DEMAND FOR SERVICE QUALITY IN RAIL FREIGHT SERVICES

T.V.S. Ramamohan Rao

Indian Institute of Technology
Kanpur, U.P. 208016, India

Introduction

The existence of several shippers of different commodities at each node of the railway network at any given point of time is an important characteristic of rail freight movements. The size of the shipments will vary considerably over different commodities and from one shipper to another. Depending on their cost considerations each of the shippers can also be conceptualized as specifying the time period over which the loaded shipment must reach its destination. (1) Over a sufficiently long time horizon, the temporal changes at every node of the network can be significant.

Much of the analysis of freight movements assumes that the spatial process is in a steady state over time and consequently employs static models to analyze the patterns of demand. However, it should be recognized that the market demand for certain commodities is periodic and indivisible so that there is a significant temporal interaction in production, inventory, and shipments. Similarly, both from the viewpoint of the shipper and the railway management it may be economical to move freight in bulk periodically rather than in small lots continuously over time. This may be a consequence of the economies of scale in production and/or shipment costs. In general, over any given time horizon, freight moves in relatively fixed lot sizes with a well-defined frequency. (2)

Almost invariably the transit time, i.e., the time taken by a loaded shipment to reach its destination, is not negligible. For, once the wagon is on wheels and moving a certain amount of time would be neeed for line haul, classification and train make-up, and unloading. As Baumol and Vinod (3) pointed out, freight in transit is equivalent to an inventory on wheels. Hence, the capital carrying cost of goods in transit constitutes an important element of the total logistic cost in the case of freight movements. Further, the existence of a non-zero transit time may have the effect of the shipper specifying certain preferences for the time taken by a loaded consignment to reach its destination and/or a change in the shipment patterns. (4) Thus, as Daughety (5) put it, if the time horizon of the shippers is long enough there is a time profile of demand for each of the commodities over each of the O-D pairs consisting of specific sizes of shipment, with a well-defined frequency of movement from a given origin in conjunction with the time taken for a loaded shipment to reach its destination. (6)

The literature on freight movements, and the demand for freight services in particular, recognized these aspects as important. However, very often, the transit time as well as the frequency of service are taken to be determined exclusively by the management of the railway network. Consequently, both these aspects have been considered to be exogenous to the decision making process of the shippers. This assumption would be tenable if the operational management of the railway network is such that the supply of freight services is fully demand responsive. (7) For, then, the distinction between the demand for and supply of the quality of service levels becomes irrelevant. (8)

However, it was noted by Daughety (9) and others that the shipper's perception of the operational aspects of management and their requirement for the level of service quality is not always the same as the actual observed values (i.e., the supply of services). Hence, the service characteristics should not be taken to be the observed values in an analysis of the demand for freight movements. Further, some railway networks are characterized by excess demand given the currently observed level of operational efficiency. (10) In such cases, it would be necessary to examine the possibility of determining the demand for service quality levels as they originate from the perceptions of the shippers before analyzing possible improvements in the operational planning and consequent supply adaptations on the railway network.

The early work of Baumol and Vinod (11) set up a possible framework for developing the demand for frequency of service along with shipment size. Similarly, the models of Allen (12) and Daughety (13) examined the effect of variations in transit time on the quantum shipped. However, as of now, a comprehensive theoretical framework to consider the shipment size, frequency of service, and transit time as interdependent decisions of the shippers does not exist. The primary purpose of the present study is to develop such a theoretical framework. (14)

The Basic Model

Consider a firm producing a single commodity at location A and selling it in the only market available at location B. Assume that the firm knows the demand curve for the product with certainty. Initially let the delay involved in moving the commodity from A to B be considered to be negligible. Then, the decision which the firm has to make is about the quantities of output which it would be willing to produce and sell at each point of time. In making this choice the firm has to consider the following alternatives:

(a) Production at A as well as the shipment will be continuous. In general, the output produced is equal to the sales volume.

(b) Production at A is continous but shipments are periodic. This will be optimal whenever there are economies of shipping in bulk and/or the demand patterns are seasonal.

(c) Production at A is in large and fixed lot sizes but shipments and sales are continuous. The existence of economies of scale in production generally dictates such a policy. (15)

(d) Both the production at A and shipments to B can be periodic. It is obvious that the actual choice depends on the (a) cost of production at A, (b) cost of shipments from A to B, (c) inventory costs at A and B, and (d) the market demand at B. (16) However, for purposes of the present analysis economies of scale in shipping will be emphasized by assuming static demand curves at B and no economies of scale in production.

