importance, are quite easy to perform and thus can be left to manual planning (perhaps supported by spreadsheets or the functionality of an ERP system) or the functionality of an APS module is insufficient resulting in the implementation of an individual software application (e.g. detailed scheduling) which then can be connected to an APS (see Kallrath and Maindl 2006) or directly to an ERP system.

10.2 Evaluation of APS

Before implementing an APS one should be aware of the strengths and limitations of (current) modules. [Table 10.1](#) highlights the most important points. DP will yield a well-structured and transparent forecasting process. The way to aggregate ex-post data regarding the granularity of time, geographic regions and product variants has to be defined. Automated forecasts will be generated and the (ex-post) forecast errors will be monitored. Employees can then focus on those forecasts which show an unusual behavior (highlighted by “alerts”).

DP is advantageous in case a great number of forecasts (>100 items) has to be made regularly. To improve forecast quality usually the input of human experts is still advisable. If there were only a handful of demand forecasts these should be key for a company (like in the energy sector) and probably will justify the most powerful forecasting models available today (like neural networks).

SNP will take into account the whole SC at an aggregate level of detail. The sources of flexibility will be exploited in order to optimize a given objective (function), like costs or contribution margin. Here, the room for optimization will be the largest by making the best use of a SC’s scarce resources. However, we should be aware that in the medium-term there may be uncertainties regarding the future (due to unknown actions of competitors and preferences of customers). Hence, different scenarios should be explored.

PP/DS will support plant managers by creating feasible schedules for the various (bottleneck) resources. Objectives considered may be costs but also time-oriented measures (like the sum of due date violations).

Meta-heuristics should be applied provided there are only a few inter-related bottleneck resources. Accordingly, production plants should be divided into manageable units (often called segments) with as little inter-relations regarding the flow of materials as possible. If there is still a great number of potential bottleneck resources within a segment simple priority rules are a last resort instead of a metaheuristic.

Automatic scheduling will be advantageous especially for scarce, automated, capital intensive production processes. However, one should bear in mind that lot-sizing and scheduling decisions are not performed simultaneously. This may be a drawback if a product’s lot sizes are not fixed (due to technical reasons) and hence may be chosen from a wide range of values (like for some flow lines).

Module	What you can expect	... and what you should not expect
Demand Planning	A well-structured, transparent forecasting process	Simple forecasting models alone will not reduce the forecast error significantly; the input of human expertise is still advisable

Continued on Next Page...

Module	What you can expect	...and what you should not expect
Supply Network Planning	Discovering bottlenecks, best use of scarce resources	The full potential of medium-term planning is not achieved with just one optimization run (you need to check several scenarios)
Production Planning / Detailed Scheduling	Quick rescheduling, consideration of sequence dependent setup times, advantageous for automated production processes	Manual production systems and those with a high degree of uncertainty will hardly benefit from PP/DS; No simultaneous lot-sizing and scheduling
Global Available to Promise	Customer requests are checked against available inventories and planned quantities; rule-based order promising	Reliable promises of due dates and quantities require stable plans
Deployment	More detailed distribution plans than in SNP considering customer orders; potential supply shortages are assigned to customers	There will be no detailed scheduling of resources (e.g. trucks)
Transportation Planning / Vehicle Scheduling	Reduced costs of transport, consideration of time windows	Complicated restrictions, like regulations of drivers working hours, may be difficult to model

Table 10.1
Advantages and
limitations of APS
modules

The quotation of reliable delivery dates is a key for most customers (even more important than short delivery lead times). These will be generated by global ATP. If a SC has little standard orders from customers and hence orders are placed at short notice and require an immediate answer regarding its fulfillment, global ATP will be very valuable. Still, the reliability of delivery dates will largely depend on the stability of production plans.

The virtue of deployment will be observed in situations in which the SNP plans for distribution are not detailed enough or outdated quickly, assuming that the re-planning interval of deployment is shorter. An appealing feature might be allocation rules (e.g. fair share rules) in case of supply shortages. However, one should bear in mind that deployment plans are based on a periodic time frame and quantities to transport – but not on individual truck movements.

