1 Introduction

The aggregate production plan refers to a plan of future aggregate resource levels that makes sure that supply is in balance with demand and costs are minimized. It is a plan that covers the intermediate term (as opposed to short term or long term) and is intended to minimize the effects of frequent changes to levels of resources such as materials and workers due to shortsighted planning. As we shall see in this chapter, the aggregate production planning problem has been, and still is, a practical problem that generates intense academic interest. We begin by emphasizing its importance for operations and then review its academic development, from its genesis in the 1950s until today. However, if aggregate plans are to have practical usefulness, they must produce feasible decisions for the key resources covered by the plan that can be used by the more detailed operational plans that follow. Consequently, we will review the “disaggregation” problem and follow that with a discussion of aggregate production planning in practice. We close the chapter with some thoughts on the future research directions of this important topic and its implications for practitioners.

1.1 The Importance of Aggregate Production Planning

In the term aggregate production plan, “production” refers to the production of goods or services. Consequently, the concept of an aggregate production plan applies equally to manufacturing as well as service firms. Its importance derives from the fact that it must reflect the operations strategy and the business plan, which is a projected statement of income, costs, and profits. The business plan embodies plans for market penetration, new product introduction, and capital investment. In this section we will describe what is aggregated, the information inputs to the plan, and the key decision variables or supply options that comprise the aggregate production plan.

1.2 Dimensions of Aggregation

The aggregate production plan shows how operations will support the operations strategy and the business plan over an extended time frame (usually twelve months) without getting
bogged down in details. To that end, three planning dimensions are aggregated: products or services, workforce, and time.

1.2.1 Products

To avoid too much detail at this level of planning, products are aggregated into product families, which are groups of products that have similar demand, process, workforce, and material requirements. For example, an automobile manufacturer may produce trucks, SUVs, sedans, and electric cars. Each of these product types come in a variety of options and colors, creating thousands of potentially different products. Nonetheless, if we look at trucks, SUVs, sedans, and electric cars as separate product families, the products within each product family would have similar process, workforce, and material requirements. Planning for four product families is much easier than planning for thousands of different products.

1.2.2 Workforce

Employees can be grouped into categories and planned for accordingly. For example, at an auto manufacturer there are assembly workers and machinists, each group with its own set of skills, usage constraints, and wages. Planning for the amount of labor hours required is much simpler for two aggregate groups as opposed to thousands of individual employees. Similarly, for a service firm such as a cable communications company, employees may be categorized into service technicians, utility construction employees, and customer service representatives. Here, the aggregate production plan is called the aggregate employment plan because tangible products and inventories are not a part of the plan.

1.2.3 Time

The planning horizon of the aggregate production plan typically covers twelve months. Additionally, the plan is reviewed on a periodic basis. However, frequent changes to aggregate workforce levels and production rates could be disruptive. Consequently, time is aggregated into months or quarters so that decisions on production and labor are made monthly or quarterly, as opposed to daily or weekly. This is in order for the decisions to be compatible with the budgetary process and other financial plans and to avoid adding unnecessary volatility to the operating environment.

1.3 Information Inputs

The aggregate production plan has implications for an entire business and as such requires informational inputs from many functional areas. It is a dynamic plan that requires constant revision as supply, demand, product mix, and new product platforms manifest themselves over time. Figure 4.1 shows six of those areas, linked together through Information Technology. Operations must supply current machine capacities and plans for future increases or decreases, workforce capability and productivities, and the current staffing and inventory levels. Accounting and Finance provide cost data and the general financial condition of the firm, which may impinge on the strategy to be employed in the aggregate production plan. The Materials and Suppliers function has the data on supplier capabilities, limits on storage capacities, and raw material availability. The labor market conditions, which affect the ease of acquiring new employees, the training capacity,
which determines the time it takes for new employees to attain full productivity, and any ethical constraints on the use of overtime or excessive layoffs are inputs provided by Human Resources. Distribution and Marketing provide information on minimal inventories needed to maintain adequate safety stocks, demand forecasts, and the trends in the market for the firm’s products or services. Finally, Engineering and R&D supply information on the introduction of new products, product design changes, and limits on machine capacities. It is clear that the aggregate production plan is complex and requires continual inputs from a variety of functional areas to be effective.

1.4 Decision Variables and Supply Options

For each period in the planning horizon, the aggregate production plan must specify the production quantity of each product or service family, the regular time workforce for each employee category, the amount of overtime, the amount of undertime (a situation that occurs when employees do not have enough work for the regular-time workweek but are retained nonetheless), the number of new hires, layoffs or part-time workers, the amount of subcontracted or outsourced work, vacation schedules, the inventory levels of each product family, and any planned backorders or stockouts. Given sales forecasts and goals, management manipulates these variables in order to meet the anticipated demand. The problem becomes more challenging if the firm faces a seasonal demand pattern because there are many options available to supply products or services, depending on the costs involved. In a manufacturing firm, seasonal inventory can be accumulated in low-demand periods for use in high-demand periods. Other options are available for both manufacturing and service firms. For example, the workforce capacity can be adjusted by hires and layoffs, or overtime can be used while keeping the regular-time workforce capacity unchanged. Undertime could be used to retain skilled employees during low demand periods. Part-time employees could be used in lieu of overtime. Another option is to use subcontractors to overcome short-term shortages in capacity, or to employ a strategy of outsourcing in long-term
situations. Vacation schedules or planned shutdowns of plants can be scheduled so that they coincide with low-demand periods. It is no wonder that such a complex problem has been the topic of considerable academic research.