When the demand curve at B is assumed to be static the time profile of sales can be fixed. Then, the firm would have to create facilities to supply the output at location B at the correct time. Given the assumption that there are economies of scale in shipping in large quantities the optimal policy would be to create appropriate inventory at both A and B. The optimal size as well as the frequency of shipments will depend upon the firm's perception of the (a) costs of production at A, (b) inventory costs at A, (c) cost of shipment from A to B, and (d) the inventory cost at B.

Hence, the optimal choice of the volume of production (and sales) per unit of time and the frequency can be formalized as follows.

Let the following quantities be defined at the outset

Y	= production per unit of time at A
$C(Y)$	= costs of production
F	= number of units of time over which production is accumulated before shipping
YF	= shipment size
$t(YF)$	= tariff rate per unit of shipment
$p=f(Y)$	= price per unit of sale at B
i_a	= cost of holding one unit of inventory for one unit of time at A, and
i_b	= inventory cost at B

Then, note that the total cost of

Inventory holding at A over one cycle of F units of time = $1/2\ i_a YF^2$

Holding inventory at B = $1/2\ i_b YF^2$

Hence, maximizing the profits per unit of time results in a choice of Y and F satisfying the equations

$$f(Y) + YF_1(Y) - C_1(Y) - iF - t(YF) - YFt_1(YF) = 0, \text{ and}$$

$$i + Yt_1(YF) = 0$$

where f_1 is the derivative of f with respect to Y and $i = 1/2(i_a + i_b)$.

The specific nature of the choices of Y and F can be inferred from a simple numerical example. Fig. 1 is drawn on the following assumptions.

$$f(Y) = p_0 - p_1 Y$$

$$t(X) = t_0 - t_1 X + t_2 X^2$$

Figure 1(a) indicates that an increase in p_0 would ceteris paribus result in an increase in Y and a reduction in F. (17) The reduction in f is primarily a result of the economies in shipping costs being exhausted beyond a certain limit. The results corresponding to increases in c_1 are just the opposite and for the same reasons. Similarly, it can be observed from Figure 1(b) that an increase in t_1 increases both Y and F significantly. But the variations in all the other parameters have only a marginal effect.

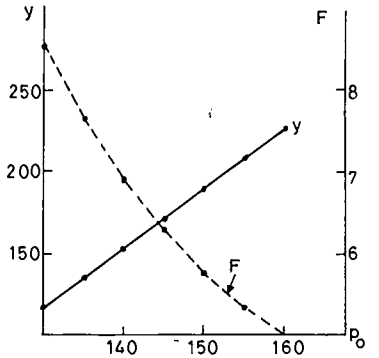
Note that an increase in F postpones the recovery of the production cost over a given length of time. Consequently it would be expected that the firm would discount future revenues. The major effect of this would be on the choice of F. However, as Fig. 2 indicates, these changes have been observed to be marginal.

In general, it can be postulated that

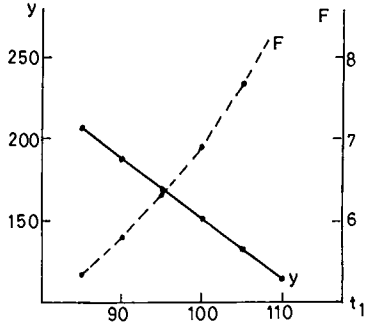
$$Y = Y(p,t);\ Y_1 > 0,\ Y_2 < 0$$

$$F = F(Y,Z);\ F_1 < 0, \text{ where}$$

p = average price in the market at B


y
F
250
200
150
8
7
6
y
F
140
150
160
p_o

(a)


y
F
250
200
150
8
7
6
F
y
90
100
110
t_1

(b)

FIG. 1

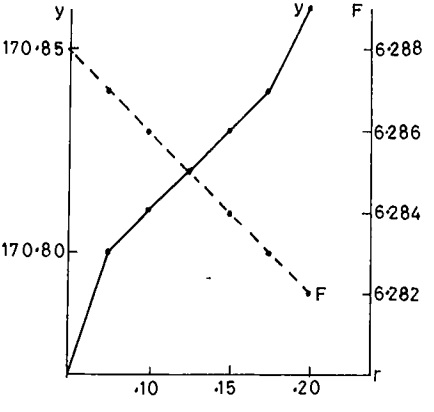

y
y
F
170·85
170·80
6·288
6·286
6·284
6·282
F
·10
·15
·20
r

FIG. 2

t = average tariff rate, and
Z = vector of other exogenous variables including the rate of discount

It may be convenient, in some applications, to write both Y and F in terms of the reduced form. The signs of the derivatives can be respecified accordingly.

