Manufacturing Planning and Control

Stephen C. Graves
Massachusetts Institute of Technology
November 1999

Manufacturing planning and control entails the acquisition and allocation of limited resources to production activities so as to satisfy customer demand over a specified time horizon. As such, planning and control problems are inherently optimization problems, where the objective is to develop a plan that meets demand at minimum cost or that fills the demand that maximizes profit. The underlying optimization problem will vary due to differences in the manufacturing and market context. This chapter provides a framework for discrete-parts manufacturing planning and control and provides an overview of applicable model formulations.

Manufacturing Planning and Control

Stephen C. Graves
Massachusetts Institute of Technology
November 1999

Manufacturing planning and control address decisions on the acquisition, utilization and allocation of production resources to satisfy customer requirements in the most efficient and effective way. Typical decisions include work force level, production lot sizes, assignment of overtime and sequencing of production runs. Optimization models are widely applicable for providing decision support in this context.

In this article we focus on optimization models for production planning for discrete-parts, batch manufacturing environments. We do not cover detailed scheduling or sequencing models (e. g., Graves, 1981), nor do we address production planning for continuous processes (e. g., Shapiro, 1993). We consider only discrete-time models, and do not include continuous-time models such as developed by Hackman and Leachman (1989).

Our intent is to provide an overview of applicable optimization models; we present the most generic formulations and briefly describe how these models are solved. There is an enormous range of problem contexts and model formulations, as well as solution methods. We make no effort to be exhaustive in the treatment herein. Rather, we have made choices of what to include based on personal judgment and preferences.

We have organized the article into four major sections. In the first section we present a framework for the decisions, issues and tradeoffs involved in implementing an optimization model for discrete-part production planning. The remaining three sections present and discuss three distinct types of models. In the second section we discuss linear programming models for production planning, in which we have linear costs. This category is of great practical interest, as many important problem features can be captured with these models and powerful solution methods for linear programs are readily available. In the third section, we present a production-planning model for a single aggregate product with quadratic costs; this model is of historical significance as it represents one of the earliest applications of optimization to manufacturing planning. In the final section we introduce the multi-item capacitated lot-size problem, which is modeled as a mixed integer linear program. This is an important model as it introduces economies of scale in production, due to the presence of production setups.

Framework

There are a variety of considerations that go into the development and implementation of an optimization model for manufacturing planning and control. In this section we highlight and comment upon a number of key issues and questions that should be addressed. Excellent general references on production planning are Thomas and McClain (1993), Shapiro (1993) and Silver et al. (1998).

Any planning problem starts with a specification of customer *demand* that is to be met by the production plan. In most contexts, future demand is at best only partially known, and often is not known at all. Consequently, one relies on a forecast for the future demand. To the extent that any forecast is inevitably inaccurate, one must decide how to account for or react to this demand uncertainty. The optimization models described in this article treat demand as being known; as such they must be periodically revised and rerun to account for forecast updates.

A key choice is what *planning decisions* to include in the model. By definition, production-planning models include decisions on production and inventory quantities. But in addition, there might be resource acquisition and allocation decision, such as adding to the work force and upgrading the training of the current work force.

In many planning contexts, an important construct is to set a *planning hierarchy*. Namely, one structures the planning process in a hierarchical way by ordering the decisions according to their relative importance. Hax and Meal (1975) introduced the notion of hierarchical production planning and provide a specific framework for this, whereby there is an optimization model with each level of the hierarchy. Each optimization model imposes a constraint on the model at the next level of the hierarchy. Bitran and Tirupati (1993) provide a comprehensive survey of hierarchical planning methods and models.

The identification of the *relevant costs* is also an important issue. For production planning, one typically needs to determine the variable production costs, including setup-related costs, inventory holding costs, and any relevant resource acquisition costs. There might also be costs associated with imperfect customer service, such as when demand is backordered.

A planning problem exists because there are limited *production resources* that cannot be stored from period to period. Choices must be made as to which resources to include and how to model their capacity and behavior, and their costs. Also, there may be uncertainty associated with the production function, such as uncertain yields or lead times. One might only include the most critical or limiting resource in the planning problem, e. g., a bottleneck. Alternatively, when there is not a dominant resource, then one must model the resources that could limit production. We describe in this article two types of production functions. The first assumes a linear relationship between the production quantity and the resource consumption. The second assumes that there is a required fixed charge or setup to initiate production and then a linear relationship between the production quantity and resource usage.

Related to these choices is the selection of the *time period* and *planning horizon*. The planning literature distinguishes between “big bucket” and “small bucket” time periods. A time period is a big bucket if multiple items are typically produced within a time period; a small bucket is such that at most one item would be produced in the time period. For big bucket models, one has to worry about how to schedule or sequence the production runs assigned to any time period. The choice of planning horizon is dictated by the lead times to enact production and resource-related decisions, as well as the quality of knowledge about future demand.

Planning is typically done in a *rolling horizon* fashion. A plan is created for the planning horizon, but only the decisions in the first few periods are implemented before a revised plan is issued. Indeed, as noted above, the plan must be periodically revised due to the uncertainties in the demand forecasts and production. For instance a firm might plan for the next 26 weeks, but then revise this once a month to incorporate new information on demand and production.

Production planning is usually done at an *aggregate level*, for both products and resources. Distinct but similar products are combined into aggregate product families that can be planned together so as to reduce planning complexity. Similarly production resources, such as distinct machines or labor pools, are aggregated into an aggregate machine or labor resource. Care is required when specifying these aggregates to assure that the resulting aggregate plan can be reasonably disaggregated into feasible production schedules.

Finally for complex products, one must decide the level and extent of the *product structure* to include in the planning process. For instance, in some contexts it is sufficient to just plan the production of end items; the production plan for components and subassemblies is subservient to the master production schedule for end items. In other contexts, planning just the end items is sub-optimal, as there are critical resource constraints applicable to multiple levels of the product structure. In this instance, a multi-stage planning model allows for the simultaneous planning of end items and components or subassemblies. Of course, this produces a much larger model.

Production Planning: Linear Programming Models

In this section we develop and state the most basic optimization model for production planning for the following context:

- multiple items with independent demand
- multiple shared resources
- big-bucket time periods
- linear costs.