2 Historical Perspective of Aggregate Production Planning Research

In this section, we will provide an historical perspective to the research literature addressing the aggregate production planning problem. It is largely quantitative in nature, reporting the development and analysis of mathematical models utilizing a wide range of quantitative techniques. It will be impossible to cover all of the research devoted to the aggregate production planning problem here. Extensive literature reviews of the aggregate production planning problem can be found in the literature (e.g., Nam and Logendran 1992). Instead, we cover the genesis of this topic in the academic literature and discuss examples of the various methodological tendrils and formulations that followed.

2.1 Linear Decision Rules

The early 1950s provided the genesis for the aggregate production planning problem in the academic research arena. Often referred to as “production smoothing,” the problem provided the attributes of substance and practicality that underlie sound research. One of the earliest and most comprehensive studies of the aggregate production planning problem was performed by Holt, Modigliani, and Simon (1955). The problem they consider derives a set of decision rules for production output and workforce levels such that expected costs are minimized over the planning horizon. Their study focused on a paint factory and used extensive data from 1949 to 1953 in order to develop the rules. This single work, and the book that followed with J. F. Muth as the third author (1960), became known as the HMMS model (the acronym for the work of Holt, Modigliani, Muth, and Simon) and set the stage for research for many years to come. Indeed, the principle result of this research, the Linear Decision Rules (LDR), reveals a concept of coordinated optimization of aggregate production, workforce, and inventory decisions that is natural and elegant, and has been used as a standard of comparison for many theoretical studies for decades afterwards (Schwarz and Johnson 1978; Sprague et al. 1990).

The problem is to minimize a total cost function which is the sum of a) regular time wages, b) overtime wages, c) hiring and firing costs, and d) inventory and back order costs. We will use the following notation:

\[ W_t = \text{number of workers required for period } t \]
\[ P_t = \text{number of units to be produced in period } t \]
\[ I_t = \text{inventory minus backlogs at the end of period } t \]
\[ R_t = \text{demand requirements for period } t \]
\[ c_k = \text{cost coefficients to be determined for a given plant, } k = 1, \ldots, 10. \]

2.1.1 Regular Time Wages

The regular time wages are assumed to be a linear function of the workforce each period,\[ c_1 W_t \] where \( c_1 \) is the regular time wage rate per period.
2.1.2 Overtime Wages

The overtime wages only occur when production exceeds the number of units the workforce can produce on regular time, \( c_3 W_t \), and thereby is only piece-wise linear with respect to the production rate in a given period. These costs can be approximated by a quadratic cost function:

\[
c_2(P_t - c_3 W_t)^2 + c_4 P_t - c_5 W_t
\]  

2.1.3 Hiring and Firing Costs

The hiring and firing costs depend on changes in the workforce from period to period and can be approximated by a quadratic function:

\[
c_6(W_t - W_{t-1})^2
\]  

2.1.4 Inventory and Backorder Costs

As inventory rises, holding costs increase. As inventory decreases, stockout and backorder costs increase. Once the “optimal” inventory level for a period has been determined, the costs can be approximated by the following quadratic function of the difference between the actual inventory and the optimal inventory in period \( t \),

\[
c_7(I_t - (c_8 + c_9 R_t))^2
\]  

The optimal inventory, which is determined by a lot size formula, is a function of the demand in period \( t \) and can be approximated by \( c_8 + c_9 R_t \).

2.1.5 Objective Function and the Rules

The objective is to minimize

\[
C_\tau = \sum_{t=1}^{\tau} c_1 W_t + c_2(P_t - c_3 W_t)^2 + c_4 P_t - c_5 W_t + c_6(W_t - W_{t-1})^2 + c_7(I_t - (c_8 + c_9 R_t))^2 + c_{10}
\]  

subject to

\[
I_t = I_{t-1} + P_t - R_t \quad t = 1, \ldots, \tau
\]

where \( c_{10} \) is the fixed cost and is the planning horizon. The solution procedure entails taking the partial derivatives of \( C_\tau \) with respect to the decision variables for each period, \( P_t \) and \( W_t \), and equating them to zero, producing 2 \( \tau \) equations in 2 \( \tau \) variables. Because the individual components of \( C_\tau \) are either linear or quadratic, the resulting form of the solution will be a set of linear equations, called the Linear Decision Rules (LDR), dependent only on the initial conditions, the cost coefficients, and the periodic demands:

\[
P_t = k_{n,t} W_t + k_{n,1} I_t + \sum_{j=1}^{\tau} \alpha_{p,j} R_j + k_{n,t}
\]  

\[
W_t = k_{w,t} W_t + k_{w,1} I_t + \sum_{j=1}^{\tau} \alpha_{w,j} R_j + k_{w,t}
\]

where the constants \( k_{n,1}, \alpha_{p,j} \) and \( \alpha_{w,j} \) depend only on the original cost coefficients.
Lee Krajewski

The HMMS model has been one of the most scrutinized and emulated models in the operations management literature. Nonetheless, the derivation of the rules is difficult and force-fitting quadratic equations to piece-wise linear costs often results in decisions that are only marginally better than informed managerial decisions (Schwarz and Johnson 1978; Lee and Khumawala 1974). In the remainder of this section, we will follow the methodological evolution of the aggregate planning problem for both manufacturing and service firms.

