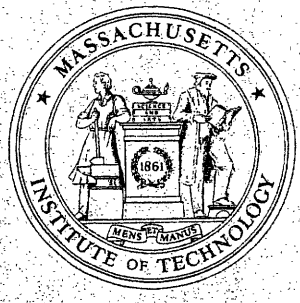
OPERATIONS RESEARCH CENTER

working paper



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SUPPLY MODELLING OF RAIL NETWORKS: TOWARD A ROUTING/MAKEUP MODEL

by

Arjang A. Assad

OR 069-77 December 1977

This research was supported, in part, by the U.S. Department of Transportation under Contract DOT-TSC-1058, Transportation Advanced Research Program (TARP).

ABSTRACT

Freight flow management in rail systems involves multicommodity flows on a network complicated by node activities (queueing and classification of cars at marshalling yards). Routing in these systems should account for technology requirements of motive power and traction as well as resource allocation (cars to blocks, blocks to trains). In this paper, we propose a hierarchical taxonomy of modelling issues and describe a class of models dealing with car routing and train makeup from the viewpoint of network flows and combinatorial optimization. We compare our model with two previous rail network models and discuss possibilities for algorithmic development.

1. INTRODUCTION

Railroads in the United States have been facing fierce competition in the area of freight transportation since 1940. This is reflected in the steadily falling market share of railroads in intercity freight transport. Over the period 1940-1970, total intercity freight ton miles have tripled while rail tone miles have barely doubled. The market shares of trucks and oil pipelines have increased by 10% at the expense of the rail market share. Finally gross revenues of railroads have been declining steadily over the past two decades [18].

The impact of this competition is felt in major reorganizations of railroads and in a renewed stimulus for implementing more rationalized planning systems, especially in view of the recent capabilities of computerization in rail systems.

On the other hand, a number of studies have pointed to poor utilization of available resources in railroads: Total origin-destination trip times are unduly large due to various delays incurred at intermediate points. Moreover, the variance in such trip times is also quite large, resulting in unreliability of the delivery process and poor customer service. A typical railcar moves only 2 hours per day on a train, spending the remaining time at yards. Only 16.5 hours are required to move 500 miles at 30 mph while, on the average, a car spends more than 20 hours to move through a yard and typically visits 5-10 yards before reaching its ultimate destination [18]. These and similar statistics indicate low car utilization and poor service to the customer. The

complicated interaction between rail policies affecting these issues show the need for a global planning process for the rationalization of rail operations. A methodical improvement of such operations may have a far greater impact than purely technological advances in rail engineering. As an example Thomet [15] estimates that doubling the speed in mail-haul movements reduces the travel-time by only 15%. A methodology for analyzing current policies in rail freight management and their coordination could not be more timely. In this context analytical models for rail systems hold much promise in planning the acquisition of facilities and resources as well as evaluating the effect of changes in the parameters of the system (such as traffic demand patterns, rates of yard activities and so forth).

Previously various subsystems of rail systems have been modelled in some detail(see[4]). The simulation approach dominates the existing literature in this area. However, to provide meaningful insights for planning purposes, a typical simulation model has to be run on a variety of different parameters, the number of which may be quite large. Such an approach may easily become expensive in terms of computation costs while it still requires the specification of a set of performance measures to compare the different outcomes. Optimization-based models avoid such shortcomings by formalizing the performance criterion thus taking full account of the tradeoffs involved in parameter changes.

Optimization models have been successfully used in transportation studies and in particular cases, such as traffic equilibrium problems (see [8]), their utility and efficacy is well established. While unable

to capture the full details of specific operations, such models can serve as a valuable aid to decision-making if used at an aggregate level. Moreover the possibility of interacting between an aggregate optimization model and a more detailed simulation model has been advocated by a number of researchers [1], [10].

Rail networks share the basic network structure of other transportation systems on the main-haul links, but are additionally complicated by the activities taking place in intermediate marshalling or classification yards (which correspond to the nodes of the network). As noted previously, the delays at such yards form the substantial portion of the total travel time of a typical freight car from its origin to its destination. As a result, any network model of railroads should faithfully reflect the activities at the nodes. Moreover the two sets of line and yard activities are interdependent and interact fully. The basic task of a comprehensive rail network model is to link these two and account for their interactions.

In this report we shall describe a mathematical programming approach to this problem and relate it to previous modelling efforts in this direction. In describing a formal model, we shall have the opportunity to point out different issues of concern to the rail community and explore to what extent they can be faithfully reflected in the model without precluding the algorithmic tractability of solving the resulting optimization problem.

The plan of this report will be as follows: In Section 2 we describe a number of issues in rail operating policies as well as major decision-making problems that they pose. We suggest a hierarchical view of the decisions involved. Section 3 provides the terminology of network flows and the

formulation of a general network model for train routing and makeup. In Section 4, we discuss various forms of the general model and discuss its capabilities as well as the issues it captures. Section 5 contains a brief review of the two existing optimization models, a comparison of these efforts with the proposal made here, and a discussion of potential advantages of our model.

2. ISSUES IN RAIL PLANNING AND THE HIERARCHICAL APPROACH

In this section we give a broad description of rail systems to establish the context of our modelling approach and to set the terminology for our discussion of planning issues.