We define the following notation

decision variables

p_{it} production of item i during time period t
q_{it} inventory of item i at end of time period t

parameters

T, I, K number of time periods, items, resources, respectively

a_{ik} amount of resource k required per unit of production of item i
b_{kt} amount of resource k available in period t
d_{it} demand for item i in period t

cp_{it} unit variable cost of production for item i in time period t
cq_{it} unit inventory holding cost for item i in time period t

We now formulate the linear program P1:

$$\text{P1:} \quad \text{Min} \sum_{\text{t=1}}^{\text{T}} \sum_{\text{i=1}}^{\text{I}} cp_{it} \, p_{it} + cq_{it} q_{it} \qquad (1)$$

$$s.t.$$

$$q_{i,t-1} + p_{it} - q_{it} = d_{it} \quad \forall \, i, \; t \qquad (2)$$

$$\sum_{\text{i=1}}^{\text{I}} a_{ik} \, p_{it} \leq b_{kt} \quad \forall \, k, \; t \qquad (3)$$

$$p_{it}, q_{it} \geq 0 \quad \forall \, i, \; t$$

The objective function (1) minimizes the variable production costs plus the inventory holding costs for all items over the planning horizon of T periods.

Equation (2) is a set of inventory balance constraints that equate the supply of an item in a period with its demand or usage. In any period, the supply for an item is the inventory from the prior period $q_{i,t-1}$, plus the production in the period p_{it}. This supply can be used to meet demand in the period d_{it}, or held in inventory as q_{it}. As we require the inventory to be non-negative, these constraints assure that demand is satisfied for each item in each period. We are given as input the initial inventory for each item, namely q_{i0}.

Equation (3) is a set of resource constraints. Production in each period is limited by the availability of a set of shared resources, where production of one unit of item i requires a_{ik} units of resource k, for k = 1, 2, ... K. Typical resources are various types of labor, process and material handling equipment, and transportation modes.

The number of decision variables is 2IT, and the number of constraints is IT + KT. For any realistic problem size, we can solve P1 by any good linear-programming algorithm, such as the simplex method.

We briefly describe next a number of important extensions to this basic model. We introduce these as if they were independent; however, we note that many contexts require a combination of these extensions.

Demand Planning: Lost Sales

For some problems we have the option of not meeting all demand in each time period. Indeed, there might not be sufficient resources to meet all demand. In effect, the demand parameters represent the demand potential, and the optimization problem is to decide what demand to meet and how. We assume that demand that cannot be met in a period is lost, thus reducing revenue. In addition, a firm might incur a loss of customer goodwill that would manifest itself in terms of reduced future sales. This lost sales cost is very difficult to quantify as it represents the future unknown impact from poor service today.

We pose a new planning problem to maximize revenues net of the production, inventory and lost sales costs. We introduce additional notation and then state the model:

decision variables

u_{it}	unmet demand of item i during time period t

parameters

r_{it}	unit revenue for item i in period t
cu_{it}	unit cost of not meeting demand for item i in time period t

$$\text{P2:} \quad \text{Max} \sum_{t=1}^{T} \sum_{i=1}^{I} \left[r_{it}(d_{it} - u_{it}) - cp_{it} p_{it} - cq_{it} q_{it} - cu_{it} u_{it} \right]$$

$$s.t. \ (3)$$

$$q_{i,t-1} + p_{it} - q_{it} + u_{it} = d_{it} \quad \forall\, i,\ t$$

$$p_{it}, q_{it}, u_{it} \geq 0 \quad \forall\, i,\ t$$

The objective function has been modified to include revenue as well as the cost of lost sales. The potential revenue, $\Sigma\ \Sigma\ r_{it}\ d_{it}$, is a constant and could be dropped in the objective function. In this case, we can restate the problem as a cost minimization problem, where the cost of lost sales includes the lost revenue.

Also, in P2, the inventory balance constraint has been modified to permit the option of not meeting demand; thus demand in a period can be met from production or inventory, or not satisfied at all. The resource constraint (3) remains unchanged.

Demand Planning: Backorders

A related problem variation is when it is possible to reschedule or backorder demand. That is, we might defer current demand until a later period, when it can be served from production. Of course there is a cost for doing this, which we term the backorder cost. Like the lost sales cost, the backorder cost includes hard-to-quantify costs due to loss of customer goodwill, as well as reduced revenue and additional processing or expediting costs due to the deferral of the demand fulfillment. We assume that this cost is linear in the number of backorders in each period.

We define additional notation and then state the model.

decision variables

v_{it} backorder level for item i at end of time period t

parameters

cv_{it} unit cost of backorder for item i in time period t

$$\text{P3:}\quad \text{Min} \sum_{\mathrm{t=1}}^{\mathrm{T}} \sum_{\mathrm{i=1}}^{\mathrm{I}} \left[cp_{it}\, p_{it} + cq_{it} q_{it} + cv_{it} v_{it} \right]$$

$$s.t.\ (3)$$

$$q_{i,t-1} - v_{i,t-1} + p_{it} - q_{it} + v_{it} = d_{it} \quad \forall\ i,t$$

$$p_{it}, q_{it}, v_{it} \geq 0 \quad \forall\ i,t$$

In comparison with P1, we now include a backorder cost on the objective function for P3. The inventory balance equation is modified to account for the backorders, which in effect behave like negative inventory. We typically would add a terminal constraint on the backorders at the end of the planning horizon; for instance, we might require $v_{iT} = 0$, so that over the T-period planning horizon all demand is eventually met by the production plan. Any initial backorders, namely v_{i0}, can be dropped by adding them to the first-period demand; that is, we restate the demand as $d_{i1} := d_{i1} + v_{i0}$, and then drop v_{i0} from the formulation.

In this formulation, when demand is backordered, the cost of this event is linear in the size and duration of the backorder. That is, if it takes n time periods to fill the backorder, the cost is proportional to n. In contrast, in some cases, the backorder cost might not depend on the duration but only on the occurrence and size of the backorder. We can modify this formulation for this case by defining a variable to represent new backorders in period t, given by max [0, v_{it} - $v_{i,t-1}$]; then we would apply the backorder cost to this variable in the objective function. This modification assumes that we fill backorders in a last-in, last-out fashion, as there is no incentive to do otherwise for this cost assumption.