2.2 Linear Programming

The introduction of linear programming as a methodology for the aggregate production planning problem opened the floodgates for many applications. In one early application, Bowman (1956) noted that the aggregate production planning problem could be put into the framework of a transportation problem. The formulation involves keeping track of when an individual unit produced will satisfy a known demand. While the transportation method provides an integer solution, a general linear programming formulation would allow for a host of constraints on the decision variables that Bowman’s model does not. Manne (1957) simplified the quadratic costs in the HMMS model by assuming that marginal production costs are constant and the workforce level is pre-determined for the duration of the planning horizon. He then approximated the total cost function with a piece-wise linear function, which then allowed for consideration of different work shifts, such as regular time and overtime. The decision variables are the production quantities on regular time and overtime and the inventory level for each period of the planning horizon.

Motivated by introducing the workforce as a decision variable and reducing the complexity of the HMMS model, Hanssmann and Hess (1960) developed a model that incorporated linear cost functions. The advantage to this model was that it allowed them to independently choose the unit costs of hiring and firing or inventory and backorders rather than trying to empirically fit a quadratic cost function to the change in workforce levels or inventory levels. Costs were modelled as step functions, relegating the objective to a piece-wise linear function that, after the definition of a new set of variables to deal with the step functions and the addition of three technical constraints for each period, could be minimized using linear programming. The model’s major focus was on showing how to transform the HMMS model to a linear program that was much easier to solve using the computer technology of the time. As time passed and computer technology advanced, aggregate production planning models using linear programming, capable of incorporating many variables each period such as subcontracting, overtime, hires, fires, inventory, production, and workforce as well as a host of constraints and conditions on those variables, including integer variables, proliferated the literature (e.g., Chung and Krajewski 1987; Chopra and Meindl 2004; Kanyalkar and Adil 2007; 2010; Leung and Chan 2009).

Linear programming can also be used for the aggregate production planning problem in service organizations. It is referred to as the aggregate employment planning problem because those organizations are focused on workforce decisions and do not have inventories of produced goods. Nonetheless, the concept is the same. For example, Krajewski and Thompson (1975) developed an aggregate employment planning model for a telephone company that sought to minimize the costs of regular time wages, overtime wages, hiring and layoff costs, and subcontracting costs in light of a seasonal demand for its services. There were four employee groups, representing different skill sets, each with full and part time employees. Newly-hired employees first became part-time employees who progressed over time to full-time status. The four sets of constraints for each period consisted of (1) definitions of the number of employees in each group after hires, layoffs, attrition rates, and the rate of progression to full-time status was accounted for, (2) relationships ensuring that enough employee capacity, including overtime and planned
delays, was available to meet demand for twenty-five different services each period, (3) limits on the amount of overtime that could be used each period, and (4) limits on the amount of delay in each service that would be tolerated each period. For a planning horizon of 12 months, there were more than 6,000 variables and 1,400 constraints, which would pose no particular problem with today’s computing power. However, such a model would have posed a significant practical problem back in the day of Hanssmann and Hess. Present day computers eliminate this problem and open the vista for continued research into new methodologies and more elaborate models.

2.3 Heuristics

While the HMMS model and the resulting LDR provide an optimal solution to the aggregate planning problem, they do so by making stringent assumptions about the cost structure and the scope of the problem. Costs are assumed to be quadratic and the environment is assumed to have only a few constraints. Alternatively, linear programming models can recognize many complex constraints but must assume linear costs and relationships between the variables. While the attraction to the elegance and simplicity of the LDR was a driving force in the literature, the limitations of the LDR regarding the quadratic costs and the difficulty in deriving the rules set the stage for the development of heuristic approaches to the aggregate production planning problem. Bowman (1963) developed equivalent versions of the LDR in equations [7] and [8] that were deemed suitable for regression and used management’s past decisions as data to estimate the coefficients for the rules. This approach avoids making assumptions about the form of the cost functions and takes advantage of the knowledge of experienced managers who are aware of the important performance criteria of a system. The rules, one for production and the other for the workforce, were dubbed the Management Coefficients Model (MCM). When compared to actual management performance in four companies, MCM was better in three of the cases, including the original paint company. MCM even beat the LDR in two of the cases, excluding the paint company. Bowman theorized that it is the variance in decision making, not the bias, which hurts because of the dish-shaped criteria surfaces. Therefore, decision rules with mean coefficients estimated from management’s past behavior should be better than actual performance.