Broadly speaking we may view rail operating policies as a sequence of decisions striving to meet demands by a suitable allocation of resources and facilities available to the railroad (which we may view as the supplier of services). On the demand side, we assume that data is available in the form of the traffic volume to be moved between a given origin-destination pair. We shall only deal with average (deterministic) estimates of such volumes. In practice, the demand requirements may be more complicated. The shipper might specify a maximum allowable delivery time or specific constraints on routing. On the supply side, the specified set of resources available to the railroad determines the feasible train routes, allowable train itineraries, crew and motive power availabilities, and yard facilities. The operating policies determine an assignment of the resources to each class of traffic (determined by its origin and destination as well as possibly traffic type). Operating policies may be roughly divided into line and yard policies. The former determines the routing of each traffic class on the

physical rail network as well as a sequence of trains to which the traffic is assigned. Yard policies specify the operations performed on different classes of traffic in the yards they visit:

At each yard the incoming traffic undergoes a sequence of operations ultimately leading to a regrouping of this traffic for outbound trains. The grouping policy at each yard specifies how the incoming traffic is reclassified into a number of groups in each of which the outbound cars share a destination further downstream along their routes. The blocking of cars into such destination-oriented blocks is also called the blocking policy. Incoming trains are inspected and then decoupled to reclassify their cars into appropriate outbound groups. Such blocks of traffic are then placed on classification or departure tracks awaiting an outbound train. The decisions involved in the process described above may be collectively called the Classification policy of the yard.

Each outbound train has a "take-list" specifying the blocks of traffic it may pick up at a given yard. The decision as to which blocks of traffic should be placed on a given train is called the Make-up policy. Obviously, the Make-up policy interacts highly with both Classification and Routing policies. It may thus be viewed as an important linking factor between yard and line decisions. In this report we shall concentrate on routing and makeup policies and their interaction with the classification work performed at a given yard. We shall not discuss other yard policies relating to the receiving and dispatching activities.

The decisions of interest in rail management vary substantially in scope, time horizon, investment requirements, and the level of managerial decision-making. Obviously the location of a major new classification yard

will involve top level management whereas timetabling and a number of yard policies may not extend beyond local operators at a given yard or a zone of several yards. This observation suggests it may be useful to adopt a hierarchical view of rail planning. We shall follow the framework proposed by Anthony [2] identifying three levels of decisions facing management, namely strategic, tactical, and operational. We shall stress that we regard the following categorization as a tentative aid to our modelling effort allowing us to identify a suitable level of aggregation in our model.

I. Strategic Decisions

These involve resource acquisition decisions of long time horizons typically requiring major capital investments. Due to the pervasive and long-lasting impact of strategic decisions on the future of the system, top level management is usually directly involved with their resolution. Prime examples of such decisions in the context of rail systems include:

a - Network Design and Improvement. Track Abandonment.

b - Location of yards and major classification facilities within large yards.

c - Highly Aggregate Routing Decisions. Long term planning of train services.

The network improvement model of LeBlanc [11] is one example of a strategic problem studied in the rail literature. The problem of choosing a set of feasible routings over a long planning horizon may fall into the category of strategic decisions if highly aggregate measures of traffic demand are used. The railroad may also want to estimate the costs and impact of providing regular service on a given set of routes and use such aggregate routing models as aids to decision-making. These models may also

serve as inputs into larger models dealing with trip distribution and modal split considerations.

II. Tactical Decisions

These decisions have medium term planning horizons and focus on rational and effective allocation of existing resources rather than major acquisitions. The intermediate level of aggregation and medium term horizon of a tactical planning model allow it to take account of broad changes in system parameters and data (such as seasonalities in the traffic volumes and imbalances resulting from lack of uniformity in the geographical pattern of shipments) without having to incorporate day-to-day changes in the data-base. Some examples of tactical decisions in rail systems follow:

a) Train Selection and Traffic Routing: What trains should run and what should the required frequency of each train be to accomodate traffic demand?

b) Train Makeup: What groups (or blocks) of traffic should a train be allowed to carry (its take-list) at a given yard of its itinerary?

c) Yard Classification Policy: Into what groups or blocks should the incoming traffic to the yard be consolidated?

d) Allocation of Classification Work among yards: What is the total amount of classification work performed in the system? How should this workload be distributed among the various yards to account for the fact that they might have different technological capabilities?

e) Train lengths: What train lengths should we consider economically attractive? Is it better to have shorter more frequent trains?

These issues have been at the heart of rail operations for some time. The decisions mentioned above influence one another to a large extent and no one model can hope to fully capture all such interactions. We wish, however, to address some of these questions with the model proposed in this report.

Note that the specification of Makeup and Classification policies is expected to be responsive only to major, stable changes in traffic patterns. Changing these policies on a daily basis in response to daily fluctuations is not advisable in view of the confusion it may cause for yard and management personnel. This suggests that a model for setting yard policies will probably be solved on a monthly or quarterly basis.