Piecewise Linear Production Cost Functions

In P1 the relevant production cost is linear in the production quantity. In many contexts the actual cost function is non-linear. In this section, we consider a convex cost function, and assume that it is well modeled as a piecewise linear function. We model concave cost functions that exhibit economies of scale in alter section.

Let Cost(p_{it}) denote the cost function for item i in period t; we present the case where this cost function is the same in each period, and introduce the following notation:

decision variables

p_{ist} production of item i during time period t, that falls in cost segment s

parameters

S number of segments in cost function

cp_{ist} unit variable cost of production for item i in time period t in cost segment s

P_{is} upper bound on cost segment s for item i

Thus, we assume that Cost(p_{it}) is given by:

$$Cost(p_{it}) = \sum_{s=1}^{S} cp_{is} p_{ist}$$

where

$$p_{it} = \sum_{s=1}^{S} p_{ist}$$

$$0 \le p_{ist} \le P_{is} \quad \forall s$$

In order for Cost(p_{it}) to be convex, we require that the unit variable costs be strictly increasing from one segment to the next:

$0 < cp_{i1t} < cp_{i2t} < ... < cp_{iSt}$

This cost function applies when there are multiple options or sources for production, and these options can be ranked by their variable costs. A common example is when one models regular time and overtime production. We have two cost segments (S=2) where the first segment corresponds to production during regular time, and the second is overtime production. The variable production cost is usually more during overtime, as workers earn a rate premium. Another example is when the firm works multiple shifts and the variable costs differ between these shifts. A third example is when there are subcontracting or outsourcing options; there are multiple costs segments, one to represent in-house production and the others to represent the outsourcing options ranked by cost.

We model the planning problem with convex, piecewise linear production cost functions by replacing the production cost in P1 with the above formulation for Cost(p_{it}):

$$\text{P4:}\quad \text{Min} \sum_{t=1}^{T}\sum_{i=1}^{I}(cq_{it}q_{it} + \sum_{s=1}^{S} cp_{is}p_{ist})$$

$$s.t.\ (2)$$

$$\sum_{i=1}^{I} a_{iks}p_{ist} \le b_{kt} \quad \forall\, k,\ t$$

$$p_{it} - \sum_{s=1}^{S} p_{ist} = 0 \quad \forall i, t$$

$$0 \le p_{ist} \le P_{is} \qquad \forall i, s$$

$$p_{it}, p_{ist}, q_{it} \ge 0 \qquad \forall\ i, s, t$$

In P4 we have modified the resource constraints (3) to accommodate the possibility that the usage of the shared resources depends on the production quantity by source or cost segment, i. e., p_{ist}, rather than just on p_{it}. In this case, a_{iks} denotes the amount of resource k required per unit of item i produced at source or cost segment s. This form permits great flexibility in modeling production costs as well as resource constraints.

In one of the first papers on production planning, Bowman (1956) formulates this problem as a transportation problem, when there are multiple time periods and multiple production options, but only one item and one resource type.

Resource Planning

Up to now we have assumed that the resource levels are fixed and given. In some cases, an important element of the planning problem is to decide how to adjust the resource levels over the planning horizon. For instance, one might be able to change the work force level, by means of hiring and firing decisions. Hansmann and Hess (1960) provide an early example of this type of model.

Suppose for ease of notation that we have just one type of resource, namely the work force. We introduce additional notation and then state the model:

decision variables

w_t work force level in time period t

h_t change to work force level by hiring in time period t
f_t change to work force level by firing in time period t

parameters
a_i amount of work force (labor) required per unit of production of item i
cw_t variable unit cost of work force in time period t
ch_t variable hiring cost in time period t
cf_t variable firing cost in time period t

$$\text{P5:} \quad \text{Min} \sum_{\text{t=1}}^{\text{T}} \left[cw_t w_t + ch_t h_t + cf_t f_t \right] + \sum_{\text{t=1}}^{\text{T}} \sum_{\text{i=1}}^{\text{I}} \left[cp_{it} p_{it} + cq_{it} q_{it} \right]$$

$$s.t. \ (2)$$

$$\sum_{\text{i=1}}^{\text{I}} a_i p_{it} - w_t \leq 0 \quad \forall t$$

$$w_{t-1} + h_t - f_t - w_t = 0 \quad \forall \ t$$

$$p_{it}, q_{it}, w_t, h_t, f_t \geq 0 \quad \forall \ i, \ t$$

We add the variable cost for the work force to the objective function, along with costs for hiring and firing workers. The hiring cost includes costs for finding and attracting applicants as well as training costs. The firing cost includes costs of outplacement and retraining of displaced workers, as well as severance costs; there might also be a cost of lower productivity due to lower work-force morale, when firings or layoffs occur.

The inventory balance constraint (2) remains the same as for P1, and we restate the resource constraint, reflecting the work force as a decision variable and as the sole resource. We then add a new set of balance constraints for planning the work force: the work force in period t is that from the prior period plus new hires minus the number fired.

We have stated P5 for a single resource, representing the work force. The model extends immediately to include other resources that might be managed in a similar fashion over the planning horizon. In addition, there might be other considerations to model such as time lags when adjusting a resource level. There might be limits on how quickly new workers can be added due to training requirements. If there were limited training resources, then this imposes a constraint on h_t. Alternatively, new hires might be less productive until they have acquired some experience. In this case, we modify the formulation to model different categories of workers, depending on their tenure and experience level.