Jones (1967) developed an approach called Parametric Production Planning (PPP) which attempts to free the manager from classifying the cost function as either quadratic or linear. Following the lead of the LDR, Jones proposed two linear feedback decision rules: one for the workforce and one for the production each period. The workforce rule for the first period includes consideration of the current workforce level, the desired workforce to meet future demands, and the difference between the desired inventory level and the current inventory level. Given the workforce decision for the first period, the production rule considers the output that the workforce can produce, the desired production to cover future demands, and the difference between the desired inventory level and the current inventory.

Each rule contains two smoothing parameters, between zero and one, that are used to (1) average the current workforce or production capacity with the desired increase or decrease, and (2) place a weight on each demand forecast over the planning horizon for determining the average desired workforce or production levels. As the rules are applied, changes in the workforce level from period to period reflect hires or layoffs, while production in excess of the workforce capacity reflects overtime.

Jones used a grid search to find the best values of the parameters. For each trial, the workforce and production decisions were evaluated by the actual cost structure of the firm, something which was not limited to linear or quadratic functions. Compared to the LDR developed for the paint company, which is optimal for the assumed cost functions, PPP resulted in savings that were
only 0.09% less than the same quadratic cost function. When compared to a linear programming solution of a problem with a linear cost function and linear constraints, the solution using PPP was 8% higher.

Smitten by the LDR for similar reasons, Taubert (1968) proposed the Search Decision Rule (SDR) to provide an option to determine production and workforce decisions in complex settings that does not rely on linear or quadratic cost assumptions. What the SDR loses in terms of optimality, it gains in terms of realism and simplicity. As with the LDR, a workforce and production decision must be made each period, from which the costs of wages, overtime, hires, layoffs, inventory, and the like can be determined. Over a planning horizon of $\tau$ periods there are $2\tau$ variables. Taubert proposed that a computer search routine be used to explore the $2\tau$ dimensional response surface of the model’s cost function. He tested the approach with an objective function of 20 variables over a 10-period planning horizon and compared the results to the optimal LDR for the paint factory. Twenty periods were simulated and the total costs were only 0.1% more than the LDR even though the LDR had a 12-month planning horizon.

### 2.4 Evaluation of Early Aggregate Planning Models

Before we continue with the methodological evolution of the aggregate planning problem in the literature, it is important to note that many extensions to the LDR were reported in the literature. For example, multiple products (Bergstrom and Smith 1970; Chang and Jones 1970; Welam 1975), marketing variables (Damon and Schramm 1972; Leitch 1974; Tuite 1968), financial variables (Damon and Schramm 1972), and learning (Ebert 1976) were added to the LDR framework. The LDR was held up to be the standard bearer, the one used as a benchmark for performance. However, LDR had its limitations, and alternatives, such a MCM, PPP, and SDR that were not limited to quadratic cost assumptions, might actually be better in some scenarios. Lee and Khumawala (1974) found that to be the case. Using a firm in the capital goods industry, they simulated the application of these four approaches and concluded that the SDR outperformed the other models. However, the LDR and MCM approaches can perform well in certain circumstances. The key is to provide a realistic test situation within which the models can be evaluated for any given firm.

The elegance of the LDR aside, and the numerous extensions and articles written about it in the twenty years after its development, Schwartz and Johnson (1978) asked the profound question as to why there is no record of it ever being implemented in practice. They hypothesized that the LDR has not been implemented because most of the cost savings of the LDR can be achieved by improved inventory management alone. Using the originally reported cost comparisons of the LDR versus the paint company management’s decisions, they show that virtually all of the savings of the LDR could have been achieved by a single one-time adjustment to the aggregate inventory buffer inventory. None of this detracts from the massive impact the LDR had on the development of the aggregate production planning literature. It does point out, however, that in addition to the potential superiority of improved inventory management, the LDR had other shortcomings in the implementation arena, including the inability to handle integer variables (Taubert 1968), the difficulty of constructing adequate aggregate cost functions that represent the firm’s actual individual product and worker costs (Krajewski et al. 1973), and the problem of disaggregation (Zoller 1971), which will be explained in Section 3.

### 2.5 Goal Programming and Other Methodological Thrusts

Given the practical shortcomings related to the early efforts at modeling the aggregate production planning problem, academic researchers focused on more realistic formulations and
more effective solution methodologies. One of their major criticisms was the singular focus on cost minimization in light of situations where management had other goals to consider, especially when those goals could not be reflected in dollars and cents. Goal programming (GP) was seen as a way to incorporate multiple goals in a linear programming format (e.g., Charnes and Cooper 1961; Rifai 1994; Romero 1991; 2004; Tamiz et al. 1998). The GP as a concept serves to introduce extra auxiliary variables called deviations that represent the distance between aspiration levels of goals, or targets, and the realized results. A GP model includes a set of technical constraints, as in a typical linear program, and a set of goal constraints with the auxiliary variables, which determine the best possible solution with respect to the goals. Each auxiliary variable can have its own coefficient in the objective function that signifies its priority as a goal. An early application of GP for the aggregate production planning problem focused on the LDR formulation, using the GP concept to approximate the quadratic cost curves (Goodman 1974). However, it was quickly realized that GP had the potential to add much more realism to the aggregate planning problem by incorporating conflicting goals in addition to minimizing cost or maximizing profit. For example, goals such as achieving production, employment, and inventory targets (Jaaskelainen 1969), maximizing customer service and plant/machine utilization (Krajewski 1984; Leung and Chan 2009), and quality and on-time deliveries (Jamalnia and Soukhakian 2009) are just a sample of the multiple goals reported in GP models for the aggregate production planning problem.