III. Operational Decisions

Such decisions deal with day-to-day operational and scheduling activities at a high degree of detail and in a fairly dynamic environment. Correspondingly only low levels of management (such as yardmasters) are directly concerened with operational decisions. Some examples are:

a) Train Timetables: Determining arrival/departure times of each train at any intermediate station of its itinerary.

b) Track Scheduling and Priority Policy: Assigning trains to tracks if track capacity is limited. Planning for meets and overtakes according to a priority scheme.

c) Engine Scheduling: Planning for daily distribution of motive power units over a specified set of train schedules.

d) Empty Car Distribution: Distributing empty cars over the rail network to meet demand and rectify imbalances due to uneven freight movement.

e) Yard Receiving and Dispatching Policies: Determining a priority scheme for processing the queue of trains incoming to a classification yard. Setting rules for departure times of outbound trains.

f) Line-haul and yard Maintenance Operations: Scheduling maintenance and inspection for cars, engines, tracks and yard facilities.

The element of timing is crucial to most of the decisions in this list. The aim of computerized information systems for railroads is to record the position of cars in real-time or on an hourly basis. Thus models for empty-car

and engine distribution may well be re-solved on a daily basis. We wish to point our that the distinction between tactical and operational issues might be blurred at times. Thus while daily empty car distribution is classified as operational, these might be a more aggregate tactical model guiding the distribution on a zonal basis. Similarly the empty car problem viewed in the context of setting optimal stock levels of empty cars at yards may assume tactical dimensions.

The main advantage of a hierarchical approach is to avoid the pitfall of dealing with all the decisions outlined above simultaneously through a monolithic model. Even if computer and algorithmic capabilities permitted the solution of a large scale detailed rail model (which is presently not the case), this approach would still be inappropriate since it would not be responsive to managerial needs at each level of the organization. Thus we do not try to integrate service and route selection problems for trains with delay timetabling and empty car distribution problems into the same model, since the two classes of problems relate to different levels of the managerial heirarchy.

3. A GENERAL RAIL NETWORK MODEL

In this section we shall formulate a rail network model with a general objective function that may be suitably specialized to capture various costs associated with the rail operations of routing, makeup, and classification. Our main concern lies in (i)-modelling the interaction of routing decisions with yard activities and (ii)-capturing the economies associated with consolidating blocks of traffic into a single train. Since most of the delays

in freight delivery occur at yards, an effective policy for reducing congestion and delays (and their variability) associated with the yards is to schedule by-pass trains that deliver blocks of traffic directly between origin and destination yards with no intermediate re-classification. The savings resulting from such a procedure, however, have to warrant the additional costs of allocating a direct (unit) train to the traffic class in question. Our model is constructed to address this issue and to provide a means of striking a rational balance between customer service and rail operating costs.

We start by reviewing the terminology of network flows and the notation our formulation will require.

3.1 Rail Network Structure

We shall deal with a network of nodes N corresponding to yard and links A_p referring to physical track sections on which main-haul rail freight may travel. We thus have a directed graph $G_p = (N, A_p)$ which we shall assume to be acyclic.

The traffic requirements from one yard to another are taken as input data. Consider a pair of nodes p and q between which a nonzero flow of traffic is specified. Such a pair is called an origin-destination (or simply OD) pair. Given the OD pair (p,q) a required flow of r^{pq} units must travel from node p to node q on the physical network G_p. The requirements r^{pq} may be measured in number of cars, or possibly in some form of equivalent tonnage.

The physical network G_p is then endowed with a route structure which specifies the set of feasible routes on which trains could be scheduled. Each route has a unique origin i and destination j, i and j being yards in the rail network, as well as a physical path (i.e. a series of links in A_p) from i to j which is completely specified. In the rail literature such a route is occasionally referred to as a train. It corresponds to the itinerary

of a unique train from i to j. A number of trains (in the sense of a string of cars provided with locomotives) may be run on each train route. Thus we distinguish between a train route (which specifies the itinerary through the physical network) and train frequency (which specifies the number of actual trains dispatched along that route). It is useful to think of train routes as a bus number (line) which has a specified itinerary. Obviously busses with the same number may be run on a given itinerary with any desired frequency, however all of these share the same routing and stop-schedule.

A train service from i to j, denoted [i,j], maintains its identity throughout its itinerary. It is therefore made up at yard i and broken up for reclassification at its destination j. No intermediate yards perform any classification activities on the train. In this model, we shall ignore the option of stopping at an intermediate yard to set off or pick up. We make this simplifying assumption mainly in order to simplify our exposition. A further refinement of our model may incorporate such stops explicitly with no substantial modifications. In our model, then, all intermediate yards between i and j are "bypassed" by the [i,j] train service. As a result, traffic groups which make connections travel on more than one train route. The train routes, or non-stop legs of traffic movement, may be represented as a network by adding "route arcs" between nodes i and j as illustrated in the following example.

Example 1: Consider the very simple line betwork of Figure 1 which is composed of four stations. Only one way traffic (eastbound) is considered accounting for the directed links as shown.