Another common variation of this model is when there are two labor classes, say, permanent employees and temporary employees. These classes differ in terms of their cost coefficients, and possibly their efficiency factors. Permanent employees have higher hiring and firing cost, as they receive more training and have more rights and protection from layoffs. But their variable production cost, normalized by their productivity, should be lower than that for temporary workers. The planning problem then entails the management and planning of both work classes over the planning horizon. For completeness, we revise P5 for two work force classes:

decision variables
w_{jt} work force level in time period t
h_{jt} change to work force level by hiring in time period t
f_{jt} change to work force level by firing in time period t
p_{ijt} production of item i during time period t, using labor class j

parameters
a_{ij} amount of labor required per unit of production of item i, using labor class j
cw_{jt} variable unit cost of labor class j in time period t
ch_{jt} variable hiring cost for labor class j in time period t
cf_{jt} variable firing cost for labor class j in time period t

$$\text{P6:}\quad \text{Min} \sum_{t=1}^{T}\sum_{j=1}^{2}\left[cw_{jt}w_{jt}+ch_{jt}h_{jt}+cf_{jt}f_{jt}\right]+\sum_{t=1}^{T}\sum_{i=1}^{I}\left[cp_{it}p_{it}+cq_{it}q_{it}\right]$$

$$s.t.\ (2)$$

$$p_{it}-\sum_{j=1}^{2}p_{ijt}=0 \quad \forall i,t$$

$$\sum_{i=1}^{I}a_{ij}p_{ijt}-w_{jt}\le 0 \quad \forall j,t$$

$$w_{j,t-1}+h_{jt}-f_{jt}-w_{jt}=0 \quad \forall j,t$$

$$p_{it},q_{it},w_{jt},h_{jt},f_{jt}\ge 0 \quad \forall\ i,j,t$$

In comparison to P5, we have decision variables for both labor classes, as well as for their hiring and firing decisions, in order to model the cost differences. We also introduce production decision variables, by labor class, to capture the differences in productivity.

Multiple Locations

In P1, there is a single supply location or production facility that serves demand for all items. Often there are multiple production facilities that are geographically dispersed and that supply multiple distribution channels. The planning problem is to determine the production, inventory and distribution plans for each facility to meet demand, which is now modeled by geographic region.

There are many ways to formulate this type of problem. We provide an example, and then comment on some variants. We define the notation and then state the model.

decision variables

p_{ijt} production of item i at facility j during time period t
q_{ijt} inventory of item i at facility j at end of time period t
x_{ijmt} shipment of item i from facility j to demand location m in time period t

parameters

T, I, K number of time periods, items, resources, respectively
J, M number of facility locations, demand locations, respectively

a_{ijk} amount of resource k required per unit of production of item i at facility j
b_{jkt} amount of resource k available at facility j in period t
d_{imt} demand for item i at location m in period t

cp_{ijt} unit variable cost of production for item i at facility j in time period t
cq_{ijt} unit inventory holding cost for item i at facility j in time period t
cx_{ijmt} unit transportation cost to ship item i from facility j to demand location m in time period t

$$\text{P7:}\quad \text{Min} \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{J}[cp_{ijt}p_{ijt} + cq_{ijt}q_{ijt} + \sum_{m=1}^{M} cx_{ijmt}x_{ijmt}]$$

$$s.t.$$

$$q_{ij,t-1} + p_{ijt} - q_{ijt} - \sum_{m=1}^{M} x_{ijmt} = 0 \quad \forall\, i, j, t$$

$$\sum_{j=1}^{J} x_{ijmt} = d_{imt} \quad \forall\, i, m, t$$

$$\sum_{i=1}^{I} a_{ijk}\, p_{ijt} \le b_{jkt} \quad \forall\, j, k, t$$

$$p_{ijt}, q_{ijt}, x_{ijmt} \ge 0 \quad \forall\, i, j, m, t$$

The decision variables for production and inventory are now specified by location, where inventory is held at the production locations. In addition, we have aggregated demand into a set of demand regions or locations, and introduce a new set of decision variables to denote transportation from the production facilities to the demand locations. The objective function captures production and inventory holding cost, which depends on the facility, plus transportation or distribution costs for moving the product from a facility to the demand location. The inventory balance constraints assure that the supply of an item at each facility is either held in inventory or shipped to a demand location to meet demand. The second set of constraints assures that the shipments satisfy the demand each period at each location. The resource constraints are structurally the same as in P1, but for multiple production locations.

A key variant of this model occurs when there are additional stocking locations, such as a network of warehouses or distribution centers. These stocking locations not only provide additional space to store inventory close to the demand locations, but also permit economies of scale in transportation from the production sites to these stocking locations. In this case, one defines shipments to and from each stocking location, and has inventory balance constraints for each stocking location. The shipment costs would capture any differences in transportation modes that might be employed.

The size of P7 creates a challenge for implementing and maintaining such a model. There are now IJT(2+M) decision variables and (IJT + IMT + JKT) constraints. A typical problem might have 20 – 100 aggregate item families, 5 - 10 facilities, 50 – 100 demand locations, 12 – 20 time periods, and 1 – 5 resource types. Thus, the model might have on the order of 100,000 to 1 million decision variables, and on the order of 100,000 constraints. As the problem is still a linear program, such problems are readily solved by commercial optimization packages. However, the real difficulty in such implementations is the development and maintenance of the parameters. There can be on the order of one million demand forecasts, 100,000 to 1 million cost coefficients, and 10,000 resource coefficients. The success of many applications often rests on whether this data can be obtained and kept accurate.

Dependent Demand Items

So far we have considered production plans for end items or finished goods, which see independent demand. But these end items are usually comprised of many fabricated components and subassemblies, for which there needs to be a production plan too. These items differ from the end items, in that their demand depends on the end-item production plan. Material requirements planning (MRP) systems are designed to characterize this dependent demand and to facilitate the planning for these dependent-demand items in a coordinated and systematic way (Vollman et al. 1992). Nevertheless, it is quite easy to incorporate dependent-demand items into the optimization models presented here.

W first need to define the “goes into” matrix A = $\{\alpha_{ij}\}$, where α_{ij} is the number of units of item i required per unit of production of item j. Then for the model P1 we replace the inventory balance constraints (2) with the following:

$$q_{i,t-1} + p_{it} - q_{it} - \sum_{\mathrm{j}=1}^{\mathrm{I}} a_{ij} p_{jt} = d_{it} \quad \forall\, i,\ t$$

In this general form, the demand for item i has two parts: exogenous demand given by d_{it}, and endogenous or induced demand given by $\Sigma\alpha_{ij}\, p_{jt}$. For end items we expect there to be no induced demand, whereas for components, there will typically be no or limited exogenous demand.

One variation to this model is when there are manufacturing lead times, whereby production of item j in time period t requires that component i be available at time t – L, where L is the lead time for producing j. The inventory balance constraints can be easily modified to accommodate this, given that the lead times are known and deterministic.

