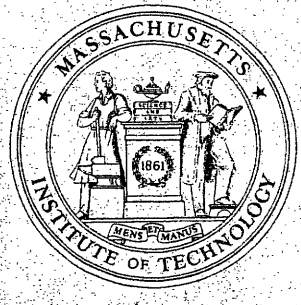
OPERATIONS RESEARCH CENTER

working paper



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Aggregate Capacity Planning
A Survey

by

Arnoldo C. Hax
Sloan School of Management
M.I.T.

OR 027-73 September 1973

ABSTRACT

The purpose of this paper is to provide a survey of the most important uses of operations research techniques in supporting decisions associated with aggregate capacity planning. Various approaches to deal with the aggregate capacity planning are being described, involving the use of mathematical programming techniques, heuristic procedures and search techniques. These approaches have been classified (according to the structure of the objective function they assume) into linear cost models, quadratic cost models, fixed cost models, and general cost models. A description is also given on the hierarchical approach to integrate aggregate planning decisions with detailed scheduling.

Decisions Affecting the Production Management Process

Economists define production as the process by which goods and services are created. In more specific terms, production can be defined as the process of converting raw materials into finished products. Of course, the terms "raw materials" and "finished products" are relative, since what constitutes a finished product for one industry, could be the raw material for another firm. An effective management of the production process should provide the finished products in appropriate quantities, at the desired times, of the required quality, and at reasonable costs.

Prior to analyzing the various models associated with the production process, it might be helpful to categorize these decisions according to the now familiar taxonomy proposed by Anthony [2] regarding strategic planning, tactical planning and operations control.

(1) Strategic Planning: Facilities Design

Strategic policies are mostly concerned with the establishment of managerial policies and with the development of the necessary resources the enterprise needs to satisfy its external requirements in a manner consistent with its specific goals. In the area of production management the most important strategic decisions have to do with the design of the production facilities, involving major capital investments for the development of new capacity and the expansion of existing capacity. These decisions include the determination of location and size for new plants, the acquisition of new equipment, and the design of working centers within each plant. Other decisions, which require strong coordination with marketing, are the selection of new products, and the design of the logistics system (including warehouse location and capacity, transportation means, etc.)

These decisions are extremely important because, to a great extent, they are responsible for maintaining the competitive capabilities of the firm, determining its rate of growth, and, eventually, defining its success or failure. An essential characteristic of these strategic decisions is that they have long lasting effects, thus forcing long planning horizons in their analysis. This in turn, forces the recognition of the impact of

uncertainties and risk attitudes in the decision making process. As we will see, this imposes some problems for the proper use of mathematical programming models which, except for parametric analyses, do not allow for uncertainties to be properly handled.

Moreover, investments in new facilities and expansions of existing capacities are resolved at fairly high managerial levels, and are affected by information which is both external and internal to the firm. Thus, any form of rational analysis of these decisions has necessarily a very broad scope, requiring information to be processed in a very aggregated form to allow for all the dimensions of the problem to be included and to prevent top managers to be distracted by unnecessary operational details.

(2) Tactical Planning: Aggregate Capacity Planning

Once the physical facilities have been decided upon, the basic problem to be resolved is the effective allocation of resources (e.g., production, storage and distribution capacities, work force availabilities, financial and managerial resources, etc.) to satisfy demand and technological requirements, taking into account the costs and revenues associated with the operation of the production and distribution process. When dealing with several plants, with many distribution centers, regional and local warehouses, with products requiring complex multistage fabrication and assembly processes, affected by strong randomness and seasonalities in their demand patterns, these decisions are far from simple. They usually involve the consideration of a medium range time horizon, divided into several periods, and the aggregation of the production items into product families. Typical decisions to be made within this context are utilization of regular and overtime work force, allocation of aggregated capacity resources to product families, accumulation of seasonal inventories, definition of distribution channels and selection of transportation and transshipment alternatives.

(3) Operations Control: Detailed Production Scheduling

After making an aggregated allocation of capacity among product families, it is necessary to deal with the day-to-day operational and scheduling decisions which require the complete disaggregation of the information generated at higher levels into the details consistent with the managerial procedures followed in daily activities. Typical decisions at this level are the assignment of customer orders to individual machines, the sequencing of these orders in the work shop, inventory accounting and inventory control activities, dispatching, expediting and processing of orders, vehicular scheduling, etc.

(4) The Need for a Hierarchical Decision Making System

To deal with these three distinct levels of decisions one has to recognize several complexities. First, the investment, location, allocation and scheduling decisions cannot be made in isolation because they interact strongly among one another; therefore, an integrated approach is required if one wants to avoid the problems of suboptimization. Second, this approach, although essential, cannot be made without decomposing the elements of the problem in some way, within the content of a hierarchical system that links higher level decisions with lower level ones in an effective manner, and in which decisions that are made at higher levels provide constraints for lower level decision making. This hierarchical approach recognizes the distinct characteristics of the type of management participation, the scope of the decision, the level of aggregation of the required information and the time framework in which the decision is to be made. In our opinion, it would be a serious mistake to attempt to deal with all these decisions at once, via a single mathematical model. Even if the computer and methodological capabilities could allow the solution of large detailed integrated production model, which is clearly not the case today, that approach is inappropriate because it is not responsive to the management needs at each level of the organization, and would prevent the interactions between models and managers at each organization echelon.

In designing a system to support the overall production management decisions it is imperative, therefore, to identify ways in which the decision process can be partitioned, to select adequate models to deal with the individual decisions at each hierarchical level, to design linking mechanisms for the transferring of the higher level results to the lower hierarchical levels which includes means to disaggregate information, and to provide quantitative measures to evaluate the resulting deviations from optimal performance at each level. Some suggestions on how to implement such an approach are provided in Hax and Meal [27].

We will now proceed to review the role mathematical programming models have in supporting the tactical and operational decisions related to aggregate capacity planning.

Aggregate Capacity Planning

Whenever the conditions affecting the production process are not stable in time (due to changing demand requirements, cost components, or capacity availability), it is imperative to plan production in an aggregate way to obtain an effective utilization of the available resources. The time horizon of this planning effort is dictated by the nature of the dynamic variations; thus if demand seasonalities are present, it is necessary to incorporate a full seasonal cycle into the planning horizon. Commonly the time horizon varies from six to eighteen months, 12 months being a suitable figure for most planning systems. Since it is usually impossible to consider every fine detail associated with the production process and still maintain such a long planning horizon, it is mandatory to aggregate the information being processed. This aggregation usually takes place by consolidating similar items into product families, different machines into machine centers, different labor skills into labor centers, and individual customers into market regions. The nature of the planning systems to be used, and the technical as well as managerial characteristics of the production activities are the elements that suggest the type of aggregation to be performed. Aggregation forces a consistent set of units to be used. It is also common to express aggregate demand in production hours.

Aggregate capacity planning attempts to satisfy demand requirements by making the best possible utilization of the resources available to the firm. Once the aggregate plan is generated, constraints are imposed on the detailed production scheduling which decides the specific quantities to be produced of each individual item. These constraints normally specify production rates or total amounts to be produced per month for a given product family. In addition, crew sizes, levels of machine utilization and amount of overtime to be used are determined.

When demand requirements do not change with time, and costs and prices are also stable, it might be feasible to bypass entirely the aggregate planning process, provided the resources of the firm are well balanced to absorb the constant requirements. However, when these conditions are not met, serious

inefficiencies might result from attempting to plan production responding only to immediate requirements and ignoring the future consequences of present decisions. To illustrate this point, it is enough to consider what happens when an order point - order quantity inventory control system [1], which treats every item in isolation, is applied in the presence of strong demand seasonalities. Firstly, at the beginning of the peak season demand starts rapidly increasing at which point a large number of items will simultaneously trigger the order point, demanding production runs on the amount specified by the order quantities. Being unable to satisfy all these orders and still maintaining an adequate service level, a normal management reaction is to reduce the production run lengths, thus creating multiple changeovers of small quantities. This, in turn, reduces the overall productivity (because of the high percentage of idle machine time due to the large number of changeovers), increases costs, and deteriorates customer service levels. Secondly, items at the end of the season will be produced in normal order quantities (typically large), thus creating inventory that would be inactive til the beginning of the next season or that would have to be liquidated at salvage values. An effective aggregate capacity planning system will prevent such inefficiencies.

Ways to Absorb Demand Fluctuations

There are several ways that can be used by managers to absorb changing demand patterns. These ways can be combined, creating a large number of possible alternative or strategies to plan production.

(1) Management can change the size of the work force by hiring and laying off which allows changes in the production rate to take place. Excessive use of these practices, however, can create severe labor problems.

(2) While maintaining a uniform regular work force, management can vary the production rate by introducing overtime and/or idle time or relying on outside subcontracting.

[1] For details on the description of an order point - order quantity inventory control system see, for example, Buffa and Taubert [9], or Magee and Boodman [38].