This is the task of TP/VS. Here, feasible schedules for trucks will result while taking into account certain objectives like minimizing total costs or the total distance traveled. Difficult constraints, like restrictions on working hours and breaks of drivers, may be difficult to implement.

Note that the use of an APS is more than a mere addition of the advantages of single modules. Its structure – based on the principles of hierarchical planning – results in a consistent solution for the whole SC covering both medium- and short-term perspectives as well as aggregate to detailed decisions. With the help of directives and feedback SCs will avoid getting stuck in local optima (of single business functions). Also, all planning processes will be

documented and transparent. Responsibilities for certain planning tasks will be assigned to respective positions in an organization's hierarchy. Hence, the risk that upper level plans are "forgotten", i.e. not acted upon on their way down to execution on the shop floor or distribution system, is drastically reduced.

"Planning" is an important task of management for being prepared for the future and to strengthen the competitive position of a company or SC. Similarly important is that plans are executed as planned. However, execution is not part of an APS. Instead, plans are transmitted to an ERP system for execution.

Furthermore, special software products are available to trace and track consignments on their way to customers. Also, supply chain event management (SCEM) has to be mentioned here which allows to check the flow and timing of goods in the SC and to compare these with pre-defined conditions (expected events). Delayed expected events – like the late arrival of goods at distinct locations in the SC - will be reported to the responsible dispatcher(s) who then can take action, such that the consignment will still reach the customer in time (for more information see Knolmayer et al. 2009, pp. 152).

To conclude, there is no final, general answer regarding the usefulness of APS, because APS are open systems capable of incorporating new planning concepts, modeling features, solution algorithms and user interfaces. Hence, new findings in research and good business practices should find their way into future developments of APS.

Bibliography

Kallrath, J.; Maindl, T. (Eds.) (2006) *Real Optimization with SAP APO*, Springer, Berlin, 4th ed.

Knolmayer, G.; Mertens, P.; Zeier, A.; Dickersbach, J. T. (2009) *Supply Chain Management Based on SAP Systems: Architecture and Planning Processes*, Springer, Berlin, 1st ed.

Index

- ABC classification, 14
- Abstraction, 123
- Accounting costs, 124
- Additive trend (demand) model, 79
- Advanced planning system, 14
 - strengths and limitations, 291
- Aggregate
 - planning, 110
 - transportation capacity, 234
- Aggregation, 23, 72, 112, 123
 - disaggregation error, 218, 239
- Anticipation, 25
- APS, *see* Advanced planning system
- ATD, *see* Available-to-Deploy
- ATP, *see* Available-to-Promise
- Automatic model selection
 - procedure 1, 93
 - procedure 2, 93
- Available-to-Deploy, 241
 - issues, 241
 - receipts, 241
- Available-to-Promise, 32, 195, 240
 - ATP-quantity, 196
 - product availability check, 204
 - rules-based ATP, 205
 - simulation, 210
 - time series, 203
- Backward scheduling, 118
- BAdI, *see* Business Add-In
- BAPI, 139
- Basic deployment model, 220, 222,
 - 224, 225, 227, 228, 230
- Batch, 19
- Big bucket model, 115
- Bin packing, 250
- Bottom-up, 72
- Business Add-In, 143
- Calculation types, 93
- Capacity, 19
- Causal
 - methods, 92
 - models, 72
- Central planning, 122
- Characteristic value combination, 50
- Chromosome, 162
- Column generation, 259
- Computational efforts, 120
- Condition technique, 204, 212
- Consumer good, 18
 - industry, 18
- Coordination, 22
- Core interface, 41
- Correction factor, 274
- Cross
 - docking, 19
 - shipping, 12
- Cross-period lot sizes, 145
- Customer, 12
 - order, 195
 - priority, 227
 - segment, 200
- Cyclic scheduling, 154
- Data
 - configuration, 172
 - master, 41, 45, 136, 139, 171
 - transactional, 42, 50
 - view, 53