Billington et al. (1983) provide a comprehensive treatment of this model and problem, and present methods for reducing the size of the problem so as to facilitate its solution. In the literature, this problem is referred to as a multiple stage problem, where end items, subassemblies, and components might represent distinct stages in a manufacturing process.

Quadratic Cost Models and Linear Decision Rule

One of the earliest production-planning modeling efforts was that of Holt, Modigliani, Muth and Simon (1960), who developed a production-planning model for the Pittsburgh Paint Company. They assume a single aggregate product, and then define three decision variables:

p_t production of the aggregate item during time period t
q_t inventory of the aggregate item at end of time period t
w_t work force level in time period t

Holt, Modigliani, Muth and Simon (HMMS) assume that the cost function in each period has four components. The first component is the regular payroll costs that is a linear function of the work force level. The second component is the hiring and layoff costs, which were assumed to be a quadratic function in the change in work force from one period to the next.

The next cost component is for overtime and idle-time costs. HMMS assume there is an ideal production target that is a linear function of the work-force level. If production is greater than this target there is an overtime cost, while if production is less than this target there is an idle-time cost. Again, HMMS assume that the cost is quadratic about the variance between the actual production and the production target for the work-force level.

The final component is inventory and backorder costs. Similar to the overtime and idle-time costs, there is an inventory target each period, which is a linear function of the demand in the period. The inventory and backorder cost is a quadratic function of the deviation between the inventory and the inventory target.

The HMMS optimization is to minimize the sum of the expected costs over a fixed horizon, subject to an inventory balance constraint. The expectation is over the demand random variables, where we are given an unbiased forecast of demand over the planning horizon. The analysis of this optimization yields two key results.

First, the optimal solution can be characterized as a *linear decision rule*, whereby the aggregate production rate in each period is a linear function of the future demand forecasts, as well as the work force and inventory level in the prior period. There is a similar linear function for specifying the work force level in each period.

Second, the optimal decision rule is derived for the case of stochastic demand, but only depends on the mean of the demand random variables. That is, we only need to know (or assume) that the demand forecasts are unbiased in order to apply the linear decision rule.

Production Planning: Lot-Size Models

In this section we consider production-planning problems for which there are economies of scale associated with the production activity or function. The most common example occurs when there is a required setup to initiate the production of an item. For instance, to initiate the production of an item might require a change in tooling, or dies, or raw material; the setup might also require a change in the production control settings, as well as an initial run to assure that the production output meets quality specifications. There might be a setup cost, corresponding to labor costs for performing the setup, plus direct expenditures for materials and tools. There might also be resource requirements for the setup, usually referred to as the setup time. The production resource cannot produce until the setup is completed; thus the setup consumes production capacity, equal to its duration, namely the setup time.

Given the presence of setups, once an item is setup to produce, we may want to produce a large batch or lot size so as to cover demand over a number of future periods and hence defer the next time when the item will be setup and produced. Whereas producing large batches will reduce the setup costs, this also increases inventory as more demand is produced earlier in time. The lot-sizing problem, as described here, is to determine the relative frequency of setups so as to minimize the setup and inventory costs, within the resource and service constraints of the production-planning problem.

We start with the simplest model and then briefly discuss variants to it. We develop and state this model for the following context:

- multiple items with independent demand
- a single shared resource
- big-bucket time periods
- linear costs, except for setup costs.

For ease of presentation we assume a single resource; the extension to consider multiple resources is straightforward.

We define the following notation

decision variables

p_{it} production of item i during time period t

q_{it} inventory of item i at end of time period t

y_{it} binary decision variable to denote setup of item i in time period t

parameters

T, I number of time periods, items, respectively

a_{i1} amount of resource required per unit of production of item i

a_{i2} amount of resource required for setup of production of item i

b_t	amount of resource available in period t
d_{it}	demand for item i in period t
B	a large constant
cp_{it}	unit variable cost of production for item i in time period t
cq_{it}	unit inventory holding cost for item i in time period t
cy_{it}	setup cost for production for item i in time period t

We now formulate the mixed-integer linear program P8:

$$\text{P8:} \quad \text{Min} \sum_{t=1}^{T} \sum_{i=1}^{I} cp_{it} p_{it} + cq_{it} q_{it} + cy_{it} y_{it} \qquad (4)$$

$$s.t.$$

$$q_{i,t-1} + p_{it} - q_{it} = d_{it} \quad \forall i,t \qquad (5)$$

$$\sum_{i=1}^{I} a_{i1} p_{it} + a_{i2} y_{it} \le b_t \quad \forall t \qquad (6)$$

$$p_{it} \le B y_{it} \quad \forall i,t \qquad (7)$$

$$p_{it}, q_{it} \ge 0; y_{it} = 0,\ 1 \quad \forall\ i,\ t$$

The objective function (4) minimizes the sum of variable production costs, the inventory holding costs and the setup costs for all items over the planning horizon of T periods.

Equation (5) is the same as (2), the inventory balance constraints that equate the supply of an item in a period with its demand or usage.

The resource constraints (6) reflect the resource consumption both due to the production quantity for each item, and due to the setup. Production of one unit of item i consumes a_{i1} units of the shared resource, while the setup requires a_{i2} units.

The constraint set (7) is for the so-called forcing constraints. These constraints relate the production variables to the setup variables. For each item and time period, if there is no setup (y_{it}=0), then this constraint assures that there can be no production (p_{it}=0). Conversely, if there is production in a period (p_{it}>0), then there must also be a setup (y_{it}=1). In (7), B is any large positive constant that exceeds the maximum possible value for p_{it}; for instance, one might set B equal to the sum of all demand.

This problem is now a mixed-integer linear program, with IT binary decision variables. For modest size problems with, say, a few hundred binary decision variable, this problem can be reliably solved by commercial optimization packages. But specialized approaches are warranted for increasing problem size and complexity. We discuss one of these approaches next.

Dual-Based Solutions

In this section we develop a dual problem for P8. We identify a generalized linear program for solving this dual problem, which is equivalent to a convexification of P8. We can solve this problem by column generation to obtain a lower bound on P8; we also discuss how this solution can be used to identify near optimal solutions to P8.