(3) While maintaining a uniform production rate, management can anticipate future demand by accumulating seasonal inventories. The trade-off between the cost incurred in changing production rates and holding seasonal inventories is the basic question to be resolved in most practical situations.

(4) Management can also resort to planned backlogs, whenever customers may accept delays in filling their orders.

(5) An alternative which has to be resolved at a higher planning level is the development of complementary product lines, with demand patterns which are counter seasonal to the existing products. This alternative, although very effective in producing a more even utilization of the firm's resources does not eliminate the need for aggregate planning.

Costs Relevant to Aggregate Capacity Planning

Relevant costs can be categorized as follows (Silver [48]):

(1) Basic production costs. These are the fixed and variable costs incurred in producing a given product type in a given time period. Included are direct and indirect labor costs, and regular as well as overtime compensations.

(2) Costs associated with changes in the production rate. Typical costs in this category are those involved in hiring, training and laying-off personnel.

(3) Inventory holding costs. A major component of the inventory holding cost is the cost of capital tied-up in inventory. Other components are storing, insurance, taxes, spoilage, obsolence, etc.

(4) Backlogging costs. Usually these costs are very hard to measure and include expediting, loss of customer good will, and cost of sales revenues resulting from backlogging.

McGarrah [40] and Holt et.al. [29] provide a good discussion on the nature and structure of these cost elements.

The Role of Models in Aggregate Capacity Planning

Models have played an important role in supporting management decisions in aggregate capacity planning. Anshen et.al. [1] indicate that models are of great value in helping management to:

1. Quantify and use the intangibles which are always present in the background of its thinking but which are incorporated only vaguely and sporadically in scheduling decisions.
2. Make routine the comprehensive consideration of all factors relevant to scheduling decisions, thereby inhibiting judgments based on imcomplete, obvious, or easily handled criteria.
3. Fit each scheduling decision into its appropriate place in the historical series of decisions and, through the feedback mechanism incorporated in the decision rules, automatically correct for prior forecasting errors.
4. Free executives from routine decision-making activities, thereby giving them greater freedom and opportunity for dealing with extraordinary situations.

In order to describe the different types of models that can be used in supporting aggregate planning decisions, it is useful to classify the models according to the assumptions they make regarding the structure of the cost components. In the following sections we will analyze first linear cost models, followed by quadratic cost models, fixed cost models, and then general nonlinear cost models.

Linear Cost Models

Some of the very first models to be proposed to guide aggregate planning decisions assume linearity in the cost behavior of the decision variables. These kinds of models are very popular even today because of the computational conveniences associated with linear programming. Moreover, these models are less restrictive than first appears because nonlinear convex costs functions can be approximated to any degree of accuracy by piecewise linear segments. We will now see two important classes of linear models.

(A) Fixed Work Force

First, let us consider the case where the work force is fixed by disallowing hiring and firing to absorb demand fluctuations during the planning horizon. Production rates can only fluctuate by using overtime from the regular work force.

The following notation will be used to describe the model in mathematical terms.

Parameters:

v_{it} = unit production cost for product i in period t

c_{it} = Inventory carrying cost per unit of product i in period t

r_t = cost per manhour of regular labor in period t

o_t = cost per manhour of overtime labor in period t

d_{it} = demand for product i in period t

k_i = manhours required to produce one unit of product i

$(rm)_t$ = total manhours of regular labor available in period t

$(om)_t$ = total manhours of overtime labor available in period t

I_{io} = initial inventory level for product i

W_o = initial regular workforce level

T = time horizon, in periods

N = total number of products

Decision Variables

X_{it} = units of product i to be produced in period t

I_{it} = units of product i to be left over as inventory at the end of period t

W_t = manhours of regular labor used during period t

O_t = manhours of overtime labor used during period t

A simple version of the fixed work force - linear cost model is

$$\text{Min } z = \sum_{i=1}^{N} \sum_{t=1}^{T} (v_{it} X_{it} + c_{it} I_{it}) + \sum_{t=1}^{T} (r_t W_t + o_t O_t) \qquad (1)$$

subject to:

$$X_{it} + I_{i,t-1} - I_{it} = d_{it} \qquad \begin{matrix} t=1,\ldots,T \\ i=1,\ldots,N \end{matrix} \qquad (2)$$

$$\sum_{i=1}^{N} k_i X_{it} - W_t - O_t = 0 \qquad t=1,\ldots,T \qquad (3)$$

$$0 \leq W_t \leq (rm)_t \qquad t=1,\ldots,T \qquad (4)$$

$$0 \leq O_t \leq (om)_t \qquad t=1,\ldots,T \qquad (5)$$

$$X_{it},\ I_{it} \geq 0 \qquad \begin{matrix} i=1,\ldots,N \\ t=1,\ldots,T \end{matrix} \qquad (6)$$

The objective function (1) expresses the minimization of variable production, inventory, and regular and overtime labor costs. If the marginal production costs v_{it} are invariant over time, the terms $v_{it}X_{it}$ do not need to be included in the objective function. (Since total production is fixed). Similarly, if the payroll of regular work force W_t constitutes a fixed commitment, the terms $r_t W_t$ should be deleted from (1).

Constraints (2) represent the typical production-inventory balance equation. Notice that (2) and (6) imply that no backordering is allowed. The next model will show how backorders can be incorporated. Moreover, (2) assumes a deterministic demand, d_{it}, for every item in every time period. One way to allow for uncertainties in the demand forecast is to specify a lower bound for the ending inventory at each period, i.e., $I_{it} \geq ss_{it}$, where ss_{it} is the safety stock associated with item i in period t.[1]

---

[1] The magnitude of the safety stocks depends on the quality of the demand forecasts and the level of customer service to be provided. For a good discussion of how to compute safety stocks see Brown [8].

Constraints (3) guarantee that the total manpower to be used at every period does not exceed the regular and overtime work force available. This model formulation assumes that manpower availability is the only constraining resource of the production process. It is a trivial matter to expand the number of resources being considered, provided that linearity assumptions are maintained.

Constraints (4) and (5) pose lower and upper bounds on the use of regular and overtime manhours in every time period.

We already have indicated how constraints (6) could be changed to incorporate safety stocks. One should bear in mind that if no terminal conditions are imposed to the inventories at the end of the planning horizon, the model will drive them to zero, i.e., it will make $I_{iT} = 0$ for all i. If total depletion of inventories is undesirable, a target inventory constraint should be added in the model. An additional constraint should also be attached if there are storage requirements that cannot be exceeded; for example, the constraint

$$\sum_{i=1}^{N} I_{it} \leq (sc)_t \qquad , t=1,\ldots,T$$

implies that the total inventory at each period cannot be greater than the total storage capacity $(sc)_t$.

When it is necessary to assign products to different working centers with limited capacities, it is required to redefine the decision variables to identify those decisions explicitly. For example, X_{ict} may be used to denote the amount of product i to be produced at working center c during period t. It is straightforward to carry out the resulting transformations in the overall model.

Even the very simple model described by expressions (1) to (6) could present enormous computational difficulties if the individual items to be scheduled are not grouped in broad product categories. If we ignore constraints (4), (5), and (6), which merely represent upper and lower bounds for the decision variables, the model consists of Tx(N+1) effective constraints.

When dealing with complex production situations, the total number of individual items, N, may be several thousands. For example, if the planning model has 12 time periods and 5,000 items the model would have about 60,000 constraints, which exceeds the capabilities of a regular linear programming code. One could argue that constraints (2), accounting for most of the effective constraints, are of a generalized upper bounded type and, therefore, the model could be computationally feasible if a generalized upper bounded code is available.

In most practical applications, however, it would not be functional to plan the allocations of the production resources at this level of detail. First, a detailed scheduling program should take into account a large number of technological and marketing considerations which cannot be include in the overall model due to their highly qualitative nature. Second as we have expressed before, many of the planning issues to be resolved with the model deal with broad allocations of resources, and excessively detail information will obscure rather than enlighten these decisions. Third, aggregate forecasts are more accurate than detailed forecasts.

It is common practice, therefore, to aggregate items in family types. The criteria for aggregation are evident from the model structure in order members of a single family type should share similar demand patterns (d_{it}), and should require similar unit production time (k_i). Once the aggregate planning decisions are made, these decision impose constraints that have to be observed when performing detailed item scheduling.

Notice that this model, as well as any other planning model, requires the definition of a planning horizon T and the partitioning of this time horizon into multiple time periods. People normally assume that this partitioning results in T equally spaced time periods. Of course, this does not need to be so. As a matter of fact, many operational planning systems will be better designed if this partitioning generates uneven time periods, so that the more recent time periods carry more detailed information. Due to the uncertain environment in which this planning effort

is being conducted, only the first time period results are usually implemented. At the end of every time period new information becomes available which is used to update the model and recompute the next time period plans.