Choice of Transit Time

Let T be a fixed amount of transit time which the shipper has to accept as exogenously determined by the supply constraints on the system. The costs and benefits associated with such a T would need to be identified before defining the optimal T.

To begin with, note that the inventory cycle at location B starts with a stock YT and a shipment of size YF is received only at time T. However, the total inventory over the cycle (0,F) remains $1/2\ YF^2$. Since this quantity is independent of T it would appear that there is no objective basis for defining the optimal T. However, note that in comparison to the previous section there is a change in the time profile of inventroy at B. Therefore the shipper, who adopts a discounting procedure, would find that there is a change in the cost of holding inventory. Properly accounting for the discounting policy the inventory at B can be written as

$$EX = (1/r^2)Y\{rT - 1 + rFe^{-rT} - rTe^{-rT} + e^{-rF}\}$$

Note that, for a gien F, this expression has the same value whenever T = 0 or T = F. That is, if T is small the present discounted value (i.e., at t = 0) of the inventory at B is high. However, as T increases the firm can postpone incurring the larger cost of the YF amount of inventory being carried. Hence, the basic advantage, if any, of a larger T is this reduction in inventory cost. But, notice that as T increases the initial inventory costs are increasing. In the final analysis the net gains would be positive only so long as these initial inventory costs are less than the advantages of postponing the receipt of the larger inventory YF. In the limit, there is no gain at all as T approaches F. That is, as T increases the gains are positive but they will be small initially, reach a peak and reduce to zero again as T tends to F.

It can be verified from Fig. 3 that the above expression attains a minimum for a finite T given F. (18) However, the foregoing argument did not take the costs created by inventory on wheels into account. (19) This may modify the choice of the optimal T.

These elementary considerations can be built into the model to determine the simultaneous choice of Y,F, and T by the shipper. As before, it can be verified that the

present discounted value of revenue at B = $(1/r)Yf(Y)(1 - e^{-rF})$

cost of production over the cycle (0,F) = $(1/r)C(Y)(1 - e^{-rF})$

inventory cost at A = $(1/r^2)i_aY(1 - e^{-rF} - rFe^{-rF})$

inventory cost at B = $(1/r^2)i_bY\{rT-1 + e^{-rF} + r(F-T)e^{-rT}\}$

cost of shipping = $YFt(YF)$

cost of inventory on wheels = $(1/r)i_wYF(1 - e^{-rF})$

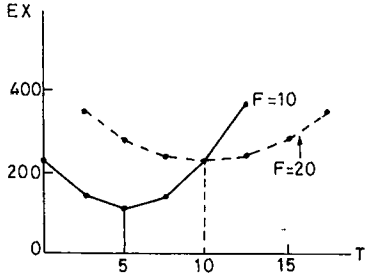
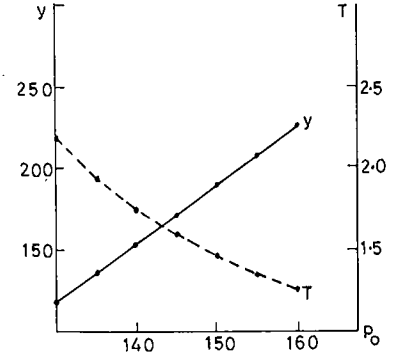

EX
400
200
0
5
10
15
T
F=10
F=20

FIG. 3


y
T
250
200
150
2·5
2·0
1·5
y
T
140
150
160
P_o
(a)

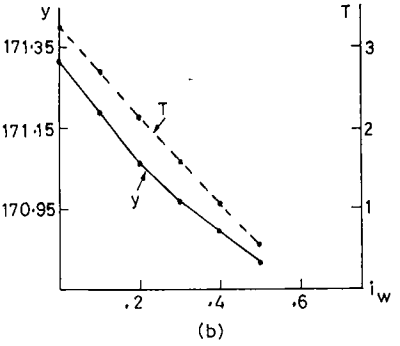

y
T
171·35
171·15
170·95
3
2
1
T
y
·2
·4
·6
i_w
(b)

FIG. 4

where i_w, the unit cost of inventory on wheels per unit of time, represents the cost of finance and loss of revenue while the goods remain in transit. Constructing the profit per unit of time and maximizing with respect to Y,F, and T will yield the optimal choices.