The aggregate production planning problem has been fertile ground for the application of diverse quantitative methodologies. For example, Jamalnia and Soukhakian (2009) modelled the objective “Total customer satisfaction should be relatively high” as a fuzzy goal in a hybrid GP approach. (See Narasimhan 1980, for details on the use of GP in a fuzzy environment.) Li et al. (2013) developed a belief-rule-based inference method for the aggregate production planning problem under uncertainty. Krajewski (1984) developed a heuristic procedure to solve a mixed-integer GP of an aggregate planning problem using data from an industrial-goods manufacturer. Finally, just when you think everything has been tried, Kumar and Haq (2005) combined a genetic global search algorithm with an ant colony algorithm, which is a heuristic optimization method typically used for job shop scheduling and shortest path problems using ideas borrowed from biological ants searching for food. Randomly wandering ants find a food source (solution to a problem) and follow a trail back to the nest, leaving markers for other ants to follow. The markers on the trail begin to evaporate over time, the longer the trail the more evaporation and the less likely other ants will follow. The shorter trails (least costly solutions) tend to be followed by more ants and the marker density becomes greater. Eventually all the ants follow a single path, which then becomes the solution to the problem. A hybrid model, consisting of a genetic algorithm combined with an ant colony algorithm for the aggregate planning problem, was better than either the genetic algorithm or the ant colony algorithm used separately.

3 Disaggregation of Aggregate Production Plans

The aggregate production planning problem, by virtue of the aggregation of many products and employee types into a smaller, more manageable number of decision variables, allows for the planning of resources over an intermediate time horizon. We have seen the intensity of the effort at modeling that problem with more realistic formulations and more powerful solution methodologies since the early days of the LDR. However, none of that will be useful in practice if the aggregate quantities cannot be translated into the plans and schedules for the individual products and employees represented in the plan. Disaggregation is the process of breaking apart aggregate decisions for production quantities and workforce sizes into feasible lower-level
schedules. Marty Starr, while making the plenary address to the delegates at the Conference on Disaggregation at The Ohio State University, perhaps put it all into perspective in the following quote (see Ritzman et al. 1979):

While doing research for this paper, I used the standard literature search which consists of looking up ‘Aggregate’ in the dictionary. My eye wandered to the entry above and I read ‘to make worse, to exasperate; annoy.’ And thus was born the slogan ‘Aggregation is preceded by Aggravation.’ I thought, if that is so, then ‘Disaggregation’ should be preceded by pleasure. When I remarked on this twist of logic to a learned colleague, I was told that Anna Freud had observed that children take pleasure in taking things (e.g. butterflies) apart. She went on to point out that children are criticized for this taking apart and consequently grow up resisting the powerful learning experience of thoughtful disassembly.

We will now explore the thoughtful disassembly of aggregate production plans.

### 3.1 Levels in Operations Planning and Scheduling

Managers develop plans for their operations covering varying time spans, from the long term to the short term. These plans form a hierarchy: the long-term plans form an umbrella under which short-term plans exist (Krajewski et al. 2016). Figure 4.2 shows the levels of disaggregation and the plans that should be consistent relative to each other. As we will see in Sections 3.2 and 3.3, valid aggregate plans should incorporate data and constraints reflecting these lower-level plans to ensure the feasibility of the production quantities and the workforce sizes.
3.1.1 Level 1

The aggregate production plan, often referred to as the sales and operations plan in practice (Wallace and Stahl 2008) is at level 1 of the hierarchy and requires the inputs from a number of functional areas. First, there is the business plan, which is a projected statement of income, costs, and profits for the next one or two years. It is typically accompanied by budgets and cash flow projections that reflect the unified plans and expectations of a firm’s operations, finance, sales, and marketing managers. These expectations include plans for market penetration, new product introduction, and capital investment. The aggregate plan must respond to these plans and expectations.

Second, the operations strategy is a statement of the way operations will implement the corporate strategy regarding the production of goods or services relative to the competitive dimensions the firm wants to emphasize. These competitive dimensions are reflected in priorities regarding cost, quality, time, and process flexibility, which affect the aggregate production plan regarding the use of seasonal inventory, backorders, overtime, hires and layoffs, vacation time, and subcontracting, for example.

Third, forecasting provides the demand projections that the aggregate production plan must address (see Chapter 3, Forecasting: State-of-the-Art in Research and Practice).

Finally, the aggregate production plan must adhere to various constraints on the resources at its disposal. For example, there may be restrictions on the use of overtime, part-time employees, layoffs, and vacations, and other human resource considerations. Further, there may be limitations on the production facilities, such as bottlenecks, plant capacity, and the like.