Broad technological, institutional, marketing, financial and organizational constraints can also be included in the model formulation. This flexibility, characteristic of the linear programming approach to problem solving, has made this type of model very useful and popular.

A simple version of the fixed work force linear programming model, having a transportation problem structure, was first proposed by Bowman [7].

(B) Variable Work Force

Whenever it is feasible to change the work force during the planning horizon as a way to counteract demand fluctuations, the composition of the work force becomes a decision variable, whose values can change by hiring and firing personnel. Therefore, the corresponding hiring and firing costs should be part of the objective function. In addition, shortages will be accepted and a backordering cost has to be part of the model formulation. The model decision variables are

$\dot{X}_{it}$ = units of product i to be purchased at period t

W_t = manhours of regular work force at period t

O_t = manhours of overtime work force at period t

H_t = manhours of regular work force hired at period t

F_t = manhours of regular work force fired at period t

I^+_{it} = units of ending inventory for product i at period t

I^-_{it} = units backordered for product i at the end of period t

Using the above notation with that introduced in the previous model, the cost incurred during period t includes the following components:

Variable manufacturing cost	$v_{it}X_{it}$
Inventory holding cost	$c_{it}I^+_{it}$
Backorder cost	$b_{it}I^-_{it}$
Regular payroll cost	r_tW_t
Overtime payroll cost	o_tO_t

Hiring cost	$h_t H_t$
Firing cost	$f_t F_t$

A simple version of the variable work force model can be formulated as:

$$\text{Min } z = \sum_{i=1}^{N} \sum_{t=1}^{T} (v_{it}X_{it} + c_{it}I_{it}^{+} + b_{it}I_{it}^{-}) + \sum_{t=1}^{T} (r_t W_t + o_t O_t + h_t H_t + f_t F_t) \quad (1)$$

subject to:

$$X_{it} + I_{i,t-1}^{+} - I_{i,t-1}^{-} - I_{it}^{+} + I_{it}^{-} = d_{it} \qquad \begin{cases} i = 1,\ldots,N \\ t = 1,\ldots,T \end{cases} \quad (2)$$

$$\sum_{i=1}^{N} k_i X_{it} - W_t - O_t \leq 0 \qquad t = 1,\ldots,T \quad (3)$$

$$W_t - W_{t-1} - H_t + F_t = 0 \qquad t = 1,\ldots,T \quad (4)$$

$$-pW_t + O_t \leq 0 \qquad t = 1,\ldots,T \quad (5)$$

$$X_{it},\ I_{it}^{+},\ I_{it}^{-} \geq 0 \qquad \begin{cases} i = 1,\ldots,N \\ t = 1,\ldots,T \end{cases} \quad (6)$$

$$W_t,\ O_t,\ P_t,\ H_t,\ F_t \geq 0 \qquad t = 1,\ldots,T \quad (7)$$

The objective function (1) expresses the minimization of all the variable costs incurred during the model planning horizon.

Constraints (2) represent the production-inventory balance equation. Notice that this is equivalent to the old balance equation

$$X_{it} + I_{i,t-1} - I_{it} = d_{it}$$

except that now

$$I_{i,t-1} = I^{+}_{i,t-1} - I^{-}_{i,t-1}$$

and $I_{it} = I^{+}_{it} - I^{-}_{it}$.

In the present model the ending inventory, I_{it}, can be either positive ($I^{+}_{it} > 0$ indicates that stock remains at the end of the period), or negative ($I^{-}_{it} > 0$ indicates an accumulation of backorders at the end of the period). Since there is a cost attached to both I^{+}_{it} and I^{-}_{it} those variables will never be both positive simultaneously.

Constraints (3) requires the total manpower, both regular time and overtime, to be at least equal to the total number of manhours used in each period, i.e. $\sum_{i=1}^{N} k_i X_{it} \geq W_t + O_t$.

Constraints (4) define the change in the work force size during period t, i.e. $W_t - W_{t-1} = H_t - F_t$. Labor has been added whenever $H_t > 0$, or has been subtracted whenever $F_t > 0$. Once again, since there is a cost attached to both hiring and firing, H_t and F_t will never simultaneously have positive values in a given time period.

Constraints (5) impose an upper bound on the total overtime available i- period t as a function of the regular work force size, i.e. $O_t \leq pW_t$, where p is the percentage of overtime allowed to the regular work force.

Constraints (6) and (7) are the satndard non-negativity requirements on the decision variables.

Many of the comments we have made in the fixed work model regarding ways to expand or simplify the models and ways to aggregate items in item families are applicable here and will not be repeated.

The first of this type of models was proposed by Hanssmann and Hess [26]. Several alternative approaches have been suggested, particularly those by Von Lazenaur [55], and O'Malley, Elmaghraby and Jeske [44].

(C) Advantages and Disadvantages of Linear Cost Models

The overwhelming advantage of linear cost models is that they generate linear programs which can be easily solved by readily available and efficient computer codes. Linear programs permit models with a large number of decision variables and constraints to be solved expediently and cheaply. In addition, linear programming lends itself very well to the performance of parametric and sensitivity analyses, a feature which can be of great help in making aggregate planning decisions. The shadow cost information can be of assistance in identifying opportunities for capacity expansions, marketing penetration strategies, new product introductions, etc.

As indicated before, the linearity assumptions which are implicit in these models are less restrictive than might appear. First, cost structures might behave linearly within the range of interest of the decision variables under consideration. Second, general convex separable functions can be treated with piecewise linear approximation. Moreover, with some ingenuity certain functions which at first seem to present nonlinear characteristics can be linearized, as indicated in the cited references of Hanssmann and Hess [26], and Von Lazenaur [55].

The most serious disadvantage of linear programming models is their failure to deal with demand uncertainties in any explicit way. In some situations this could constitute a serious drawback. However, Dzielinski, Baker and Manne [14] have reported favorable experiences in using linear programming models under fairly uncertain and dynamic environments.

Quadratic Cost Models (Linear Decision Rules)

Whenever quadratic cost models are used to solve the aggregate capacity planning problem, the decision rules that are generated posses a linear structure (because the differentiation of a quadratic function produces a linear function). Thus, these models are also known as Linear Decision Rules. The first model of this kind was developed by Holt, Modigliani, Muth and Simon (HMMS), see reference [29]. Subsequently, several extensions have been offered. We will now discuss the basic concepts underlying the HMMS model.

The HMMS model calls for a complete aggregation of all product types into a single category. This might require the use of appropriate compatible units that allow for this transformation to be made. Thus, there are essentially two decision variables:

P_t = aggregate production rate for period t

W_t = work force size at period t

The remaining decision variable

I_t = ending inventory at period t, is specified automatically by the values of P_t, and W_t, and the relationship that exists among the three variables. The optimum decision rules, therefore, require specification of the aggregate production and work force for each period that minimize a quadratic cost function.

(A) Cost Components

Let us now review in some detail the components of this quadratic cost function. The following cost categories are identified:

(1) Regular Payroll Costs

These costs are assumed to increase linearly with the workforce size, according to the following relationship:

$$c_1 W_t + c_{13}$$

Where c_1 and c_{13} are cost coefficients to be determined externally to the model. Since c_{13} is a constant, it can be eliminated from further consideration.

(2) Hiring and Firing Costs

Both hiring and firing costs are assumed quadratic in the work force variation $(W_t - W_{t-1})$, thus allowing an increasing cost rate to be incorporated. The specific relationship is a U-shaped curve given by:

$$c_2(W_t - W_{t-1} - c_{11})^2$$

where c_2 and c_{11} are constants to be evaluated. c_{11} is introduced to allow for asymmetry is the cost function.

(3) Overtime and Idle Costs

Given a work force size W_t there is a desirable production rate $c_4 W_t$. If the production rate exceeds that amount, there will be overtime cost; if it is lower than that amount there will be an idle cost. The exact nature of these cost relationships is given by the expression

$$c_3(P_t - c_4 W_t)^2 + c_5 P_t - c_6 W_t + c_{12} P_t W_t$$

where the three last terms are given to improve the accuracy of the cost relationships.

(4) Inventory and Backorder Costs

The relationship which characterizes the inventory related costs is assumed to be of the following form:

$$c_7 [I_t - (c_8 + c_9 d_t)]^2$$

where

d_t = expected units of aggregate product demand at period t.

The target inventory level is $c_8 + c_9 d_t$; when deviations occur from this target, either carrying or backorder costs are incurred which increase with the square of these deviations. In the original HMMS work, c_9 was set to zero.

The estimation of the cost coefficients is an expensive and time consuming activity requiring statistical analysis, accounting information and managerial inputs. Extensive work has been done to improve the quality of these estimates (Van dePanne and Bosje [53], Kriebel [36]) and develop aggregate cost functions which represent the cost characteristics of the individual items (Bergstrom and Smith [4], Krajewski et al [35]).