(see next page)

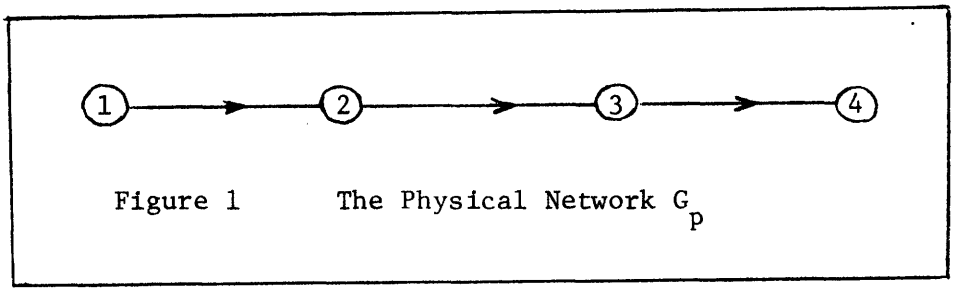


Figure 1 The Physical Network G_p

We may augment this network G_p by all direct arcs from one node to another further downstream. Thus we may, for example, add the direct arc (1,4) which corresponds to a train going from 1 to 4 with no intermediate processing. In this case we get 6 arcs in the enlarged network G as shown in Figure 2. Note that traffic from one node to another may use a sequence of trains, that is, it might make connections. Thus 1 to 4 traffic may be placed on the train [1,3] and then transferred to train [3,4]. Figure 2 exhibits all possible train routes between the four stations. Obviously the network G no longer represents links but, rather train services between yards.

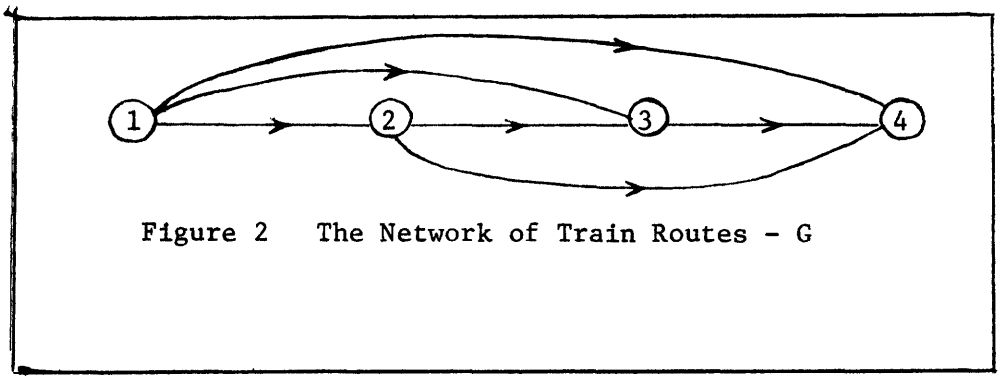


Figure 2 The Network of Train Routes - G

In general a direct arc [i,j] on the route network G constructed above represents a feasible train service with a specified itinerary on the physical links joining yards i and j. For the special case of a line network of n consecutive stations on a line, G will have n(n-1)/2 arcs. In practice, of course, other management considerations, such as inspection requirements and constraints on non-stop travel time, may rule out some of these routes.

3.2 Network Flows We now consider a given network G = (N,A) of possible routes (represented by the arcs A of G). For a given origin-destination pair (p,q) we define the following decision variables:

x_{ij}^{pq} = number of cars travelling from p to q on the service [i,j].

y_{ij} = number of engines (units of motive power) provided on the service [i,j].

Both of these variables are measured as average values over a given period (say a day or a week). In what follows we may also envisage x_{ij}^{pq} to mean the tons of freight going from p to q on [i,j]. On the enlarged route network G,we may view the variables x_{ij}^{pq} as arc flows since the arc (i,j) of G corresponds to the train [i,j]. Then for each OD pair (p,q) we have network flows x_{ij}^{pq} which are subject to flow balance constraints. As the flows of traffic between different OD pairs should be distinguished from one another (not mixed), we are dealing with a multicommodity flow on the network G (see[3]).

Let us also specify for an arc (i,j) of G an allowable car/locomotive ratio α_{ij}. Thus if $\alpha_{ij} = 50$ a single motive power unit can haul a maximum

of 50 cars over the route of the train [i,j]. We must impose conditions on the flows x_{ij}^{pq} to ensure that this restriction is not violated. Thus we will have two sets of constraints corresponding to flow conservation and motive power constraints. We shall illustrate these by an example before passing on to the general formulation.

Example 2: Let us consider the network G of Figure 2 and write out the corresponding constraints. Note that our OD pairs are (1,2), (1,3), (1,4), (2,3), (2,4), and (3,4) in this case.

a) Flow balance equations: (for each OD pair as noted on the left).

$$(1,2) \quad \{ \quad x_{12}^{12} = r^{12}$$

$$(1,3) \quad \begin{cases} x_{12}^{13} + x_{13}^{13} = r^{13} \\ -x_{23}^{13} - x_{13}^{13} = -r^{13} \end{cases}$$

$$(1,4) \quad \begin{cases} x_{12}^{14} + x_{13}^{14} + x_{14}^{14} = r^{14} \\ x_{12}^{14} - x_{23}^{14} - x_{24}^{14} = 0 \\ x_{14}^{14} + x_{23}^{14} - x_{34}^{14} = 0 \\ -x_{14}^{14} - x_{24}^{14} - x_{34}^{14} = -r^{14} \end{cases}$$

$$(2,3) \quad \{ \quad x_{23}^{23} = r^{23}$$

$$(2,4) \quad \begin{cases} x_{23}^{24} + x_{24}^{24} = r^{24} \\ x_{23}^{24} - x_{34}^{24} = 0 \\ -x_{23}^{24} - x_{24}^{24} = -r^{24} \end{cases}$$

$$(3,4) \quad \{ \quad x_{34}^{34} = r^{34}$$

b) Motive Power Restrictions for each train (arc).