Fig. 4 is drawn to indicate the sensitivity of these choices to the various parameter values. Generally, the decisions of the firm are significantly affected by the variations in p_0, c_1, and i_w. From Figure 4(a) it may be noted that T decreases with an increase in p_0. The behavior of Y and F is similar to what has been observed earlier. Further, compared to Fig. 1, there is not much change in Y and consequently the value of F as well. However, the new result of this section is the trend in T reflected in Figure 4(a) as i_w varies. While T is the most sensitive to increases in i_w, it was found that F is less so and Y is affected only marginally. Hence, only the changes in T are depicted in Figure 4(b).

In general it can be concluded that

$Y = Y(p,t)$
$F = F(Y, i_w, Z)$, and
$T = T(F, i_w)$

The Two Market Case

The basic results of the preceding sections can be readily generalized to shipper demands on more general networks. For, from each node of the network many commodities will be moving to the same destination or the same firm would be shipping freight to different destinations on the network. The congestion created by the increase in freight density along any arc of the network can be taken into account either by considering the physical characteristics of the freight movements or by modifying the perception of the shippers with respect to the cost of inventory on wheels.

The nature of the extensions required to handle more general problems can be illustrated by considering a two market case. Thus, for the firm under consideration, instead of only one market at B let there be a second market at location C. Postulate that (Y, F, T) are the optimal choices in the direction of the market at B while (Y*, F*, T*) are optimal for location C. Similarly, the demand curve at C will be written as $p^* = g(Y^*)$. Consider a total length of time FF* units. Over this span of time there are F* cycles in the B direction and F cycles in the C direction. The modelling can be completed by observing the following:

(a) Y + Y* units of output will be produced at A at each unit of time throughout the time span.

(b) Out of this Y units will be stored and shipped to B once in F units of time. Similarly, Y* units are put in inventory to be shipped to C once in F* units of time.

It may now be verified that the present discounted value of revenue at

$B = Yf(Y)E/r$
$C = Y^*g(Y^*)\,E/r$

cost of production at A $= C(Y + Y^*)E/r$
cost of inventory at A $= i_a E\{G(Y,F) + G(Y^*,F^*)\}/r^2$
cost of inventory at B $= i_b EH(Y,F,T)/r^2$
cost of inventory at C $= i_c EH(Y^*,F^*,T^*)/r^2$
cost of inventory wheels $= i_w E\{J(Y,F,T) + J(Y^*,F^*,T^*)\}/r$

where $E = 1 - e^{-rFF^*}$

$$G(Y,F) = Y(1 - e^{-rF} - rFe^{-rF})/(1 - e^{-rF})$$

$$H(Y,F,T) = Y\{e^{-rF} - 1 + rT + r(F-T)e^{-rT}\}/(1 - e^{-rF}), \text{ and}$$

$$J(Y,F,T) = YF(1 - e^{-rT})/(1 - e^{-rF})$$

Hence, the profit per unit of time can be maximized as before to obtain the optimal choices from the viewpoint of the firm.

Before proceeding further it should be noted that the following special cases can arise:

(a) The demand curves in both the markets may be identically the same. Then, it would be generally expected that the firm would cater equally to both the markets. This solution would be obtained if the marginal cost of production remains constant over the entire range of output under consideration. However, if the marginal costs increase rather steeply the firm may cater to only one of the markets. The present analysis should be suitably modified to examine the possibility that extreme point solutions can be optimal.

(b) An extreme point solution can also arise when the market at one of the locations is far more advantageous relative to the other. Extensive numerical experimentation indicated that these exceptions do arise. However, these will not be pursued further.

Keeping these limitations in perspective, Fig. 5 is drawn to broadly identify the sensitivity of the optimal solutions to changes in various parameters. (20) The following salient features may be noted:

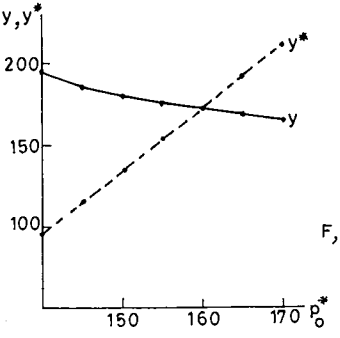
(a) As before, the choices of Y,F as well as T are sensitive to the variations in p_0, p^*_0, c_1, and i_w. In general, greater movement at more frequent intervals has been recorded for the more lucrative market.