(B) Model Formulation

Given the cost structure discussed above, the aggregate capacity planning model can be formulated as:

$$\text{Min } z = \sum_{t=1}^{T} [(c_1-c_6)\, W_t + c_2(W_t-W_{t-1}-c_{11})^2 + c_3(P_t-c_4\, W_t)^2+c_5 P_t + c_{12}\, P_t W_t + c_7(I_t - c_8 - c_9\, d_t)^2] \quad (1)$$

subject to

$$P_t + I_{t-1} - I_t = d_t \qquad t=1,\ldots,T \quad (2)$$

$$P_t,\ W_t \geq 0 \qquad t=1,\ldots,T \quad (3)$$

The objection function (1) should be ragarded as the minimization of expected costs. One of the interesting features of the model is that it does not assume the demand d_t to be deterministic. Holt et al [29] proved that if the demand forecasts are unbiased and represent expected values, the linear decision rules resulting from the minimization of (1) subject to constraints (2) and (3), provide minimum expected costs.

(C) The Linear Decision Rules

The above model will have a unique global minimum if the objective function is strictly convex. This condition usually is met by all the cost functions encountered in practice since the cost components normally have increasing marginal costs.

Optimal solution to the model are found by the use of Lagrangians. Several applications have been reported which illustrate the nature of the resulting rules (See Buffa and Taubert [9]. In general, the form of the rules can be characterized by equations of the following type

$$P_t = a_o d_t + a_1 d_{t+1} + \ldots + a_{T-t} d_t + bW_{t-1} + c - d I_{t-1} \quad (4)$$

$$W_t = e_o d_t + e_1 d_{t+1} + \ldots + e_{T-t} + fW_{t-1} + g - hI_{t-1} \quad (5)$$

Equation (4) describes the nature of the aggregate production rate which is dependent on future demand forecasts, previous work force size, and beginning inventory. Same comments apply to expression (5) that illustrates the form of the aggregate work force decision. The weights given to the demand forecasts (the a's and e's) decreased rapidly with time.

(D) Extensions to the HMMS Model

Several extensions to the initial HMMS model have been reported in the literature. Bergstrom and Smith [4] generalized the approach to a multiproduct formulation, and incorporated revenues in the objective function. Chang and Jones [10] also dealt with the multiproduct problem, and suggested procedures to solve situations when production cannot be started and completed in a given time period. Sypkens [51] included plant capacities as an additional decision variable.

(E) Advantages and Disadvantages of Quadratic Cost Models

The major advantages of quadratic cost models are that they allow for more realistic cost structure in the planning process, and provide linear decision rules which are easy to solve and implement, and they allow uncertainties to be handled directly since the linear decision rulse minimize the expected cost, provided that unbiased expected demand forecasts are given.

The more serious drawbacks are the strong need for aggregation, the elaborate estimation procedures that are required to assess the numerical of the cost coefficients, and the numerical difficulties encountered when the number of decision variables and constraints increase, which limits the model dimensions to a small size.

Computational results (Van dePanne and Bose [53]) seem to indicate that decision rules are fairly insensitive to large errors in estimating cost parameters. This is a very attractive property due to the difficulty in providing accurate costs values.

In spite of the encouraging results reported on large savings that have been obtained by applying linear decision rules to actual managerial situations, these techniques have not been adopted by practicing managers. Probably the disadvantages listed seem to outrule the advantages that linear decision rules have vis-a-vis linear programming models. Comparisons made by Kolenda [34] between HMMS and Hanssmann-Hess type of models rank these two approaches as very closed in overall efficiency. Given the enormous computational capabilities of linear programming, this has to result in a more widespread use of linear cost models.

Lot Size Models (Fixed Cost Models)

Whenever the manufacturing process is characterized by batch-type production operations (as opposed to continuous production), a cost is incurred when setting up the production facilities for a given run. Including the setup cost in the planning process creates many problems. First, every item that generates a setup (or a family of items sharing a common setup) has to be identified and treated independently. This expands the number of variables and constraints so that the dimensions of the model generate a large scale system which can only be coped with by using special computational techniques. Second, the inclusion of setup costs produces a problem of lot-size indivisability since a given batch has to be run incurring a single setup. This introduces the presence of integer variables in the model formulation. Finally, setup costs give rise to fixed cost components in the objective function. Moreover, the downtime which is characteristic of every setup operation introduces additional nonlinearities in the constraint set. The resulting large scale, integer, nonlinear programming model is hard to resolve computationally. We will now review some of the most effective approaches that have been suggested to solve this problem.

(A) The Uncapacitated Lot Size Model

The standard economic lot size formula, also known as the EOQ (economic order quantity) formula, determines the production amount for an individual item when setup and inventory holding costs identify the cost trade-offs.[1] This formula does not account for any interaction that may exist among the individual items to be scheduled for production. In particular, it ignores the capacity limitations which impose some of the most critical constraints for production planning.

Moreover, the EOQ formula assumes the demand to be constant and known during the whole planning horizon. When the demand is known but changes during the various time periods of the planning horizon, the EOQ lot-size can provide very misleading recommendations. Wagner and Within[57] suggested a dynamic programming model for a dynamic version of the economic lot size. We will review their approach here because it plays an important role in the capacitated lot size models to be discussed later.

A simplified version of the uncapacitated lot size problem can be described as follows:

$$\text{Minimize} \quad z = \sum_{t=1}^{T} [s_t \, \delta(X_t) + c_t I_t]$$

$$\text{subject to} \quad X_t + I_{t-1} - I_t = d_t \quad , t=1,\ldots,T$$

$$X_t \geq 0 \quad , t=1,\ldots,T$$

$$I_t \geq 0 \quad , t=1,\ldots,T$$

$$\text{where} \quad \delta(X_t) = \begin{cases} 0 \text{ if } X_t = 0 \\ 1 \text{ if } X_t > 0 \end{cases}$$

---

[1] For discussion on the various types of EOQ formulae that have been proposed in the literature, see Magee and Boodman [38].

and as before:

X_t = amount to be produced in period t
I_t = ending inventory at period t
s_t = setup cost in period t
c_t = inventory holding cost in period t
d_t = demand during period t

Notice that we have eliminated the variable production costs since we are allowing no backorders and the total production is fixed. It is a trivial matter to add these costs, if necessary. A dynamic programming solution to this problem is straightforward. The functional equation that represents the minimum cost policy (including only setup and inventory holding costs) for periods t through T is:

$$f_t(I_{t-1}) = \min_{\substack{X_t \geq 0 \\ X_t + I_{t-1} \geq d_t}} \left[s_t \delta(x_t) + c_t(x_t + I_{t-1} - d_t) + f_{t+1}(x_t + I_{t-1} - d_t) \right]$$

In the last period T, the functional equation becomes:

$$f_T(I_{t-1}) = \min_{\substack{X_T \geq 0 \\ X_T + I_{T-1} = d_T}} \left[s_T \, \delta(x_T) \right]$$

Thus, a backward induction process can be applied to compute the optimum lot sizes during the planning horizon.

Wagner and Whitin proved that it is enough to consider production sequences that only produce integral periods of demand, where production only takes place when the level of inventory is zero. This implies that a given demand requirement is satisfied from the production run that occurs in the nearest preceding time period in which a setup was incurred and that production meets all demand inventory again goes to zero. Thus, when dealing with a time horizon of T time periods, the total number of production sequences to consider is 2^{t-1}. The dynamic programming approach requires the analysis of only $T(T+1)/2$ of these sequences. This number can be further reduced by

using the Planning Horizon Theorem, which states that whenever it is optimal to make $I_t=0$, periods 1 through t can be considered by themselves. This makes advantageous to conduct a forward induction process and partition the original problem into subproblems, whenever an optimal policy with $I_t=0$ is found.

The functional equation that characterizes the forward induction procedure and takes advantage of the dominant production sequences can be now specified. Let f(t) be the minimal cost program from period 1 to t, then

$$f(t) = \min \left\{ \min_{1 \le j \le t} \left[s_j + \sum_{h=j}^{t-1} \sum_{k=h+1}^{t} c_h d_k + f(j-1) \right], \; s_t + f(t-1) \right\}$$

where $f(1) = s_1$ and $f(0) = 0$

in here s_j represents the setup cost at period j, and $\sum_{h=j}^{t-1} \sum_{k=h+1}^{t} c_h d_k$ provides the inventory carrying cost from periods j+1 to t.

Numerical examples illustrating how to carry out the forward induction procedures are provided in the original reference of Wagner and Whitin [57].

The concept of dominant production sequences has been greatly exploited for computational purposes when dealing with capacitated lot size models, as we will see in the subsequent sections.

Wagner [56] expanded this approach to include changing purchasing or manufacturing costs during the multiperiod planning horizon; in addition, Eppen, Gould and Pashigian [18] developed a new planning horizon theorem. Zangwill [61] showed how to treat backordering cost; and Bomberger [5], Stankard and Gupta [49], and Hodgson [28] extended the dynamic programming approach to cover some interactions among multiple items.