$$\begin{aligned} x_{12}^{12} + x_{12}^{13} + x_{12}^{14} &\leq \alpha_{12} \cdot y_{12} \\ x_{13}^{13} + x_{13}^{14} &\leq \alpha_{13} \cdot y_{13} \\ x_{14}^{14} &\leq \alpha_{14} \cdot y_{14} \end{aligned}$$

$$\begin{aligned} x_{23}^{13} + x_{23}^{14} + x_{23}^{23} + x_{23}^{24} &\leq \alpha_{23} \cdot y_{23} \\ x_{24}^{14} + x_{24}^{24} &\leq \alpha_{24} \cdot y_{24} \\ x_{34}^{14} + x_{34}^{24} + x_{34}^{34} &\leq \alpha_{34} \cdot y_{34} \end{aligned}$$

$$\text{all} \quad x_{ij}^{pq} \geq 0 \quad ; \quad y_{ij} \geq 0 \text{ and integral.}$$

For the general formulation, we may use the following notation.

Let I_i and O_i be the set of incoming and outgoing trains(respectively) at yard i, that is:

$$I_i = \{k \mid (k,i) \in A\}$$

$$O_i = \{k \mid (i,k) \in A\}$$

Moreover for a train [i,j], let T_{ij} be its take-list, i.e. the set of all OD pairs whose traffic could be put on that train. In network terminology this lists all the commodities which may flow on the arc (i,j) of G. This set may be specified easily from the network configuration. Indeed let us call a *node j accessible to node i* if there is a path in G going from i to j.

Then

$$T_{ij} = \{ (p,q) \mid i \text{ is accessible to } p \text{ and } q \text{ is accessible to } j \}.$$

For the simple case of a line network of n consecutive yards where $N=\{1,\ldots,n\}$ and $A = \{ (i,j) \mid 1 \leq i < j \leq n \}$ we have

$$I_i = \{k \mid 1 \leq k < i\} \quad , \quad O_i = \{k \mid i < k \leq n\}$$

and

$$T_{ij} = \{ (p,q) \mid 1 \leq p \leq i < j \leq q \leq n \}.$$

The flow balance and motive power constraints are:

$$\sum_{j \in O_i} x_{ij}^{pq} - \sum_{j \in I_i} x_{ji}^{pq} = \begin{cases} r^{pq} & \text{if } i=p \\ -r^{pq} & \text{if } i=q \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

for all nodes i used OD pairs (p,q)

$$x_{ij} \triangleq \sum_{(p,q) \in T_{ij}} x_{ij}^{pq} \leq \alpha_{ij} \cdot y_{ij} \qquad (2)$$

$$\text{all } x_{ij}^{pq} \geq 0 \ . \ y_{ij} \geq 0 \text{ and integral} \qquad (3)$$

for all $(i,j) \in A$.

In (2) x_{ij} denotes the total flow of traffic assigned to train service [i,j].

Note that for fixed values of y_{ij} the constraint set is that of a capacitated multicommodity flow problem. Our approach is to maintain this feasible set (which is shared with many other transportation models involving flow of goods and carriers) and incorporate complications into the objective function. The objective function will contain two types of costs: (i) main-haul or over-the-road costs and (ii) yard costs.

3.3 Model Formulation For a given train service [i,j], let X_{ij} be a vector with components x_{ij}^{pq} for all (p,q) in the take list T_{ij} listed, say, according to lexicographic ordering of pq. This vector will fully describe the makeup (or composition) of the train [i,j], i.e. how many cars of each traffic class (p,q) is placed on the train. At a given yard j we also form a supervector $\underline{\bar{X}}_j$ containing the vectors X_{ij} for all trains [i,j] incoming to yard j (i.e. for all $i \in I_j$). This vector contains the composition of all traffic brough into yard j. As an example for a yard j of the line network,

$$\underline{\bar{X}}_j = (X_{1j}, X_{2j}, \ldots, X_{j-1,j})$$

and

$$X_{ij} = (x_{ij}^{1j}, \ldots, x_{ij}^{1n}, x_{ij}^{2j}, \ldots, x_{ij}^{2n}, \ldots x_{ij}^{ij}, \ldots, x_{ij}^{in})$$

We consider two types of costs: Train costs relating to running a train [i,j] over its route with a specified load assumed to be expressible as a cost function $\psi_{ij}(X_{ij}, y_{ij})$. Yard costs at a typical yard j with an incoming traffic $\underline{\bar{X}}_j$, given by the functional form $\phi_j(\underline{\bar{X}}_j)$. Then a formal model may

be set up as follows:

$$(P): \quad \text{Min } Z = \sum_{(i,j)\in A} \psi_{ij}(X_{ij}, y_{ij}) + \sum_{j\in N} \phi_j(\bar{X}_j) \qquad (4)$$

subject to equations (1), (2), and (3).

At this point we should pause to list some of the costs which we would expect the above cost functions to incorporate. In rail cost accounting a number of different cost factors such as maintenance and constraint crew costs may be attached as unit costs to a number of aggregate measures including: total gross-ton-miles, locomotive miles, train miles, over-the-road engine and car hours. In an optimization model, it is preferable to identify major components of the cost individually to bring out impact of alternative policies on the costs more clearly.