- warehouse, 139
- Decoding, 163
- Decomposition, 22
- Decoupling point, 31
 - influence on demand fulfillment, 196
- Delivery point, 12, 20
- Demand fulfillment, 32, 197
 - demand supply matching, 198
 - order promising, 197
 - batch, 198
 - real-time, 198, 199
 - shortage planning, 198
- Demand planning, 36
- Deployment, 37, 201, 209
 - horizon, 240
 - planning, 239
 - planning attributes, 223
- Deployment solution methods
 - fair-share, 243
 - push rules, 244
- Deterministic simulation, 138
- Directives, 135
- Disaggregation, 25
 - error, 218, 239
 - routine, 135
 - rules, 72
- Distribution center, 11, 19
- Divergent product structure, 19
- Duration of storage, 131

- Effective demand, 24
- Encoding, 163
- ERP system, 139, 196, 201
- Error tracking signal, 70
- Evolutionary
 - algorithm, 260
 - local search, 260
- Ex-post-simulation, 76
- Exponential smoothing
 - (Brown's) double, 84
 - Adaptive-response-rate single, 83
- Extreme solution, 131

- Fashion-4-You, 111
- Feed-forward-bottom-up, 142

- Feedback, 22
- Filling line, 11
- Fitness value, 163
- Flow line, 157
- Forecast, 18, 67, 71, 72
 - collaborative, 73
 - composite, 92
 - consensus-based, 73
 - error, 69
 - judgmental, 73
 - model, 91
 - profile, 97
 - strategy, 91
- Forecasting, 32
- Forecasting method
 - causal, 36
 - statistical, 36
- Frozen horizon, 26

- Gene, 162
- GIS system, 137
- Global ATP, *see* Global available-to-promise
- Global available-to-promise, 39, 195

- Hard constraint, 267
- Heuristic, 118
 - construction, 158, 259
 - improvement, 158, 259
 - nearest neighborhood, 158
- Hierarchical integration, 274
- Hierarchical planning, 21
 - system, 21, 23
- Hierarchy, 47, 71
- Holt's method, 85

- Implementation, 289
- In-depth stream, 144
- Individual, 163
- Information flow, 18, 29
- Initialization, 164

- JIT line, 156
- Job, 151
- Job shop, 157

- Key figure, 93

- Lagrangian relaxation, 259
- Lead time, 20
- Level demand model, 78
- Like-profile, 75
- Linear programming, 31, 109, 112, 119
- Linear regression
 - multiple, 83
 - seasonal, 86
 - simple, 83
- Loading resource, 262
- Loading/unloading times, 262
- Location, 45, 278
 - product, 124
- Lost sales, 127
- Lot, 19
 - size, 31, 146
 - sizing, 133, 143
- Lot-for-lot, 135
- LP, *see* Linear programming
- Make and pack, 19
- Management by exception, 92
- Mass production, 19
- Master
 - Data, *see* Data
 - data element, 140
 - forecast profile, 91
 - planning, 30, 32, 110, 111
 - production schedule, 30, 110
- Material requirements planning, 14
- Matrix
 - supply chain planning, 28
- Mean
 - absolute deviation, 69
 - absolute percentage error, 70
 - error, 69
 - squared error, 70
- Means of transport, 277
- Medium-term, 109
 - planning, 111
- Meta-heuristic, 158, 259
- Middle-out, 72
- MIP, *see* Mixed integer programming
- Mixed integer programming, 110, 119
- Mixed resources, 142
- Model, 44
 - dimensions, 115
- MPS, *see* Master production schedule
- Multilevel planning, 72
- Multiple routes, 272
- Mutation, 165
- One of a kind production, 157
- One point crossover, 164
- Operational planning, 29
- Optimization profile, 280
- Optimizer profile, 141
- Order crossover, 164
- Overtime, 133
 - costs, 11
- Paced assembly line, 157
- Pair selection, 164
- Pallet loading, 252
- Pegging, 52, 143
- Penalty costs, 263
- Perishability, 224
- Pick-the-best, 76
- Planned orders, 135
- Planning
 - book, 53, 141
 - interval, 29
 - production, 38
 - run, 186
 - system, 14, 18
 - transportation, 38
 - unit, 22
- Population, 163
- PPM, *see* Production process model
- Procurement, 20
- Product, 47
 - availability check, 204
 - changeover, 19
 - decomposition, 121
 - group, 23
 - storage definition, 49
- Production
 - coefficient, 15
 - data structure, 48
 - order, 14
 - process, 11, 19