(B) The Capacitated Lot Size Model

The capacitated lot size model deals with a multi-item production planning problem under changing demand requirements during the multi-period planning

horizon. The items are competing for limited capacity, and setup costs become an important element of the total cost to be minimized.

As before, we will analyze first the fixed work force problem when only overtime can be added to expand the manpower availability, and subsequently we will examine the variable work force problem, when hiring and firing is permitted to change the total production rate.

(1) Fixed Work Force Model

Using the notation presented in the previous pages, a simple version of the fixed work force - capacitated fixed cost model can be expressed as follows:

$$\text{Min } z = \sum_{i=1}^{N} \sum_{t=1}^{T} [s_{it}\delta(X_{it}) + v_{it}X_{it} + c_{it}I_{it}] + \sum_{t=1}^{T} (r_t W_t + o_t O_t) \tag{1}$$

subject to:

$$X_{it} + I_{i,t-1} - I_{it} = d_{it} \qquad \begin{cases} t=1,\ldots,T \\ i=1,\ldots,N \end{cases} \tag{2}$$

$$\sum_{i=1}^{N} [a_i\delta(X_{it}) + k_i X_{it}] - W_t - O_t \leq 0 \qquad t=1,\ldots,T \tag{3}$$

$$0 \leq W_t \leq (rm)_t \qquad t=1,\ldots,T \tag{4}$$

$$0 \leq O_t \leq (om)_t \qquad t=1,\ldots,T \tag{5}$$

$$X_{it},\ I_{it} \geq 0 \qquad \begin{cases} i=1,\ldots,N \\ t=1,\ldots,T \end{cases} \tag{6}$$

where

$$\delta(X_{it}) = \begin{cases} 0 \text{ if } X_{it} = 0 \\ 1 \text{ if } X_{it} > 0 \end{cases} \tag{7}$$

Most of the comments we made when dealing with the fixed work force - linear cost model are also applicable now and will not be repeated. This model does not allow backorders, although it will be to incorporate this added feature in the model formulation.

In expression (3) above, the term a_i represents the setup time consumed in preparing a production run for item i. The presence of $\delta(X_{it})$ both in the objective function (1) and in the constraints (3) completely breaks the linearity conditions of our previous models, and makes the computation of this model much more difficult. We will now examine some of the methods that have been proposed to solve the model.

(1.1) Fixed Cost Model

Whenever the down time consumed by the setup operation is negligable, $a_i=0$ in expression (3) and the lot size fixed work force model becomes a fixed cost linear programming model, also known as the fixed charge model. Since the objective function of the fixed charge model is concave and the constraint set is convex, the global minimum will occur at an extreme point. However, generally, many local minima will also exist at extreme points which make a simplex type algorithm that terminates at a local minimum not very effective to use.

Several approaches have been suggested to deal with this problem. Exact solution methods can be classified in two different categories: extreme point ranking procedures (Gray [23], and Murty [41]), and branch and bound solution to mixed integer programming formulations of the problem. (Jones and Soland [32], and Steinberg [50]). Exact methods are computationally limited to relatively small size problems, and therefore have little practical value at the present time. As a result of this limitation, several heuristic approaches have been proposed that generate near-optimal solutions. Generally, these heuristics start by producing a good extreme point solution, and, by examining the adjacent extreme points, a local minimum is determined. Then, a move is made to an extreme point away from this local minimum, and the process is repeated until no further improvement is obtained or after completing a specified number of iterations. Effective heuristics have been provided by Balinski [3], Cooper and Drebes [11], Denzler [13], Roussean [46], and Steinberg [50].

(1.2) Linear Programming Approach

When the downtime, a_i, required to setup a production run for every item is not negligable, the resulting large scale non-linear capacitated lot size model becomes extremely hard to solve in a direct way. Realizing

this computational difficulties, Manne [39] suggested to reformulate the problem as a linear programming model. This approach was subsequently refined by Dzielinski, Baker and Manne [14], Dzielinski and Gomory [15], and Lasdon and Terjung [37].

Essentially, the approach consists in incorporating setup costs by defining a set of possible production sequences. For a given item i a production sequence over the planning horizon T is a set of T non-negative integers that identifies the quantities of item i to be produced at each time period during the planning horizon so that the demand requirements for that item are met. As explained in the uncapacitated lot size model, it is enough to consider 2^{T-1} dominant sequences for each item.

Let us define

X_{ijt} = amount to be produced of item i by means of production sequence j in period t; i=1,...,N; j=1,...,J; t=1,...,T.

and as usual, let

d_{it} = demand for item i in period t.

To illustrate how these sequences are constructed, let us assume we have only three time periods. The number of dominant sequences for item i is $2^{3-1} = 4$; these four strategies for a given item i can be defined as follows:

Amounts to be produced at each sequence

Sequence No.	Time period t=1	t=2	t=3
j=1	$X_{i11}=d_{i1}+d_{i2}+d_{i3}$	$X_{i12} = 0$	$X_{i13} = 0$
j=2	$X_{i21}=d_{i1}+d_{i2}$	$X_{i22} = 0$	$X_{i23} = d_{i3}$
j=3	$X_{i31} = d_{i1}$	$X_{i32} = d_{i2}+d_{i3}$	$X_{i33} = 0$
j=4	$X_{i41} = d_{i1}$	$X_{i42} = d_{i2}$	$X_{i43} = d_{i3}$

It is easy to compute the total production, inventory holding and setup costs, t_{ij}, for each sequence. In the above example these costs are the following:

Setup and holding costs for each sequence (t_{ij})

Sequence	t_{ij}
j=1	$t_{i1}=s_{i1} + v_{i1}(d_{i1}+d_{i2}+d_{i3}) + c_{i1}(d_{i2}+d_{i3}) + c_{i2}(d_{i3})$
j=2	$t_{i2}=(s_{i1}+s_{i3}) + v_{i1}(d_{i1}+d_{i2}) + v_{i3}d_{i3} + c_{i1}d_{i2}$
j=3	$t_{i3}=(s_{i1}+s_{i2}) + v_{i1}d_{i1}+v_{i2}(d_{i2}+d_{i3}) + c_{i2}d_{i3}$
j=4	$t_{i4} = (s_{i1} + s_{i2} + s_{i3})$

In general

$$t_{ij} = \sum_{t=1}^{T} [s_{it}\ \delta(X_{ijt}) + v_{it}X_{ijt} + c_{it}\ I_{it}] \tag{8}$$

the total labor resources consumed by the production quantities X_{ijt} can be written as:

$$\ell_{ijt} = \begin{cases} 0 & \text{if } X_{ijt} = 0 \\ a_i + k_i X_{ijt}\ , & \text{if } X_{ijt} > 0 \end{cases} \tag{9}$$

If we assume, to simplify matters, that we have a prescribed work force at every time period, $(rm)_t$, that we cannot exceed, the fixed work force lot size model can be formulated as follows:

$$\text{Minimize} \quad z = \sum_{i=1}^{N} \sum_{j=1}^{J} t_{ij}\theta_{ij} \tag{10}$$

subject to

$$\sum_{i=1}^{N} \sum_{j=1}^{J} \ell_{ijt}\theta_{ij} \leq (rm)_t \quad , \quad t=1,\ldots,T \tag{11}$$

$$\sum_{j=1}^{J} \theta_{ij} = 1 \quad , \quad i=1,\ldots,N \tag{12}$$

$$\theta_{ij} \geq 0 \quad \begin{cases} i=1,\ldots,N \\ j=1,\ldots,J \end{cases} \tag{13}$$

where

J = total number of dominant production sequences, and

θ_{ij} = fraction of the j^{th} production sequence used to produce item i

Expression (10) states the objective of the model as the minimization of variable production, setup and inventory holding costs. It is easy to expand the model to include regular and overtime labor costs, shortage costs, and hiring and firing costs. Constraints (11) force the total manpower consumed in the production schedules not to exceed the maximum labor availability at each time period. It is also simple to consider several types of production resources, and to include a variable work force as a decision variable with overtime capabilities (See Dzielinski, Baker and Manne [14], and Dzielinski and Gomory [15] for these model extensions).

Constraints (12) indicate that for every item i one production sequence, θ_{ij}, should be selected. Constraints (13) are the trivial non-negativity requirements on the decision variables. Naturally, θ_{ij} should be further restricted to take only values 0 or 1. However, the inclusion of these integrality constraints would make a problem of this size impossible or very expensive to compute. What we will seek, therefore, is a continuous approximation to an integer programming problem. Fortunately this approximation is usually quite satisfactory. Since there are T+N constraints in the model, there will be at most T+N positive variables in the optimum linear programming solution, and at least one of these variables will be associated with each of the N constraints (12). Thus, there could be at most T instances for which more than one θ_{ij} is positive. If only one θ_{ij} is positive for a given item i that value of θ_{ij} should be 1, due to constraints (12). Consequently, whenever N is much greater than T, which occurs in most practical applications, the θ_{ij} fractional problems do not have much significance.