The approach is based on the original work of Manne (1958) who suggested the generalized linear program given below. Dzielinski and Gomory (1965) extended the model of Manne to include resource planning decisions (i. e., hiring and firing labor) and applied the Dantzig-Wolfe decomposition method to introduce column generation. Lasdon and Terjung (1971) reformulated the linear program so as to provide a more efficient and effective column generation approach for solving the linear program; they also address a variant of P8, where there are small time buckets for planning production and multiple resources corresponding to scarce machines and dies.

To develop the dual problem, we first define a Lagrangean function L(π) by dualizing the resource constraints (6):

$$L(p) = \text{Min } -\sum_{t=1}^{T} b_t p_t + \sum_{t=1}^{T}\sum_{i=1}^{I} (cp_{it} + a_{i1} p_t) p_{it} + cq_{it} q_{it} + (cy_{it} + a_{i2} p_t) y_{it}$$
$$s.t. (5), (7)$$
$$p_{it}, q_{it} \geq 0; y_{it} = 0, 1 \quad \forall\, i, t$$

where $\pi = (\pi_1, \pi_2, \ldots \pi_T) \geq 0$ is a vector of dual variables.

We state the following observations:

- The Lagrangean separates by item, where we have a single-item dynamic lot-size problem with no capacity constraints for each item. This is the so-called Wagner-Whitin (1958) problem and can be solved by dynamic programming.
- If $q_{i0} = 0$, the extreme points to L(π), namely the single-item dynamic lot-size problem, have the property whereby $p_{it}\, q_{it-1}=0$. As a consequence the optimal production quantities are a sum of consecutive demands. That is, if $p_{it} > 0$, then $p_{it} = d_{it} + \ldots + d_{is}$ for t ≤ s ≤ T. We will refer to this as a “Wagner-Whitin schedule.”
- Without loss of generality, we will assume that $q_{i0} = 0$, and the above property applies to optimal solutions. If $q_{i0} > 0$, then we use this initial inventory to reduce the item’s demand. That is, we restate the demand as follows:
  $d_{it} := 0$ *for* $t = 1, 2, \ldots s-1$
  $d_{is} := d_{i1} + d_{i2} + \ldots + d_{i,s-1} - q_{i0}$
  where s such that $d_{i1} + d_{i2} + \ldots + d_{i,s-1}$ £ $q_{i0} < d_{i1} + d_{i2} + \ldots + d_{is}$.

We can use the Lagrangean function to define a dual problem to P8:

$$D8: \quad Max\ L(p)$$
$$s.t.\ p \geq 0$$

The dual solution need not and usually will not identify a primal feasible solution to P8. In such instances, a duality gap exists and the dual solution provides a lower bound to P8. One might consider two procedures for resolving the duality gap. First, one could use the dual problem for generating bounds in a branch and bound procedure; the effectiveness of this depends on the tightness of the bounds and the number of integer variables in P8. The second approach incorporates the solution of the dual problem into a heuristic procedure. For each iteration in the solution of the dual, we might generate, somehow, a corresponding feasible solution to P8. The best such feasible solution can be compared with the solution of the dual to assess its near optimality; the procedure stops once the best feasible solution is sufficiently close to the optimum or after a predetermined number of iterations.

We might solve the dual D8 directly by means of a method such as subgradient optimization, or a dual-ascent procedure. Alternatively, we can follow the general derivation provided by Magnanti et al. (1976) to reformulate the dual problem as a generalized linear program. We denote the extreme points of the convex hull defined by constraint sets (5) and (7), the non-negativity constraints and the binary constraints by

$$u^j = (u_1^j, u_2^j, \ldots u_I^j) = ((p_{1t}^j, q_{1t}^j, y_{1t}^j), \ldots (p_{I,t}^j, q_{I,t}^j, y_{I,t}^j)),$$

where for each item i, u_i^j is a Wagner-Whitin schedule.

We define the cost for the j^th extreme point and resource usage for the j^th extreme point in time period t as

$$cu^j = \sum_{\mathrm{i}=1}^{\mathrm{I}} \sum_{\mathrm{t}=1}^{\mathrm{T}} cp_{it} p_{it}^j + cq_{it} q_{it}^j + cy_{it} y_{it}^j$$

$$a_t^j = \sum_{\mathrm{i}=1}^{\mathrm{I}} a_{i1} p_{it}^j + a_{i2} y_{it}^j.$$

We can rewrite D8 in terms of the extreme points as the following equivalent linear program:

$$Max\ z$$
$$s.t.$$
$$z + \sum_{t=1}^{T} (b_t - a_t^j) p_t \leq cu^j \quad \forall j$$
$$p_t \geq 0, \forall t$$

where z is unconstrained in sign. The dual of this problem is:

$$
\begin{aligned}
&P9: \quad Min \sum_{j=1}^{J} cu^j x_j \\
&s.t. \\
&\sum_{j=1}^{J} x_j = 1 \\
&\sum_{j=1}^{J} a_t^j x_j \le b_t \qquad \forall t \\
&x_j \ge 0, \forall j
\end{aligned}
$$

where J is the number of extreme points; the decision variable x_j indicates the fraction of the schedule given by the jth extreme point. Problem P9 is a convexification of the primal problem P8 in which we replace the feasible region defined by constraints (5), (7), the non-negativity and binary constraints by the convex hull of this region. The solution of P9 provides a lower bound for P8. However, P9 is not all that useful due to the large number of variables, on the order of 2^{IT}; and the solution to P9 will typically be fractional, and not suggestive of good, near-optimal feasible solutions.