3.1.2 Level 2

At the next level, resource planning disaggregates the aggregate quantities of product or service families, workforce, and time to arrive at material and resource requirements over a time horizon of less than a year, broken down into time periods of days or weeks. For a manufacturing firm, the resource plan gets specific as to individual products, purchased materials, and resources. A major input is the master production schedule (MPS), which specifies the timing and size of production quantities for each product included in each product family. The material requirements plan (MRP) then determines the timing and size of orders for the components and purchased materials and workstation labor hours. Similarly, the resource plan for a service firm disaggregates projected service demands into daily or weekly requirements for the employees and service facilities using a bill of resources (Roth and Van Dierdonck 1995). Whether it is manufacturing or service, the aggregate production plan plays a direct role in the detailed planning that follows.

3.1.3 Level 3

Finally, scheduling translates the resource plan into specific operational tasks on a detailed basis (see Chapter 5 “Scheduling in Manufacturing and Services”). Given the resources provided by the aggregate plan, and the detailed requirements provided by the resource plan, facility, workforce, and job schedules can be developed. For example, an aggregate workforce plan may allocate fifteen police officers for the night shift in a particular district for a given month. The resource plan may determine the daily police protection requirements for a typical week, and the workforce schedule might assign eight of them to work Monday through Friday and the other seven to work Wednesday through Sunday to meet the varying needs for police protection in that district. Similarly, given the material requirements plan for a group of jobs in a manufacturing plant, the specific sequence of the jobs on a bottleneck machine can be scheduled.
Figure 4.2 implies that information flows in two directions: from the top down and the bottom up. If an aggregate plan cannot be developed to satisfy the objectives of the business plan with the existing resources, the business plan may have to be revised. Likewise, if a feasible material requirements plan or workforce schedule cannot be developed, the aggregate production plan may have to be revised. In the next two sections we will explore some of the research addressing the disaggregation of aggregate plans, which will demonstrate the importance of recognizing lower-level decisions in preparing a feasible aggregate production plan.

3.2 Manufacturing

This section first describes the initial attempts at developing aggregate production plans that recognize lower-level decisions and then shows some of the research that is based upon that earlier work and the direction it is taking.

3.2.1 Hierarchical Production Planning (HPP)

Much in the way the LDR was a catalyst for the early work in aggregate production planning, so was hierarchical production planning (HPP) a catalyst for the work on disaggregation (Hax and Meal 1975; Bitran and Hax 1977; 1981; Bitran et al. 1981; 1982). HPP follows the way in which decisions are made within the organizational hierarchy. Aggregate (strategic and tactical) decisions are made first and impose constraints within which more detailed (operational) decisions are made. In turn, the detailed decisions provide feedback to evaluate the quality of the aggregate decisions. HPP is as much a philosophy for solving the disaggregation problem as it is an actual technique. That is, the approach to solving the problem should be from the top down; the techniques used for solving sub problems could vary. There are three levels of aggregation:

- **Product types**: groups of products that have similar unit costs, direct costs, holding cost, productivities, and seasonality.
- **Families**: groups of products that belong to the same product type and share similar setups.
- **Items**: the final products delivered to the customer and the highest level of specificity; they differ in characteristics such as color, packaging, and size.

HPP develops an aggregate plan for product types, disaggregates it for product family production lots, and then disaggregates the family production to item quantities. There is management interaction at each stage of the process. Hax and Meal (1975) proposed a heuristic to perform the three levels of computations. Later, Bitran and Hax (1977) suggested the use of convex knapsack problems to disaggregate the product type quantities into family run quantities and then family run quantities into item run quantities.

3.2.2 Setups, Resource Profiles, and Distribution Plans

HPP did not consider setup costs in the aggregate plan. Chung and Krajewski (1987) formulated a mixed-integer programming model that incorporated setup costs at the aggregate planning level that interfaced with a goal programming model for the master production schedule (MPS) via a rolling horizon feedback mechanism. The MPS used resource planning profiles to ensure feasibility regarding work center capacities. Qiu et al. (2001) added setup costs at the aggregate level of the HPP model and incorporated an expert system to generate production plans and schedules for a fiber producer.
Inevitably, the evolution of problem solving in this area focused on expanding the problem scope to take advantage of the increased computing power available. Kanyalkar and Adil (2007; 2010) formulated a monolithic mixed-integer linear goal program (MILGP) that produced a time and capacity aggregated production plan, a detailed procurement plan, and a detailed distribution plan simultaneously to ensure the feasibility of all planning levels. Two heuristics were developed to solve the MILGP. Fahimnia et al. (2012) developed a mixed integer non-linear formulation (MINLF) that integrates the aggregate production plan and detailed distribution plan. They use a genetic algorithm, which is a search heuristic that mimics the process of natural selection, to solve a real-case scenario. This article contains many more references for the production-distribution disaggregation topic. See Mitchell (1996) for a detailed description of the genetic algorithm methodology.

3.3 Services

Aggregate planning for services differs from that of manufacturing because there are no tangible products to store as seasonal inventory. Typically, aggregate employment plans (also referred to as the staff-sizing problem) specify monthly employee levels (regular, part-time, temporary), overtime commitments, subcontracting needs, vacation plans, undertime levels, and hiring and layoff quantities (Ritzman et al. 1976). The aggregate staff sizes must be disaggregated into staff schedules, which specify for each employee a set of assigned work periods over a given time horizon. Other than creating monolithic models of the staff-sizing and staff-scheduling problems and using decomposition to arrive at feasible decisions (e.g., Ruefli 1971; Sweeny et al. 1978), several recursive approaches have been proposed. Top-down recursive approaches were developed by Nolan and Sovereign (1972) and Carlson et al. (1979) in which mathematical programs determine staff sizes and then, those results are tested in simulation models for feasibility regarding the staff schedules. If the staff-sizing plan was found to be infeasible, additional constraints are added to the mathematical program and a new staff plan generated.