As we have indicated before, it is easy to expand this model to include not only manpower availabilities but any number, K, of limited resources. When this is the case, the split of production sequences is not significant whenever N (the number of items to be scheduled) is much greater than K x T (the number of resources times the number of time periods). This condition is usually satisfied in practice.

Regardless of the integrality problems posed by the variables θ_{ij}, the resulting linear program is hard to solve by conventional methods. In some situations there might be several thousands of items to schedule, and a model with that many rows can be impossible to compute with regular simplex procedures. In addition, each item generates 2^{T-1} dominant production sequences. If T=12, there will be $2^{12-1} = 2048$ variables for each item, and if there are one thousand items to schedule, the model will have more than two million θ_{ij} variables.

To bypass these difficulties, Dzielinski and Gomory [15] suggested a Dantzig-Wolfe decomposition approach where the sub-problems led to uncapacitated lot size models of the Wagner-Whitin type. These subproblems, which can be computed quite simply, are used to generate attractive entering production sequences so that there is no need to specify all the θ_{ij} variables from the very beginning.

The decomposition approach, however, has one severe limitation for this type of problem. As it is well known, the decomposition technique finds a near optimum solution relatively fast, but a large number of iterations might be spent in obtaining the optimum. In most applications it is not very critical to get the final optimum. Lower bounds can be evaluated to determine how good an approximation to the optimum the current solution is, and stopping rules can be designed accordingly. In our problem, however, it is important to obtain the optimum since only then the integrality requirements for the production sequences are satsified, and a feasible solution to the original problem is found.

To resolve this limitation, Lasdon and Terjung [37] maintained the column generation procedure suggested by Dzielinski and Gomory (thus bypassing the computational problem introduced by the large number of columns), but instead of defining a decomposition master program, they solved the original linear programming formulation using generalized upper bounding techniques (Dantzig and Van Slyke [12]) taking advantage of the structure of the initial model.

We will now proceed to explain how the column generation procedure works. Let π_t , t=1,...,T, be the set of dual variables associated with constraints (11), and π_{T+i}, i=1,...,N, be the dual variables associated with constraints (12). The reduced costs corresponding to problem (10) to (13) are given by the expression

$$\bar{t}_{ij} = t_{ij} - \sum_{t=1}^{T} \pi_t \ell_{ijt} - \pi_{T+1} \tag{14}$$

To choose the entering variable we want to find

$$\min_i \quad \min_j \quad \bar{t}_{ij}$$

by introducing the values of t_{ij} and ℓ_{ijt} given by expressions (8) and (9), respectively, and rearranging the terms this minimization requires us to compute for each product i the optimum sequence j since that

$$\min_j \left[\sum_{t=1}^{T} (s_{it} - \pi_t) \delta(X_{ijt}) + (v_{ij} - \pi_t k_i) X_{ijt} + C_{it} I_{it} \right] \tag{15}$$

Since $\pi_t \leq 0$, the above coefficients are all positive. The problem then involves a minimization of setup, variable manufacturing and inventory carrying costs so that the production quantities X_{ijt} satisfy the demand requirements for item i over the multiperiod planning horizon. As we can recall, this is the uncapacitated lot size problem that can be resolved by the dynamic programming approach of Wagner and Whitin.

To determine which column to enter in the basis, we should subtract π_{T+1} from the optimum value of expression (15) corresponding to each item i. The minimum of these quantities identifies the entering column. The new linear programming problem thus generated is solved by the standard generalized upper bounded techniques.

Since the Lasdon and Terjung approach constitutes a continuous approximation to an integer programming problem, it is only applicable when the number of items, N, is much greater than the number of time periods, T. To eliminate

this shortcoming, Newson [43] suggested a heuristic procedure which is independent of column generation techniques and treats the lot size problem as a shortest route problem.

(2) Variable Work Force Model

In this model the work force size becomes also a decision variable. Using the notation defined previously, the model can be formulated as follows:

$$\text{Min } z = \sum_{i=1}^{N} \sum_{t=1}^{T} [s_{it}\, \delta(X_{it}) + v_{it}X_{it} + C_{it}I_{it}] + \sum_{t=1}^{T} (r_t W_t + o_t O_t + h_t H_t + f_t F_t) \tag{1}$$

subject to:

$$X_{it} + I_{i,t-1} - I_{it} = d_{it} \qquad \begin{cases} i=1,\ldots,N \\ t=1,\ldots,T \end{cases} \tag{2}$$

$$\sum_{i=1}^{N} [a_i\, \delta(X_{it}) + k_i X_{it}] - W_t - O_t \leq 0 \quad , \quad t=1,\ldots,T \tag{3}$$

$$W_t - W_{t-1} - H_t + F_t = 0 \quad , \quad t=1,\ldots,T \tag{4}$$

$$-_p W_t + O_t \leq 0 \quad , \quad t=1,\ldots,T \tag{5}$$

$$X_{it} \quad I_{it} \geq 0 \qquad \begin{cases} i=1,\ldots,N \\ t=1,\ldots,T \end{cases} \tag{6}$$

$$W_t,\ O_t,\ H_t,\ F_t \geq 0 \quad , \quad t=1,\ldots,T \tag{7}$$

where:

$$\delta(X_{it}) = \begin{cases} 0 \text{ if } X_{it} > 0 \\ 1 \text{ if } X_{it} = 0 \end{cases} \tag{8}$$

The interpretation of the model should be now straight forward to the reader. One could easily add backorder costs following the procedure suggested in the linear cost - variable work force model.

The solution procedures used to deal with this model are identical to those employed with the lot size - fixed work force model; that is, a fixed cost model is generated whenever the downtime incurred in manufacturing setup (a_i) is negligable, otherwise the linear programming approximations suggested by Dzielinski and Gomory, or Lasdon and Terjung can be applied.

Newson [43] proposed to attack the problem in two stages. The first stage deals with the detailed scheduling decision for each individual item over the multiperiod planning horizon, neglecting the manpower constraints. For a given product i this stage can be formulated as follows:

$$\text{Min} \quad z_i = \sum_{t=1}^{T} [s_{it}(X_{it}) + v_{it}X_{it} + C_{it}I_{it}]$$

subject to

$$X_{it} + I_{i,t-1} - I_{it} = d_{it} \quad , \quad t=1,\ldots,T$$

$$X_{it}, I_{it} \geq 0 \quad , \quad t=1,\ldots,T$$

After solving this model for each of the N items, the capacity required by the detailed schedule for each time period t is computed as:

$$\hat{P}_t = \sum_{i=1}^{N} [a_i \delta(X_{it}) + k_i X_{it}] \qquad t=1,\ldots,T$$

Then the second stage model, dealing with the aggregate capacity decision is solved. The model is defined as follows:

$$\text{Min} \quad z(\hat{P}) = \sum_{t=1}^{T} (r_t W_t + o_t O_t + h_t H_t + f_t F_t)$$

subject to

$$W_t + O_t - \hat{P}_t \geq 0 \quad , \quad t=1,\ldots,T$$

$$W_t - W_{t-1} - H_t + F_t = 0 \quad , \quad t=1,\ldots,T$$

$$-pW_t + O_t \leq 0 \quad , \quad t=1,\ldots,T$$

$$W_t, O_t, H_t, F_t \geq 0 \quad , \quad t=1,\ldots,T$$

Newson suggested a heuristic iterative process that relates two models sequentially until a terminal criterion is met.

(C) Advantages and Disadvantages of Lot Size Models

The primary advantage of these models is that they incorporate the scheduling issues associated with lot size indivisabilities in the capacity planning decisions. Moreover, the linear programming approximations are computationally feasible and efficient.

The greatest limitation these models have is imposed by the nature of the problem they attack, which forces a great level of detailed information to be processed. An alternative approach to coordinate the aggregate capacity planning and detailed scheduling decisions is represented by the construction of hierarchical planning systems, to be discussed later on.

General Cost Models

The linear, quadratic and lot size models we have analyzed, although fairly appropriate for a great number of applications, impose several restrictions on the nature of the cost functions to be used. Some authors have argued that realistic industrial situations tend to exhibit cost functions which are nonlinear and discontinuous and, therefore, cannot be treated by any of the methods outlined previously. Buffa and Taubert [9] report the following factors as mainly responsible for this cost behavior: supply and demand interactions, manufacturing or purchasing economics of scale, learning curve effects, quantum jumps in costs in costs with addition of a new shift, technological and productivity changes, labor slowdowns, etc.