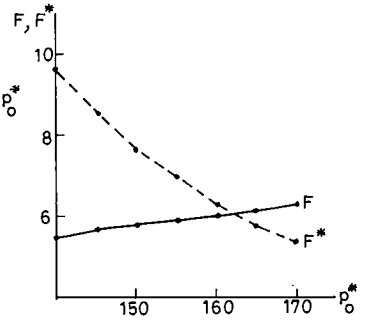
(b) Given a value of p_0, the variations in Y*, F* and T* are significant when p^*_0 changes. But the same amount of change in p^*_0 does not induce any change in T though Y and F are affected somewhat. The relative changes, even in Y and F are smaller in comparison to those observed in (Y*, F*, T*). The markets at each of the destinations have the dominant effect on the movements to the respective markets though the secondary effects cannot be ignored altogether.

Note that in the case of many commodities the same market may receive shipments from two or more production centers. From an analytical viewpoint this case is symmetric to the two market case. Hence, it can be maintained that the analysis of this section is quite general.

Summary and Comment

The very early work of Baumol and Vinod (21) emphasized the need to examine shipment size as well as the frequency of shipment as independent choices of a firm. They identified the inventory cost considerations as the major factor in such choices. The present work is not very different from this perspective though the interrelationship of various markets in such decision making environments is brought into focus.


y, y*
200
150
100
y*
y
150
160
170
p*o


F, F*
10
8
6
F
F*
150
160
170
p*o

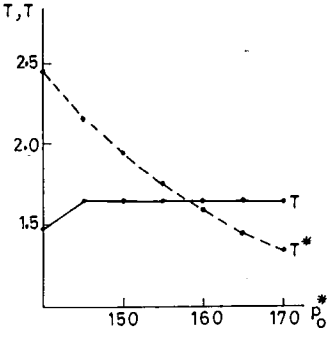

T, T
2.5
2.0
1.5
T
T*
150
160
170
p*o

FIG. 5 (a)

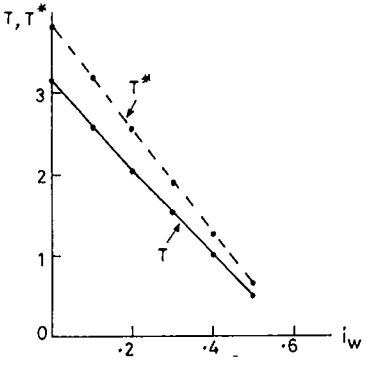

T, T*
3
2
1
0
T*
T
.2
.4
.6
iw

FIG. 5 (b)

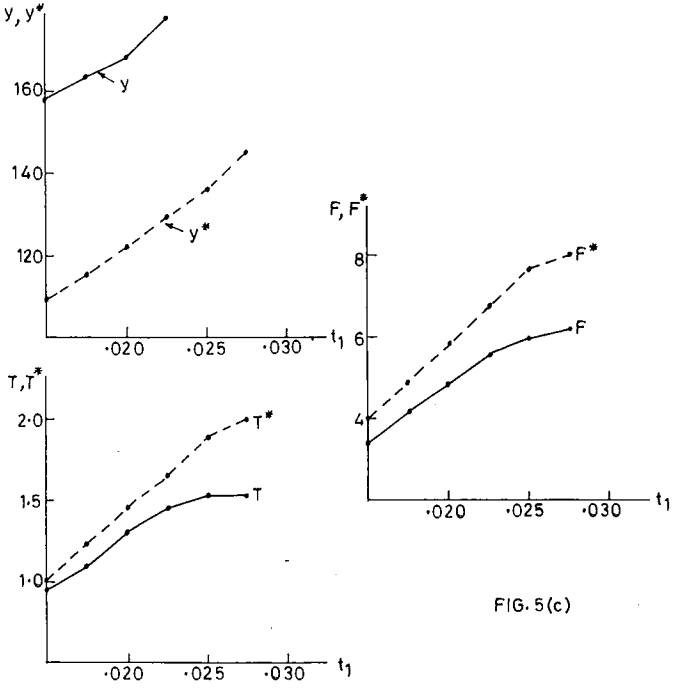


FIG. 5(c)

The work of Allen (22) appeared to offer a basis in economic theory to examine the ex ante choices of transit time by the shippers. However, the assumption that tariff rates are adjusted to transit time variations is untenable and the model cannot be reconstructed meaningfully if this assumption is omitted. Hence, it was necessary to reexamine the gains and costs, to the shipper, of variations in transit time. Once again the inventory cost considerations were invoked to develop a theoretically satisfactory result. The inventory on wheels characterization in Baumol and Vinod (23) appears to be important in the shipper's valuation of the changes in transit time as well.