I. Trains costs will include:

i) Crew Costs - A crew should be engaged over the entire length of the train route.

ii) Fuel Costs - These will depend on the train weight (total tonnage) and length (number of cars) as well as on the train's speed and the geograhical terrain of its route. It is reasonable to take fuel costs as being proportional to train weight over a given link.

iii) Costs of Motive Power - We attach costs to providing and running an engine over each link.

iv) Over-the-Road Delay Costs - The delay incurred on the main-haul legs, due to travel-time, congestion, meets and passes and so forth, may be attached a dollar value to reflect the value of capital tied up in the system pipelines and, more importantly, the shipper's devaluation rate.

II. Yard Costs will include:

i) Inspection and Classification Costs - These are attached to the operations involved in inspecting and breaking up incoming trains for reclassification into outbound groups. They may reflect the use of yard equipment (yard-engine-hours) or labor resources (inspection crews).

ii) Yard Delay Costs - We may associate time (devaluation) costs to the total delay suffered by cars in the yard due to the queueing effects and waiting times of various yard operations: receiving, classification, outbound inspection and assembly, accumulation and connection delays.

Naturally in a yard with fixed resources (crew and equipment) the main component of yard costs is formed by delays incurred by different classes of traffic at that yard.

Let us try to suggest functional forms for the cost functions ψ_{ij} and ϕ_j that would account for the components listed above. Starting with train costs ψ_{ij}, let

$$\psi_{ij}(X_{ij},y_{ij}) = c^e_{ij} \cdot y_{ij} + c^c_{ij} \cdot x_{ij} \qquad (5)$$

where

c^e_{ij} = cost of providing a unit of motive power for train [i,j]

c^c_{ij} = hauling cost per car on the route of [i,j]

and x_{ij}, we recall, is the total load on train [i,j] as defined in equation (2). Thus we have variable costs corresponding to the number of cars on the train (and track parameters for the train route) as well as engine costs which exhibit a discrete character peculiar to rail systems. Later, we shall point

out why we believe modelling this discreteness in the cost function is important. For the moment, note that for fixed values of x_{ij}^{pq} (and hence x_{ij}), given that the cost function is strictly increasing in y_{ij}, the values of the y_{ij} variables may be deduced from equation (2) to be

$$y_{ij} = [\alpha_{ij}^{-1} \cdot x_{ij}]^{+}$$

where $[\,.\,]^{+}$ denotes rounding up to the nearest integer. Consequently the graph of ψ_{ij} as a function of x_{ij} exhibits a stepwise nature as shown in Figure 3. The magnitude of the jumps is c_{ij}^{e} and slope of the linear sections equals c_{ij}^{c}.

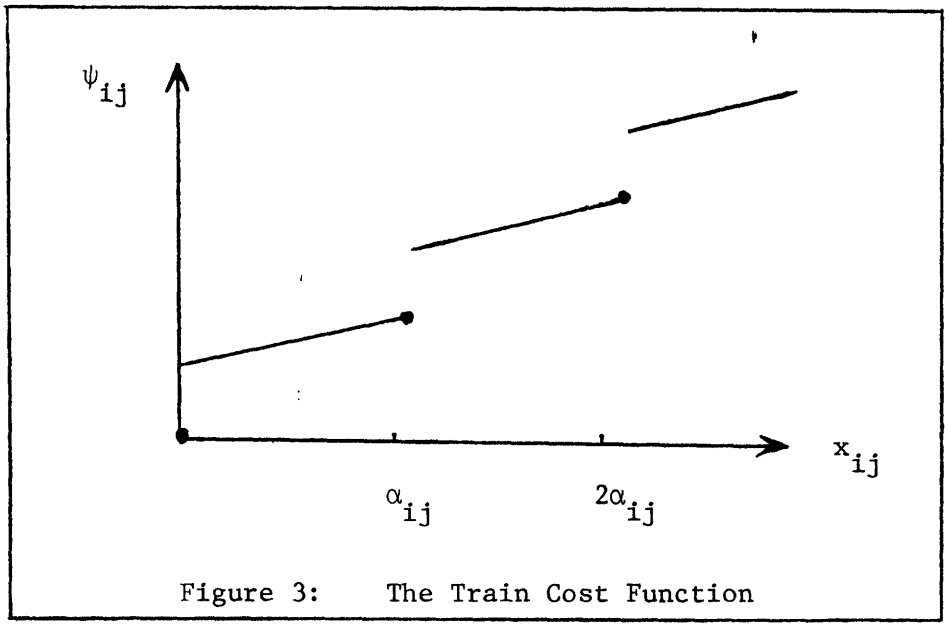


Figure 3: The Train Cost Function

We shall now turn to the yard cost function ϕ_j. To choose the simplest expression for these costs, we might let ϕ_j depend on the total throughput of yard j, that is

$$\phi_j(\underline{X}_j) = f_j\left(\sum_{i \epsilon I_j} x_{ij}\right) . \qquad (6)$$

To allow for congestion effects, we may take f_j to be a convex increasing function. Moreover the network structure of the problem will be preserved if the node for yard j is split into two nodes joined by a 'throughput arc' with flow $\sum_{i \in I_j} x_{ij}$ and arc cost $f_j(\sum_{i \in I_j} x_{ij})$. For a tactical model, however, we consider this expression to be too crude. It does not distinguish the delaying effects of one class of traffic on others.