Several aggregate capacity planning methods have been suggested which attempt to be more responsive to the complexities introduced by the specific environment in which these decisions have to be made. Generally, these more realistic approaches do not guarantee that an optimum solution will be found and can be roughly classified according to the following categories:

- Heuristic decision rules, which attempt to bring in the decision maker's intuition of the problem under consideration by incorporating "rules of thumb" that can contribute to the solution of the problem;
- Search decision rules, which consist in the application of hill climbing techniques to the response surface defined by a nonlinear cost function and the problem constraints; and
- Simulation decision rules, which represent the problem under consideration by a set of programmed instructions. The decision maker is able to test various approaches in an iterative fashion, where the outcome of each run suggests what the subsequent run might be. Simulation is particularly suitable to treat the uncertainties that can be present in a decision.

We will now review the major contributions that have been proposed in each of these categories.

(A) Heuristic Decision Rules

Perhaps the most important attempt to incorporate management behavior in a systematic fashion to the aggregate capacity planning problem is Bowman's management coefficient approach [6]. Bowman suggested that managers tend to determine production rates, inventory levels and work force levels in a way which is responsive to the relevant costs that affect those decisions. However, they tend to overreact to the daily pressures of their work, occasionally creating expensive and erratic decisions which are important to prevent by maintaining them close to their average pattern of past behavior. Moreover, since most cost functions exhibit a flat shape around the optimum, small deviations from the optimum are not going to generate heavy penalties.

From this, Bowman concluded that a decision rule with mean coefficients estimated from management's past performance should produce better than actual results, and better than those results generated from analytical studies.

The actual structure of the decision rule to use can be suggested from analytical considerations, like the linear decision rules obtained from quadratic cost functions, by intuitive reasoning, or by a combination of both. Bowman suggested the following example of a production scheduling rule:

$$P_t = \sum_{i=1}^{t+T} a_i S_i + x(P_{t-1} - S_t) + y\,(I_N - I_{t-1})$$

where

P_t = production scheduled in period t

S_t = sales forecast in period t

x_y = smoothing constants $(0 \leq x \leq 1), (0 \leq \quad \leq 1)$

I_N = "normal" inventory

I_{t-1} = ending inventory at period t-1

a_i = weighting coefficient for sales forecast S_i,

$a_t > a_{t+1} > \ldots > a_{t+T}$

T = planning horizon

The numerical values of the coefficients a_i, x, and y are obtained not by analytical methods (like in the HMMS models) or by simulation techniques, but by performing regression analysis on past management behavior.

Bowman reported encouraging results by comparing the performance of his approach against linear decision rules and actual past costs in four industries.

(B) Search Decision Rules

Jones [33] combined a heuristic approach, to define the nature of the decision rules, and a search approach, to compute the coefficients of the decision rules, in developing a method for aggregate capacity planning that he called Parametric Production Planning. He started by postulating the existence of two linear decision rules, to address work force level and production level decisions, respectively.

The work force decision rule takes the form a a smoothing expression:

$$W_t = W_{t-1} + A(W_D - W_{t-1})$$

where:

W_{t+1} = current work force level

W_t = planned work force level for the upcoming period

W_D = desired work force level to meet upcoming demand forecast

A = coefficient determining the fraction of the difference in the planned and current work force to be realized.

The desired work force W_D is expressed as a weighted sum of the workforce required to meet future sales during the planning horizon, T.

$$W_D = \sum_{i=1}^{T} b_i \, K(S_{t+i-1})$$

where

b_i = weighting coefficient for sales forecast S_{t+i-1}

$K(S_t)$ = number of workers required to produce S_t units at minimum cost.

After experimenting with several weighting functions, Jones suggested the following expression to determine values of the S_i coefficients:

$$b_i = \frac{B^i}{\sum_{i=1}^{T} B^i}$$

where

B = coefficient between 0 and 1 that determines the relative weight to be given to future forecasts.

Note that all the b_i coefficients, i=1,...,T, are expressed as a function of a single parameter B.

Moreover, Jones included a term to prevent inconsistencies in inventory depletion or buildup. Jones suggested the following corrective term to be added to the work force decision rule:

$$b_i \, K(I_t^* - I_{t-1})$$

where

I_t^* = optimal inventory level at the end of the upcoming period (to be computed externally to the model).

The resulting work force decision rule becomes:

$$W_t = W_{t-1} + A \sum_{i=1}^{T} \left[b_i K(S_{t+i-1}) - W_{t-1} + b_1 \, K(I_t^* - I_{t-1}) \right]$$

The production decision rule is similar to the work force rules, except that the production rates are expressed in production units rather than in number of workers:

$$P_t = K^{-1}(W_t) + C \sum_{i=1}^{T} \left[d_i \, S_{t+1-1} - K^{-1}(W_t) + d_1 (I_t^* - I_{t-1}) \right]$$

where

$K^{-1}(W_t)$ = number of units that can be produced by W_t workers at minimum cost

C = coefficient between 0 and 1 indicating the fraction of the desired production increase or decrease to be achieved

d_i = weighting coefficient for sales forecast S_{t+i-1}

The d_i coefficients are defined by an expression similar to the one used for the b_i coefficient, i.e.,

$$d_i = \frac{D^i}{\sum_{i=1}^{T} D^i}$$

The numerical values of the four coefficients A, B, C and D are obtained by applying search techniques over a five dimensional space determined by the firm's profitability and the four parameters. The profitability is determined by taken into consideration the general cost structure relevant to the production rate and work force decisions.

There are a large number of different search techniques available for optimization purposes. An extensive coverage of these techniques have been reported by Wilde [58]. Among those techniques the one that seems most promising is the Direct Search procedures developed by Hooke and Jeeves [30]. Jones suggested that the response surface determined by the four coefficients and the associated profitability measure is unimodel, shallow and smooth, which are highly desirable attributes for search techniques to be applied. Jones reported some encouraging results after testing the performance of his approach.

Another important application of search to aggregate capacity planning was developed by Taubert [52]. There are some basic differences between Taubert and Jones' approaches. Taubert searches on the values of production rates, work force and inventory levels during each time period, while Jones searches only on the values of four coefficients (A,B,C, and D). The dimensionality of Taubert's search depends, therefore, on the number of time periods contained in the planning horizon, which creates more computational difficulties.

Taubert suggests also the possibility of combining search with branch and bound procedures by partitioning the set of feasible solutions, using the branch and bound methods, into simpler aggregate scheduling bounding problems that can be solved by applying search techniques. This approach, which would have some attractive potentials, has not yet been tested. Another simple application of search to aggregate capacity planning was done by Goodman [21].

(C) Simulation Decision Rules

For a long time simulation has been recognized as an important modelling tool to deal with situations where analytical models either become computationally infeasible or provide a too simplified representation of a real world problem.(1) Vergin [54] developed a general purpose simulator that is able to capture some of the special conditions that are present in practical scheduling problems that, by necessity, have been ignored in the analytical approaches to the capacity planning problem. The simulation can be adjusted to incorporate special conditions of a particular firm.

The simulation process starts with an initial schedule, which can be suggested by experience or can represent the current conditions of the firm. An objective function, which has no restrictions in terms of its structure, is used to evaluate the performance of each schedule. A change is introduced in employment levels, overtime, inventories, subcontracting, etc. until a local minimum is achieved.

Vergin conducted a study on three manufacturing firms affected by strong seasonalities and reported a much better performance of simulation schedules against both operating schedules and linear decision rules schedules.

(D) Advantages and Disadvantages of General Cost Models

One of the greatest advantages of the general cost models we have surveyed is the added realism they are capable of introducing to reflect more accurately the production planning environment, including uncertainties

(1) For good references on simulation see Emshoff and Sisson [16], and Naylor et.el. [42].

and special cost structure and constraints. In addition they are more closely associated with the actual decision process, which makes them more acceptable by managers and easier to explain and justify.

However, these advantages have a price. Usually the models are expensive to develop and to run, and the computational procedures used to solve them seldomly guarantee overall optimization. Some of the models require a high degree of aggregation, which creates problems of implementation when decisions need to be disaggregated at the lower levels. Moreover, general cost models do not lend themselves to handle a large number of interactive constraints, which can be easily managed by linear programming methods.

REFERENCES

1. Anshen, M., C.C. Holt, F. Modigliani, J.F. Muth, and H.A. Simon, "Mathematics for Production Scheduling," Harvard Business Review, March-April, 1958.

2. Anthony, R.N., Planning and Control Systems: A Framework for Analysis, Harvard University Graduate School of Business Administration, Boston, 1965.

3. Balinski, M.L., "Fixed Cost Transportation Problem," Naval Research Register Quarterly, Vol. 8, January, 1961.

4. Bergstrom, G.L., and B.E. Smith, "Multi-Item Production Planning - An Extension of the HMMS Rules," Management Science, Vol. 16, No. 10, June, 1970.

5. Bomberger, E.E., "A Dynamic Programming Approach to a Lot Size Scheduling Problem," Management Science, Vol. 12, No. 11, July, 1966.