We can reformulate P9 by noting that we can express the set of extreme points U = $\{u^j\}$ in terms of extreme points for the individual items. That is,

$$U = U_1 \times U_2 \times ... \times U_I$$

where

$$U_i = \{u_i^k\}$$

is the set of extreme points or Wagner-Whitin schedules for item i. For the kth extreme point for item i, we define its cost and resource usage parameters:

$$
\begin{aligned}
&cu_i^k = \sum_{t=1}^{T} cp_{it} p_{it}^k + cq_{it} q_{it}^k + cy_{it} y_{it}^k \\
&a_{it}^k = a_{i1} p_{it}^k + a_{i2} y_{it}^k .
\end{aligned}
$$

We now can rewrite P9 in terms of the Wagner-Whitin schedules for the items:

$$
\begin{aligned}
&P10: \quad Min \sum_{i=1}^{I}\sum_{k=1}^{K_i} cu_i^k x_{ik} \\
&s.t. \\
&\sum_{k=1}^{K_i} x_{ik} = 1 \qquad\qquad \forall i \\
&\sum_{i=1}^{I}\sum_{k=1}^{K_i} a_{it}^k x_{ik} \le b_t \qquad \forall t \\
&x_{ik} \ge 0, \forall i, k
\end{aligned}
$$

where K_i denotes the number of Wagner-Whitin schedules for item i and the decision variable x_{ik} is the fraction of the k^th Wagner-Whitin schedule for item i in the solution. The first constraint assures that these fractions sum to one, so that the demand for each item is met. The second constraint enforces the resource constraint.

The optimal solution to P10 solves the dual D8, and provides a lower bound to P8. If the optimal solution to P10 is all integer (x_{ik} = 0 or 1 for all i, k), then this solution is optimal to P8. When the solution to P10 is not integer, it provides the basis for finding near optimal solutions. We describe this solution strategy in two parts: first how we solve P10 and second how we use the solutions to P10 to obtain good feasible solutions to P8.

Manne (1958) first proposed solving P10 as an approximation to solving P8. However, as with P9, there can be an enormous number of decision variables, on the order of $I2^T$ in this case. Rather than generate all of these decision variables and their parameters, we solve P10 by means of column generation [Dzielinski and Gomory (1965), Lasdon and Terjung (1971)]. At each iteration, we solve a master problem, namely a reduced version of P10 with a subset of columns. Then, using the dual values from the master problem, we solve a Wagner-Whitin problem for each item to find a new candidate schedule to enter into the master problem. The procedure iterates between the master problem and the item sub-problems until the solutions to the Wagner-Whitin sub-problems yield no new item schedules. One would typically terminate this procedure when the gap between the lower bound provided by the master problem and a known feasible solution to P8 is suitably small.

Manne (1958) noted that although solutions to P10 are usually fractional, there is an integer solution for most items. An optimal solution to P10, as well as the master problem in the column generation procedure, has I+T basic variables, as there are I+T constraints. Thus, there are at most I+T fractional variables in the solution. Each item must have at least one basic variable, in order to satisfy the convexity constraint in P10. As a consequence, no more than T items can have two or more basic variables. Thus, at least I-T items have a single basic variable, which must be integer due to the convexity constraint.

For most planning problems the number of items would be much greater than the number of time periods. For instance, we might be planning for 50 – 100 items, with 12 or 13 time periods. For I=100 and T=13, a solution to P10 provides a single Wagner-Whitin schedule for at least 87 items. For the remaining items, the solution suggests a convex combination of two or more Wagner-Whitin schedules. These schedules satisfy the demand constraints in P8, but require more setup time and cost than assumed by P10. [P10 assumes that we only incur a fraction of the setup cost and time, for a schedule at a fractional level of activity.] Thus, the set of schedules from P10 might violate the resource constraints in P8. Manne notes that in some contexts, the resource constraints are soft constraints and minor violations can be ignored. In other cases, one might apply a heuristic to modify the solution from P10 so as to make it satisfy the resource constraints. One expects that it is relatively easy to find a feasible solution to P8 from the solution to P10, as most of the items have a single schedule. Furthermore, one expects that the feasible solutions are near optimal, again due to the fact that the solution to P10 is near integer. Computational experience in Manne, Dzielinski and Gomory and Lasdon and Terjung supports these claims. Also, Trigeiro et al. (1989) describe and test a heuristic smoothing procedure for constructing feasible solutions.

Variable Redefinition

Eppen and Martin (1987) report on an alternative solution approach to P8, based on variable redefinition. They reformulate P8 so that its LP relaxation provides a very tight bound, comparable to that from the dual-based approaches of Lagrangean relaxation or column generation. They then solve the mixed integer program using general-purpose optimization codes to obtain optimal or near-optimal solutions for problems with up to 200 products and ten time periods.

Their approach is based on the optimal property of Wagner-Whitin schedules: when a production activity occurs, the production quantity covers the demand in an integer number of consecutive periods beginning with the period of production. They then define decision variables for each such production opportunity for each item. With these new variables, the schedule for each item is a shortest path problem through a network of T nodes. These shortest path problems are coupled by resource constraints that cut across the individual items.

For completeness we present the Eppen-Martin model for multi-item capacitated lot-sizing.

decision variables

z_{itk} binary decision variable to denote production of item i during time period t, where the production quantity is to satisfy demand for periods t through k

y_{it} binary decision variable to denote setup of item i in time period t

parameters

T, I number of time periods, items, respectively

a_{itk} amount of resource required in period t, if $z_{itk} = 1$

b_t amount of resource available in period t

cz_{itk} variable production and inventory holding cost for item i, for production in period t to satisfy demand from period t to k for item i.

cy_{it} setup cost for production for item i in time period t

We now formulate the mixed-integer linear program P11:

$$P11\text{:}\quad Min \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=t}^{T} cz_{itk} z_{itk} + \sum_{i=1}^{I}\sum_{t=1}^{T} cy_{it} y_{it}$$

$$s.t.$$

$$\sum_{i=1}^{I}\sum_{k=t}^{T} a_{itk} z_{itk} \le b_t \qquad \forall t$$

$$\sum_{k=1}^{T} z_{i1k} = 1 \qquad \forall i$$

$$\sum_{k=t}^{T} z_{itk} - \sum_{k=1}^{t-1} z_{ik,t-1} = 0 \quad \forall i,t$$

$$\sum_{k=t}^{T} z_{itk} \le y_{it} \qquad \forall i,t$$

$$z_{itk} \ge 0, y_{it} = 0,\ 1 \qquad \forall i,t,k$$

The first set of constraints corresponds to the resource constraint (6) in P8. The next two sets model the flow balance for the underlying shortest path problem for each item on nodes 1, 2, ...T, where z_{itk} corresponds to the flow on the arc from node t to node k+1. The last set is the forcing constraint, equivalent to (7) in P8. Note that the demand parameters d_{it} do not appear in P11. Rather they are embedded in the definition of the cost parameters cz_{itk}, and the resource coefficients a_{itk}. Also, whereas we define z_{itk} to be a binary variable, we do not make this an explicit requirement when solving P11. Eppen and Martin find that P11 provides a tight lower bound to P8 and also identifies near-optimal feasible solutions.