Following Thomet [15,16] we introduce a delay function at a given yard j of the form

$$W_j + v_j \cdot x_{ij}$$

where W_j is the fixed delay for processing a train coming into the yard and v_j is the variable delay (in units of time per car). Such a delay is incurred for any train [i,j] which is classified and processed at yard j. To account for the effect of train composition in the processing costs, we should note that trains composed purely of cars for yard j need not be classified at that yard.

Such cars need not be placed on the outbound traffic groups to continue their journey, but will rather be delivered to local industrial sidings from the destination yard and thus exit the main rail network. As an example, in Figure 2 the trains [1,4], [2,4], and [3,4] all have cars destined for yard 4 and have no cars with a different destination. Consequently, these trains will not be classified at yard 4, but will just be transferred to local demand points serviced by yard 4. The train [1,3] however may have cars destined for yard 3 and 4. If any cars with destination 4 are placed on that train they must be classified at yard 3. We also note that the total number of cars destined for a yard (to stay) is a constant determined

by traffic requirements and so should not affect the optimization. This can be seen formally from the equation

$$\sum_{i \in I_j} \sum_{(p,j) \in T_{ij}} x_{ij}^{pj} = \sum_{p} r^{pj} = \text{constant} \qquad (8)$$

where the last summation is over all origins p that have traffic destined for yard j (i.e. all p to which node j is accessible). Let us define a variable θ_{ij} corresponding to the classification status of train [i,j] as follows:

$$\theta_{ij} = \begin{cases} 1 & \text{if train } [i,j] \text{ should be classified at yard } j \\ 0 & \text{otherwise} \end{cases}$$

Moreover let $<$ be a partial ordering on the set of nodes N of our network where $i < j$ means is accessible to i, that is to say there is a directed path from i to j in G. (For the single track network this coincides with the usual ordering on integers). Then we see

$$\theta_{ij} = \delta\Big(\sum_{\substack{(p,q) \in T_{ij} \\ j < q}} x_{ij}^{pq}\Big) \qquad (9)$$

where $\delta(x)$ is the delta function,
that equals 0 for $x = 0$ and 1 for $x > 1$, commonly used in fixed charge problems. Equation (9) simply says that the train [i,j] should be classified at yard j whenever it includes any traffic travelling beyond yard j. Then the delay at yard j for processing trains [i,j] is

$$\tau_{ij}^{d} = (W_j + v_j x_{ij}) \theta_{ij} \qquad (10)$$

We could translate this delay term to money costs by attaching a time cost

d_{ij} to each unit of time delay. This should reflect in dollar terms the undesirability of delaying cars at yard j. The yard cost function may then be written as:

$$\Phi_j(\bar{X}_j) = \sum_{i \in I_j} d_{ij}\,\theta_{ij}(W_j + v_j\,x_{ij}) \qquad (11)$$

We wish, however, to point out a complication in transferring delay times into cost terms: The devaluation costs d_{ij} should logically depend on the train composition. As in inventory holding costs, d_{ij} usually reflects the total dollar value of freight which undergoes delay. If the average devaluation rate of traffic corresponding to OD pair (p,q) is c_d^{pq} dollars/car/day, then the train devaluation rate would be

$$d_{ij} = \sum_{(p,q) \in T_{ij}} c_d^{pq} \cdot x_{ij}^{pq} \qquad (12)$$

Even if all the rates c_d^{pq} are equal (say to c_d), d_{ij} still depends on the train volume x_{ij}, leading to a quadratic cost function for each train. Consequently measuring delay more realistically in terms of car-days, rather than just days, leads to a more complicated objective function.

We have now presented some possible cost functions for the general model(4). In later sections we shall pursue further simplifications in the cost function with a view twoards algorithmic issues.

4. CAPABILITIES AND LIMITATIONS OF THE NETWORK MODEL

In this section we shall relate the general model presented above to the tactical issues listed in Section 2. Rail systems are distinguished from many other transportation modes by their inherent ability to move a large

number of shipments as a single unit. We feel an adequate model of rail freight transportation should take explicit account of such economies associated with train length and the corresponding blocking policy of consolidating diverse classes of traffic into large groups in which members share some leg of their itinerary. The train length economies operate in two directions: On the one hand they set lower limits on the total traffic assigned to a given train so that it would be profitable to run. On the other, they allow us some flexibility of "stretching" the train capacity by assigning additional engines so that the train could accomodate more traffic. This latter feature is absent from the competing trucking mode, for example.

On the negative side of the economies suggested by large shipments may well be negated by the necessary intermediate sorting and grouping operations on the cars which tend to increase operating costs delays while lowering customer service and equipment utilization. Our model should address both the economies and diseconomies mentioned above.