- process model, 48, 124, 139, 176
- segment, 151
- Production process model, 176, 177
- Quality of forecasts, 69
- Quota arrangement, 48
- Random noise, 78
- Rank based strategy, 164
- Ratio-to-moving averages decomposition, 85
- Raw material, 19, 20
- Real-time order promising, 199
- Recombination, 164
- Remaining slack time, 160, 161
- Replenishment lead time, 19
- Resource, 46
 - decomposition, 121
 - network, 46
- Returns planning, 273
- Robust planning, 28
- Rolling schedule, 26, 111
- Root mean squared error, 70
- Roulette wheel strategy, 164
- RTMAD, *see* Ratio-to-moving averages decomposition
- Rules-based ATP, 205
- Runout time, 25
- Safety stock, 31, 32, 111
- SAP[®] business suite, 35
- Scenario, 123
- Schedule, 50
- Scheduling
 - detailed, 38
 - vehicle, 38
- Seasonal
 - coefficient, 79
 - cycle, 79
 - demand, 12
 - demand model, 79
- Setup
 - costs, 11, 19
 - group, 50
 - matrix, 50
 - time, 11, 16, 19, 25
 - sequence-dependent, 17
- Shelf life, 12, 127
- Shortage modeling, 229
- Shortest processing time, 160
- Single (or simple) exponential smoothing, 82
- SKU, *see* Stock keeping unit
- Smoothing parameter, 82
- SNP, *see* Supply network planning
- SNP optimizer, 141
- Soft constraint, 267
- Sourcing flexibility, 226
- Standard product, 18
- Statistical, 72
- Steering costs, 124
- Stock keeping unit, 19
- Stopping rule, 165
- Strategic planning, 29
- Substitutability, 225
- Supply chain
 - engineer, 140
 - planning matrix, 28, 110
- Supply network planning, 37, 110
- Tactical decision, 29
- Time
 - bucket, 30
 - decomposition, 121
 - series, 68
 - series analysis, 68, 72
- Time-based disaggregation types, 93
- Top-down, 72
- Transactional data, 172
- Transportation
 - cost, 267
 - lane, 47, 279
 - load building, 250, 261
 - planning, 209
 - service provider, 39
 - zone, 46, 282
- Traveling salesman problem, 253
- Trend, 79
- TSP, *see* Traveling salesman problem
- Uncertainty, 67, 228
- Univariate methods, 92

- Unweighted moving average, 81
- Utilization, 16
 - rate, 124
- Variable types, 119
- Vehicle resources, 278
- Vehicle routing problem, 252
 - capacitated, 257
 - distance-constrained, 257
 - time dependent, 258
 - with backhauls, 257
 - with pickup and delivery, 258
 - with time windows, 257
- Vendor managed inventory, 231
- Version, 44
- VMI, *see* Vendor managed inventory
- VRP, *see* Vehicle routing problem
- Warehouse replenishment, 31
- Weighted moving averages, 81
- Winters' method, 85
- Working time, 16

About Contributors

The authors are experts in supply chain management and advanced planning with a long list of corresponding papers in international scientific journals. Four are university professors and one is an expert consultant at SAP, Germany.

Bernhard Fleischmann is an emeritus at the University of Augsburg where he held a chair for Production and Logistics until summer 2010. 1978-1991 he was a professor of Operations Research at the University of Hamburg. 1971-1978 he worked in the Operations Planning department of Unilever Germany. His research interests include the development and application of systems for production planning, transportation and distribution planning and inventory management. He can be contacted at bernhard.fleischmann@wiwi.uni-augsburg.de.