A bottom-up approach suggested by Abernathy et al. (1973) uses a Monte Carlo simulation model to determine detailed work center requirements such that service levels will be met at a specified level of risk. These data provide constraints for a workforce allocation model. The resulting staff sizes are checked to determine if macro cost objectives are satisfied and, if not, new requirements are determined and the process repeats. Henderson et al. (1982) propose a recursive bottom-up approach that uses a heuristic staff scheduling model developed by Showalter et al. (1977) to generate a plot of service levels over a range of feasible staff sizes for each period of the planning horizon. From this analysis a target and a minimum staff size reflecting desired and minimal service levels are specified for each period for an aggregate employment goal programming model, which then specifies optimal staff sizes over the time horizon. In addition to a service goal, the model can incorporate goals on cost, mix of part-time and temporary employees, and workforce stability.

4 Aggregate Production Planning in Practice

Aggregate production planning is often referred to as sales and operations planning (S&OP) in practice (Wallace and Stahl 2008). The new title is meant to place emphasis on the fact that the aggregate plan is a top management plan, and not just the responsibility of operations management. S&OP enables top management to proactively manage customer service levels, inventory investment, and lead times for fulfilling customer orders, thereby giving them control over these strategic aspects of operating the business. S&OP is a six-step decision-making process that is
dynamic and continuing, as aspects of the plan are updated periodically when new information becomes available and new opportunities emerge.

4.1 Step 1: Roll the Plan Forward

S&OP involves making monthly or quarterly decisions for a planning horizon that often extends a year or more into the future. Each period the plan is updated. This is often referred to as a “rolling horizon” in the aggregate production planning literature (e.g., Baker 1977; Chung and Krajewski 1987). Actual sales, production, inventory, and costs are recorded for the most recent past period and the database is prepared for the development of a new plan.

4.2 Step 2: Forecast and Demand Planning

Managers from sales and marketing provide forecasts of demand for each product family. The forecasts can be from statistical models of past demand patterns, tempered by the promotions and efforts at demand planning initiated by marketing. For service providers, the demand forecasts could be the result of a bottom up recursive approach (e.g., Abernathy et al. 1973; Henderson et al. 1982).

4.3 Step 3: Update the Sales and Operations Plan

Given the data from steps 1 and 2, the sales and operations plan is updated for each period of the planning horizon, recognizing any new constraints imposed by supplier shortages, plant capacities, storage limitations, and the like. Management may impose limitations on the amount of backorders, the use of subcontractors or outsourcing, and inventories. As there could be many plans that satisfy these constraints, the aggregate planning models discussed earlier can be helpful, especially those that can incorporate multiple goals such as cost, customer service, and workforce stability.

4.4 Step 4: Consensus Meetings

The process of updating the sales and operations plan typically results in several alternatives that differ in their achievement of the various goals. Plan stakeholders should hold a meeting with the goal of achieving consensus on how the firm will balance supply with demand. The plan would be presented at the executive S&OP meeting. The stakeholders include supply chain management, key suppliers and partners, supplier management, plant management, production control management, logistics management, and controller. Lacking consensus, several alternatives should be prepared. Further, an updated financial view of the total business can be prepared by accumulating the plans for each product family, converted to dollars.

4.5 Step 5: Executive S&OP Meeting

The recommendations for each product family are presented to the firm’s president and vice presidents of each functional area. The plans are reviewed relative to the business plan, new product plans, and other corporate plans. The executives may ask for changes. For example, the demands planned for a product family may be too aggressive, or not aggressive enough. Consensus at this level means that everyone marches to the same drummer, whether they are in total agreement or not.
4.6 Step 6: Update and Revise Final Plans

The changes mandated at the executive S&OP meeting are incorporated in the final aggregate plan for each product family. The new plans are communicated to all the stakeholders, where the detailed schedules are prepared. However, the sales and operations plan is a dynamic plan that must be constantly reviewed and revised as necessary. Feedback loops made possible by information technology allow managers to use a rolling planning horizon to make revisions on a periodic basis.

5 Conclusions

Now that we have reviewed the problem of aggregate production planning and discussed its application in practice, we conclude with a discussion of future directions of research in this area and its implications for practitioners.

5.1 Future Research

This chapter has shown that the aggregate production plan is a key to the effective execution of a firm’s operations strategy. While the analysis of the aggregate production plan had its origins in the 1950s, there are still six topics that could benefit from further research.