These simple changes in the conceptualization are amenable to further generalization. It appears that even the aggregate service quality characteristics of the transportation network, for which none of the shippers can individually express a preference, can be taken into account in the framework of the models developed here. The reader is referred to Rao and Sriraman (24) for a more detailed empirical analysis in the context of Indian railways.

Limitations of the algebraic procedure appear to be quite formidable if very general theoretical results are desired beyond the broad qualitative answers provided by the present study.

REFERENCES

1. This argument assumes that the supply procedures are flexible. Such an ex ante analysis would be useful in identifying the level of service which the system should provide even if the observed level differs from it.

2. The total shipments over the given time horizon can be looked upon as the product of average shipment size and frequency. Many studies acknowledge that the shipper has an independent choice of the shipment size as well as the frequency of shipments. See, in particular, Baumol, W.J., and H.D. Vinod (1970), An Inventory Theoretic Model of Freight Transport Demand, Management Science, March, pp. 413-421, Friedlaender, A.F., and R.H. Spady (1981), Freight Transport Regulation: Equity, Efficiency, and Competition in the Rail and Trucking Industries, Cambridge: M.I.T. Press, and Roberts, P.O. (1977), Forecasting Freight Demand, in E.J. Visser (ed.) Transport Decisions in an Age of Uncertainty, Hague: Martinus Nijhoff.

3. Baumol and Vinod, ibid.

4. It would be important to note that changes in transit time not only affect these decisions but can also have an influence on the shipment sizes and frequency of shipments. The three decisions are significantly interrelated.

5. Daughety, A.F. (1979), Freight Transport Demand Revisited: A Microeconomic View of Multi-model, Multi-characteristic Uncertainty and Demand for Freight Transport, Transportation Research, December, pp. 281-288.

6. But the transit time observed in practice, in its turn, will be a function of the speed of trains, the number of hours a wagon is on wheels and moving during a day, the delays in classification yards, the time taken in loading, unloading, etc. The shipper may not be able to express a preference for each of these dimensions in particular.

7. For a discussion on this aspect see Rao, T.V.S.R. (1982), Supply of Freight Services on a Railway Network, International Journal of Transport Economics, August, pp. 171-192.

8. For a detailed analysis of this issue and an empirical evaluation in the context of Indian railways the reader is referred to Rao, T.V.S.R., and S. Sriraman (1985), Disequilibrium in Rail Freight Services, New Delhi: Ajanta Books International.

9. Daughety, op.cit.

10. Rao, T.V.S.R., and S. Sriraman (1985), Efficiency in Railway Transportation: The Indian Experience, in K.J. Button (ed.) International Railway Economics, Aldershot: Gower Press, documented the existence of this phenomenon on Indian railways and the reasons for it in the context of both freight and passenger services.

11. Baumol and Vinod, op.cit.

12. Allen, W.B. (1977), The Demand for Freight Transportation: A Micro Approach, Transportation Research, February, pp. 9-14.

13. Daughety, op. cit.

14. The basic framework is similar to Rao, T.V.S.R. et al. (1981), Length of Haul in Rail Freight Movements, International Journal of Transport Economics, April, pp. 89-100, and extends the work of Allen, op.cit., Baumol and Vinod, op.cit.

15. Inventory models which approximate such a choice are considered in Arvan, A., and L.N. Moses (1982), Inventory investment and the Theory of the firm, American Economic Review, March, pp. 186-193.

16. One or more of these patterns may be relevant even if the demand curves are uncertain.

17. The values of the other parameters are $p_1 = 0.1$, $c_0 = 400.0$, $c_1 = 95.0$, $c_2 = 0.04$, $t_0 = 17.5$, $t_1 = 0.025$, $t_2 = 0.00001$, and $i = 0.6$. note that one day is the unit of time throughout the rest of the analysis.

18. Throughout the rest of the numerical illustration and the results of the subsequent sections both F and T are in days.

19. It can be readily verified that there is no reduction in the inventory cost at B if $i_w > i_b$. Under these conditions the optimal value of T is zero.

20. Note that the default values of the parameters are $p_0 = 160.0$, $p_1 = 0.1$, $p^*_0 = 0.1$, $i_a = 0.6$, $i_b = 0.6$, $i_c = 0.6$, $i_w = 0.3$, and $r = 0.15$.

21. Baumol and Vinod, op.cit.

22. Allen, op.cit.

23. Baumol and Vinod, op.cit.

24. Rao and Sriraman, op.cit.