Martin Grunow is professor in Production and Supply Chain Management at Technische Universität München (since 2010). From 2006 to 2010, he was professor and head of the Operations Management group and co-leader at the FoodDTU Center at the Technical University of Denmark. Earlier, he worked at the Technical University Berlin and at Degussa's R&D department (a large multinational in the special chemicals sector). His research interests are in production and logistics management with a focus on supply chain management in the process industries. Martin Grunow is a member of a programme committee of The Danish Council for Strategic Research and of The International Academy of Production Engineering. He can be contacted at martin.grunow@tum.de.

Herbert Meyr worked at the department of Production and Logistics at the University of Augsburg from 1994-2003, at the Vienna University of Economics and Business Administration from 2003-2006, and at the Technical University of Darmstadt from 2007-2010. He currently holds a chair for Supply Chain Management at Universität Hohenheim (Stuttgart) and acts as head of the working group "Supply Chain Management" of the German operations research society (Gesellschaft für Operations Research e.V., GOR).

His research interests are model building and integration aspects of supply chain planning. He can be contacted at h.meyr@uni-hohenheim.de.

Hartmut Stadtler is professor of Business Administration at the University of Hamburg. Formerly, he held a chair at Darmstadt University of Technology (1990-2004). From 1987-1990 he was employed as a consultant in the field of production management. He has published numerous articles about operations and materials management in international journals like International Journal of Production Research, Management Science, Operations Research, OR Spectrum and Production and Operations Management. He acts as department editor for the Zeitschrift für Betriebswirtschaft and is in the editorial board of the International Journal of Production Research. He can be contacted at hartmut.stadtler@uni-hamburg.de.

Dr. Christopher Sürrie is currently working as expert consultant in the area of supply chain optimization at SAP Deutschland AG & Co. KG, Walldorf, Germany. In this position he has been involved in numerous projects implementing the SAP SCM software focusing on its optimization engines in production and transportation. Prior to this he worked at the department of Production and Supply Chain Management at Darmstadt University of Technology, Germany and completed his PhD thesis in the area of production planning in process industries which was awarded several prizes. He can be contacted at c.suerie@sap.com.

Research Assistants

Christopher Haub is working as a research assistant at the Institute of Logistics and Transportation at the University of Hamburg. He is currently preparing his PhD thesis aiming at the coordination of production and sales plans.

Sebastian Geier is working as a research assistant at the Chair for Production & Supply Chain Management at the University of Augsburg. From 2008 until summer 2010 he worked at the chair for Production and Logistics at the University of Augsburg. He is currently preparing his PhD thesis on demand fulfillment in an assemble-to-order production.

Bryndís Stefánsdóttir is a research assistant at the Chair of Production and Supply Chain Management, Technische Universität München, Germany. She is currently doing her PhD on Advanced Planning Systems in the Food Industry.

Poorya Farahani is a PostDoc researcher at the Chair of Production and Supply Chain Management, Technische Universität München, Germany. He did his PhD from 2007 to 2011 at the Technical University of Denmark focusing on operations management in food supply chains.

Other Contributors

Below is a list of students who have prepared their Diploma or Master thesis on either testing some modules of an APS or by creating a Frutado learning unit and the corresponding PhD students. Many thanks to all!

Students

Julia Beckschäfer (global ATP)
Steffen Christ (SNP, PP/DS)
Bastian Dittmer (SNP)
Dirk Gerhardt (Deployment)
Baptiste Lebreton (inventor)
Shan Lin (Deployment, TP/VS)
Jarno Lüth (e-learning concept, SNP)
Steffen Schorpp (TP/VS)
Jan-Yves Weseloh (PP/DS)
Jens Philipp Weber (concept)
Sebastian Zier (DP)

PhD Students

Martin Albrecht
Magnus Fröhling
Saskia Köstl (born Kauder)
Florian Kröger
Mario Lueb
Carolin Püttmann
Florian Seeanner
Michael Wagner
Volker Windeck
Julian Wulf