5.1.1 Employment Planning in Manufacturing

Aggregate plans, whether they are for manufacturing or service firms, include plans for employment levels, hires, layoffs, overtime, undertime, and vacations. With the advent of technology to operating processes, skilled labor is becoming in short supply. For example, the McKinsey Global Institute (McKinsey & Company 2012) projected a potential shortage of 40 million skilled manufacturing workers globally by 2020. Manufacturing firms will compete with service firms for this labor. Aggregate plans must incorporate the potential for hiring shortfalls and the need to entertain other sources of additional resources, such as placing more emphasis on offshore sources. For manufacturing firms, increased logistic lead times from offshoring may require additional inventories and more production during the off season. Further, firms may have to provide more extensive training to their new hires, increasing the lag time until they are fully productive.

5.1.2 Employment Planning in Services

Another employment planning topic worthy of future research is the growing expansion of services in manufacturing firms. Customers are increasingly looking to manufacturers for after-market services. Consequently, aggregate employment plans should recognize the additional resources manufacturing services may require as well as their timing. Finally, aggregate plans should incorporate any human resource policies the firm may have in place, such as limitations on layoffs and union-related limitations on the use of employees.

5.1.3 Aggregation

The consumer demand for products and services is constantly shifting and fragmenting. For example, new consumers in emerging economies often require very different products and services than their counterparts in more developed economies. This forces firms to provide more
variety and stock keeping units (SKUs). The consumers in more established economies constantly demand more variety and faster product life cycles, all of which creates problems for producers who are trying to plan for adequate resources in the next six to eighteen months. One of the advantages of aggregate planning is to reduce the complexity of planning for a variety of products or services. In this emerging environment where fragmentation is necessary for competitiveness, new approaches for aggregating products, services, and employees is needed. This is especially true with the new strategic product platform concepts that are being utilized.

5.1.4 Uncertainty

When planning operations, the one thing for certain is that there will be uncertainty. It could come in the way of poor forecasts or evolving fragmentation of products or services, as mentioned in Section 5.1.3. Alternatively, the firm could be the victim of a massive natural disaster, such as a hurricane, which would affect its own operations or that of its suppliers. Each source of uncertainty has its own effect on the aggregate production plan. Future research should explore the ways to mitigate uncertainty in the aggregate production plan.

5.1.5 Sustainability and Reverse Logistics

Many firms now recognize the need to be responsible stewards of the earth’s resources and have started planning for the entire birth to death cycle of products. A product begins at the new product development process, makes its way to the customer and then enters a reverse logistics process that either prepares the product for direct reuse (e.g., car rental), remanufactures it by tearing it down and rebuilding it with new parts as needed (e.g., diesel engine), or completely disassembles it and the useable parts cleaned, tested, and returned to the manufacturing process (Mollenkopf and Closs 2005; Ferguson 2010). Because of the reverse logistics process, each new product generates resource requirements beyond the point of its initial sale. Sustainability and the reverse logistics process pose interesting research opportunities regarding the aggregate production plan. Recognizing supplier capabilities throughout the birth to death life cycle and the logistical implications of material flows are important. Appropriate time lags and forecasts must be incorporated in the aggregate production plan to adequately plan for resources.

5.1.6 Supply Chain Visibility

Suppose the sales and operations plan for Firm X specifies the aggregate sales and production plan for the next 12 to 18 months. In addition to providing the operating targets for Firm X, it also provides a picture of the requirements for its direct suppliers, often called Tier One suppliers. Each of these suppliers must develop a sales and operations plan that provides requirements for their suppliers (Tier Two), and so on. The sales and operations plan for Firm X has a ripple-down effect on the entire supply chain. Viewing the problem myopically, Firm X could choose to use a chase strategy, seeking to follow the peaks and valleys of its seasonal demand pattern. Or, it could use a level strategy, using seasonal inventories to meet peaks in demands throughout the year. Regardless of the strategy in play, the sales and operations plan for Firm X affects the cost of its supply, especially if it causes suppliers to react in ways that are not the most efficient. If Firm X has the most power in the supply chain, it can dictate its plan and require suppliers to react. If not, Firm X might have to revise its sales and operations plan to conform to what the supply chain can provide. Future research efforts could be directed at the sales and operations plan in a supply chain context in much the same way we did in Section 3.0 for synchronizing within-firm
 Aggregate Production Planning

plans and schedules. That is, the aggregate production plan should recognize the aggregate production plans of key suppliers in the supply chain; each key supplier could be considered a sub problem to Firm X’s aggregate plan. Recognizing multiple supplier tiers would add considerable complexity. Approaches such as bottom–up recursive planning might prove useful. The use of multiple criteria and the recognition of power in this context may also be fruitful. Further, the development of computer-based systems to handle the complexities of supply-chain visibility, along with the associated uncertainties and need to incorporate sustainability efforts, should be a primary research agenda.

5.2 Implications for Practitioners

Aggregate production planning has been, and will continue to be, a key activity in the planning cycles of manufacturing as well as service firms. Over time, the aggregate plan has morphed into the sales and operations plan or the aggregate employment plan depending on the firm. Its boundaries have expanded to include finance, marketing, sales, and human resources. Computing power has increased to the point where the granularity of the aggregate plan has reached lower level plans and schedules. All of this has helped to improve customer service and profitability. It is for these reasons that practitioners should look forward to the results of continued research on this topic.

References and Bibliography


Lee Krajewski


Aggregate Production Planning