Extensions to the Lot Size Model

We can extend formulation P8 to incorporate all of the problem variations that were introduced for the linear-programming production-planning model. In addition we mention here three other common extensions.

In P8 we assume that any quantity could be produced, subject to the resource constraint. Furthermore, the resource usage for a production quantity consists of a fixed setup time and a variable amount linear in the production quantity. In some contexts, production occurs as a batch process, e. g., a heat treat or diffusion process. Each batch produces a fixed amount of product and consumes a fixed amount of the limited resource. We can model this by introducing an integer decision variable for the number of batches produced of an item in a time period.

A second variation is when the setups are sequence dependent; that is, the setup for an item will depend upon what was just previously processed. There is not an easy way to modify P8 to accommodate this feature. Indeed, in general, the standard representation of sequence-dependent setups is to map this into a traveling salesman problem, which results in a new level of complexity.

The third variation is when setups can be carried over from one period to the next. In P8 we assume that this is not possible; that is, every period that we produce an item we incur a setup. In some contexts, though, we might be able to preserve the last setup in the period. Thus, we would incur only one setup if we produce item i last in one period and first in the next period. Karmarkar et al. (1987) examine a single-product version of this problem. Another variation of this problem is when there are multiple products and we assume small time buckets, so that at most one item is produced in a period. Lasdon and Terjung formulate and address this problem by means of column generation. and Eppen and Martin show how to solve both of these problems efficiently by variable redefinition.

Word Count: 7686

References

Billington, P. J., J. O. McClain and L. J. Thomas, “Mathematical Approaches to Capacity-Constrained MRP Systems: Review, Formulation and Problem Reduction,” *Management Science*, Vol. 29, No. 10 (October 1983), pp.1126-1141.

Bitran, G. R. and D. Tirupati, “Hierarchical Production Planning,” In *Handbooks in Operations Research and Management Science, Volume 4, Logistics of Production and Inventory*, edited by S. C. Graves, A. H. G. Rinnooy Kan and P. H. Zipkin, Amsterdam, Elsevier Science Publishers B. V., 1993, pp. 523-568.

Bowman, Edward H., “ Production Scheduling by the Transportation Method of Linear Programming,” *Operations Research*, Vol. 4, No. 1, (February 1956), pp. 100-103.

Dzielinski, B. P. and R. E. Gomory, “Optimal Programming of Lot Sizes, Inventory and Labor Allocations,” *Management Science*, Vol. 11, No. 9 (July 1965), pp. 874-890.

Eppen, G. D. and R. K. Martin, “Solving Multi-Item Capacitated Lot-Sizing Problems Using Variable Redefinition,” *Operations Research*, Vol. 35, No. 6 (November-December 1987), pp. 832-848.

Graves, S. C., “A Review of Production Scheduling,” *Operations Research*, Vol. 29, No. 4 (July-August 1981) pp. 646-675.

Hackman, S. T. and R. C. Leachman, “A General Framework for Modeling Production,” *Management Science,* Vol. 35, No. 4 (April 1989), pp. 478-495.

Hansmann, F. and S. W. Hess, “A Linear Programming Approach to Production and Employment Scheduling,” *Management Technology*, Vol. 1, No. 1, (1960), pp. 46-51.

Hax, A. C. and H. C. Meal, “”Hierarchical Integration of Production Planning and Scheduling,” In *Studies in Management Sciences, Vol. 1: Logistics*, edited by M. A. Geisler, New York, Elsevier, 1975, pp. 53-69.

Holt, C. C., F. Modigliani, J. F. Muth and H. A. Simon, *Planning Production, Inventories and Work Force*, Englewood Cliffs NJ, Prentice-Hall, 1960.

Karmarkar, U. S., S. Kekre and S. Kekre, “The Dynamic Lotsizing Problem with Startup and Reservation Costs,” *Operations Research*, Vol. 35, No. 3 (May-June 1987), pp. 389-398.

Lasdon, L. S. and R. C. Terjung, “An Efficient Algorithm for Multi-Item Scheduling,” Operations Research, Vol. 19, No. 4 (July-August 1971), pp. 946-969.

Magnanti, T. L, J. F. Shapiro, and M. H. Wagner, “Generalized Linear Programming Solves the Dual,” *Management Science*, Vol. 22, No. 11 (July 1976), pp. 1195-1203.

Manne, A. S., "Programming of Economic Lot Sizes," *Management Science*, Vol. 4, No. 2 (January 1958), pp. 115-135.

Shapiro, J. F., "Mathematical Programming Models and Methods for Production Planning and Scheduling," In *Handbooks in Operations Research and Management Science*, *Volume 4, Logistics of Production and Inventory*, edited by S. C. Graves, A. H. G. Rinnooy Kan and P. H. Zipkin, Amsterdam, Elsevier Science Publishers B. V., 1993, pp. 371-443.

Silver, E. A., D. F. Pyke, and R. Peterson, *Inventory Management and Production Planning and Scheduling*, 3rd Edition, New York, John Wiley Inc., 1998.

Thomas. L. J. and J. O. McClain, "An Overview of Production Planning," In *Handbooks in Operations Research and Management Science, Volume 4, Logistics of Production and Inventory*, edited by S. C. Graves, A. H. G. Rinnooy Kan and P. H. Zipkin, Amsterdam, Elsevier Science Publishers B. V., 1993, pp. 333-370.

Trigeiro, W. W., L. J. Thomas and J. O. McClain, "Capacitated Lot Sizing with Setup Times," *Management Science*, Vol. 35, No. 3 (March 1989), pp. 353–366.

Vollman, T. E., W. L. Berry and D. C. Whybark, *Manufacturing Planning and Control Systems*, 3rd edition, Burr Ridge Ill., Richard D. Irwin Inc., 1992.

Wagner, H. M. and T. Whitin, "Dynamic Version of the Economic Lot Size Model," *Management Science*, Vol. 5, No. 1 (October 1958), pp. 89-96.