6. Bowman, E.H., "Consistency and Optimality in Managerial Decision Making," Management Science, January, 1963.

7. _____, "Production Scheduling by the Transportation Method of Linear Programming," Operations Research, Vol. 4, No. 1, February, 1956.

8. Brown, R.G., Decision Rules for Inventory Management, Holt, Rinehart and Winston, 1967.

9. Buffa, E.S., and W.H. Taubert, Production-Inventory Systems: Planning and Control, Richard D. Irwin, Inc., 1972.

10. Chang, R.H., and C.M. Jones, "Production and Workforce Scheduling Extensions," AIIE Transactions, Vol. 2, No. 4, December, 1970.

11. Cooper, L. and C. Drebes, "An Approximate Solution Method for the Fixed Change Problem," Naval Research Logistics Quarterly, Vol. 14, No. 1, March, 1967.

12. Dantzig, G.B., and R.M. Van Slyke, "Generalized Upper Boundary Techniques," Journal of Computer System Science, Vol. 1, 1967.

13. Denzler, D.R., "An Approximate Algorithm for the Fixed Change Problem," NRLQ, Vol. 16, No. 3, September, 1969.

14. Dzielinski, B.P., C.T. Baker, and A.S. Manne, "Simulation Tests of Lot Size Programming," Management Science, Vol. 9, No. 2, January, 1963.

15. Dzielinski, B.P., and R.E. Gomory, "Optimal Programming of Lot Sizes, Inventory and Labor Allocations," Management Science, Vol. 11, No. 9, July, 1965.

16. Emshoff, J.R., and R.L. Sisson, Computer Simulation Models, Macmillan, 1970

17. Eppen, G.D., and F.J. Gould, "A Lagrangian Application to Production Models," Operations Research, Vol. 16, No. 4, July-August, 1968.

18. Eppen, G.D., F.J. Gould, and B.P. Pashigian, "Extensions of the Planning Horizon Theorem in the Dynamic Lot Size Model," Management Science, Vol. 15, No. 5, January, 1969.

19. Everett, H., "Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources," Operations Research, Vol. 11, No. 3, May-June, 1963.

20. Florian, M., and M. Klein, "Deterministic Production Planning with Concave Costs and Capacity Constraints," Management Science, Vol. 18, No. 1, September, 1971.

21. Goodman, D.A., "A New Approach to Scheduling Aggregate Production and Work Force," AIIE Transactions, Vol. 5, No. 2, June, 1973.

22. Gorenstein, S., "Planning Tire Production," Management Science, Vol. 17, No. 2, October, 1970.

23. Gray, P., "Exact Solution of the Fixed-Change Transportation Problem," Operations Research, Vol. 19, October, 1971.

24. Green, P.,"Heuristic Coupling of Aggregate and Detailed Models in Factory Scheduling," MIT, unpublished Ph.D. thesis, 1971.

25. Groff, G.K., and J.F. Muth, Operations Management: Analysis for Decisions, Richard D. Irwin, Inc., 1972.

26. Hanssmann, F. and S.W. Hess, "A Linear Programming Approach to Production and Employment Scheduling," Management Technology, No. 1, January, 1960.

27. Hax, A.C., and H.C. Meal, "Hierarchical Integration of Production Planning and Scheduling," Massachusetts Institute of Technology, Alfred P. Sloan School of Management, Working Paper 656-73, May, 1973. Forthcoming in Management Science.

28. Hodgson, T.J., "Addendum to Standard and Gupta's Note on Lot Size Scheduling," Management Science, Vol. 16, No. 7, March, 1970.

29. Holt, C.C., F. Modigliani, J.F. Muth, and H.A. Simon, Planning Production Inventories and Work Force, Prentice-Hall, Inc., 1960.

30. Hooke, R., and T.A. Jeeves, "Direct Search Solution of Numerical and Statistical Problems," Journal of the Association for Computing Machinery, Vol. 8, April, 1961.

31. Jagannathan, R., and M.M. Rao, "Class of Deterministic Production Planning Problems," Management Science, Vol. 19, No. 11, July, 1973.

32. Jones, A.P., and R.M. Soland, "An Approximative Algorithm for the Fixed-Change Transportation Problem," Naval Research Logistics Quarterly, Vol. 9, No. 1.

33. Jones, C.H., "Parametric Production Planning," Management Science, Vol. 13, No. 11, July, 1967.

34. Kolenda, J.F., "A Comparison of Two Aggregate Planning Models," unpublished master's thesis, Wharton School of Finance and Commerce, 1970.

35. Krajewski, L.J., V.A. Mabert, and H.E. Thompson, "Quadratic Inventory Cost Approximations and the Aggregation of Individual Products," Management Science, Vol. 19, No. 11, July, 1973.

36. Kriebel, C.H., "Coefficient Estimation in Quadratic Programming Models," Management Science, Vol. 13, No. 8, April, 1967.

37. Lasdon, L.S., and R.C. Terjung, "An Efficient Algorithm for Multi-Item Scheduling," Operations Research, Vol. 19, No. 4, July-August, 1971.

38. Magee, J.F., and D.M. Boodman, Production Planning and Inventory Control, McGraw-Hill, 1967.

39. Manne, A.S., "Programming of Economic Lot Sizes," Management Science, Vol. 4, No. 2, January, 1958.

40. McGarrah, R.E., Production and Logistics Management, Wiley, 1963.

41. Murty, K.G., "Solving the Fixed Change Problem by Routing Extreme Points," Operations Research, Vol. 16, No. 2, March-April, 1968.

42. Naylor, T., et al, "Computer Simulation Techniques," Wiley, 1966.

43. Newson, E.F.P., "Lot Size Scheduling to Finite Capacity," unpublished doctoral thesis, Sloan School of Management, MIT, 1971.

44. O'Malley, R.L., S.E. Elmaghraby, and J.W. Jeske, "An Operational System for Smoothing Batch-Type Production," Management Science, Vol. 12, No. 10, June, 1966.

45. Orrbeck, M.G., D.R. Schuette and H.E. Thompson, "The Effect of Worker Productivity on Production Smoothing," Management Science, Vol. 14, No. 6, February, 1968.

46. Rousseau, J.M., "A Cutting Plane Method for the Fixed Cost Problem," unpublished doctoral thesis, Sloan School of Management, MIT, August, 1973.

47. Shwimer , J., "Interaction Between Aggregate and Detailed Scheduling in a Job Shop," MIT, unpublished Ph.D. thesis, 1972.

48. Silver, E.A., "A Tutorial on Production Smoothing and Work Force Balancing," Operations Research, Vol. 15, No. 6, November-December, 1967.

49. Stankard, M.F., and S.K. Gupta, "A Note on Bomberger's Approach tc Lot Size Scheduling: Heuristic Proposed," Management Science, Vol. 15, No. 7, March, 1969.

50. Steinberg, D.I., "The Fixed Change Problem," Naval Research Logistics Quarterly, Vol. 17, No. 2, June, 1970.

51. Sykpens H.A., "Planning for Optimal Plant Capacity," unpublished master's thesis, Sloan School of Management, MIT, 1967.

52. Taubert, W.H., "A Search Decision Rule for the Aggregate Scheduling Pattern," Management Science, Vol. 14, No. 6, February, 1968.

53. Van de Panne, C., and P. Bosje, "Sensitivity Analysis of Lost Coefficient Estimates: The Case of Linear Decision Rules for Employment and Production," Management Science, Vol. 9, No. 1, October, 1962.

54. Vergin, R.C., "Production Scheduling Under Seasonal Demand," Journal of Industrial Engineering, Vol. 7, May, 1966.

55. Von Lanzenaur, C.H., "Production and Employment Scheduling in Multistage Production Systems," Naval Research Logistics Quarterly, Vol. 17, No. 2, July, 1970.

56. Wagner, H.M., "A Postcript to Dynamic Problems in the Theory of the Firm," Naval Research Logistics Quarterly, Vol. 7, No. 1, March, 1960.

57. Wagner, H.M. and T.M. Whitin, "A Dynamic Version of the Economic Lot Size Model," Management Science, Vol. 5, 1958.

58. Wilde, D.J., "Optimum Seeking Methods," Prentice-Hall, 1964.

59. Winters, P.R., "Constrained Inventory Rules for Production Smoothing," Management Science, Vol. 8, No. 4, July, 1962.

60. Zangwill, W.I., "A Deterministic Multi-Period Production Scheduling Model with Backlogging," Management Science, Vol. 13, No. 1, September, 1966.

61. Zangwill, W.I., "A Backlogging Model and Multi-Echelon Model of a Dynamic Economics Lot Size Production System - a Network Approach," Management Science, Vol. 15, No. 9, May, 1969.

62. Zoller, K., "Optimal Disaggregation of Aggregate Production Plans," Management Science, Vol. 17, No. 8, April, 1971.