When economies of scale are present, it is imperative that a model aiming to incorporate them be on the same hierarchical level as where such economies are realized. This consideration will serve as a guide for choosing a suitable level of aggregation and planning horizon for the model. To clarify this issue consider a strategic model for route selection on a rail network. Suppose we measure our flow variables x_{ij}^{pq} in units of cars/year and correspondingly think of our requirements r^{pq} as average annual requirements. Moreover, let us choose aggregate cost functions $\psi_{ij}(x_{ij})$ for the routes [i,j] and $\phi_j(u_j)$ for the yards j where u_j is the total throughput of yard j as described in (6). Then the model in (4) will attempt to find an optimal routing of traffic while accounting for link and yard congestion. (We assume the functions

ψ_{ij} and ϕ_j to be increasing and convex). With this choice of variables, y_{ij} will have a large value (say for a train per day over the route [i,j], y_{ij} will be about 360 trains/year). Thus we may relax integrality on y_{ij} with no substantial loss of optimality using the relation

$$y_{ij}^{cont} = \alpha_{ij}^{-1} x_{ij}$$

the constraints (2) may be dropped from the model and the program (P) becomes a convex network minimization problem. While such a model is of interest for aggregate routing purposes and for facilities location (for example the location or expansion of classification yards), it abstracts away from the blocking problem of assigning blocks of traffic to trains. We shall give an example of economies that such a model will be too aggregate to reflect.

Example Consider three nodes (yards) i,j,and k of a rail network with requirements as shown in Figure 4:

20 cars from i to j

100 cars from i to k

100 cars from k to j

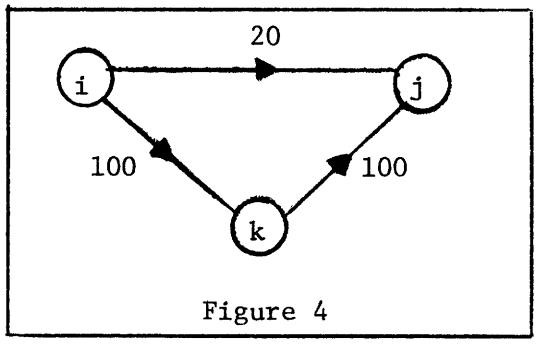


Figure 4

Suppose moreover that we have a car/engine ratio of 40 for trains travelling over the routes [i,j], [i,k], and [k,j]. Sending cars directly from each origin to destination requires a total of 7 engines: one over the route (i,j) and 3 over each of the routes (i,k) and (k,j). However, diverting the traffic from i to j through yard k changes the loads on links (i,k) and (k,j) from 100 to 120. Then one engine is saved over the route (i,j) and, moreover,

the other engines will be fully utilized (to capacity). There is of course a corresponding processing and classification cost that cars of (i,j) will incur at yard k, but the balance may well be in favor of traffic diversion.

The above example shows that the integrality issue in the provision of motive power units is a key factor in two questions that have long plagued rail operations: that of train length (short versus long trains) and of high equipment utilization. Maintaining the discreteness is our cost function allows for a more rational evaluation of the economies of shipping in train-loads. We see the economies are realized on the level of loading (and blocking) decisions for cars.

We now list some of the issues that a successful solution of the routing/makeup model of Section 3 will clarify.

a) Service Selection and Frequency - The variables y_{ij} determine the choice of trains to provide service. If $y_{ij} = 0$ no trains will be sent along the route [i,j]. For $y_{ij} = 1$ only one train will be run. For larger values of y_{ij} longer (or more frequent) trains will travel on the service [i,j].

b) Train Makeup - The optimal values of the flow variables x_{ij}^{pq} will specify the optimal train composition on a given route [i,j]. In our model these variables will also provide guidelines for the blocking policy.

c) Yard Grouping Policy - At each yard the incoming traffic may be reclassified according to outbound destinations. The number of groups formed at the yard can be derived from the total number of outbound services operative at optimality. We may think of each classification track as containing all the traffic belonging to the takelist of a given outbound train.

d) Yard Workload - The total workload of each classification yard may be obtained from the total yard delay as expressed by the terms τ_{ij}^{d} in equation (10). The yard parameters W_j and v_j should be calibrated for each yard and will differ from one yard to another according to yard type (flat or hump yard) and the technological profile of the yard. The model strives for an efficient allocation of classification work among yards. For example, classification activity may be diverted from a small flat yard to a large automatic hump yard where it may be performed much more quickly.

e) Train Length - The train length variable x_{ij} will be influenced by the discrete marginal costs c_{ij}^{e} attached to y_{ij}. The discreteness of these variables will specify a number of "regimes" of possible train lengths with different costs to the railroad.

We thus see that the routing/makeup model, if solved, will address a number of the tactical issues raised in Section 2 and consequently has the potential of being a valuable planning tool.

We shall close this section with some technical observations concerning our model for routing and makeup: As it stands, the model involves continuous flow variables x_{ij}^{pq}, integer variables y_{ij}, and rather complicated cost terms in the objective function as specified by equation (11). Let us start by noting how this formulation may be transformed to a network flow model with arc costs involving set-ups. In the process we shall also eliminate the 0-1 variables θ_{ij} which specify the classification status of the train[i,j].

Consider a given yard j. A train [i,j] will be classified at j if it carries

freight destined for yards beyond yard j i.e. for some nodes k >j). On the other hand, a train composed entirely of traffic due for yard j will not require reclassification at that yard. This situation can be partially captured by splitting the node for yard j into two nodes j'and j" as shown in Figure 5.

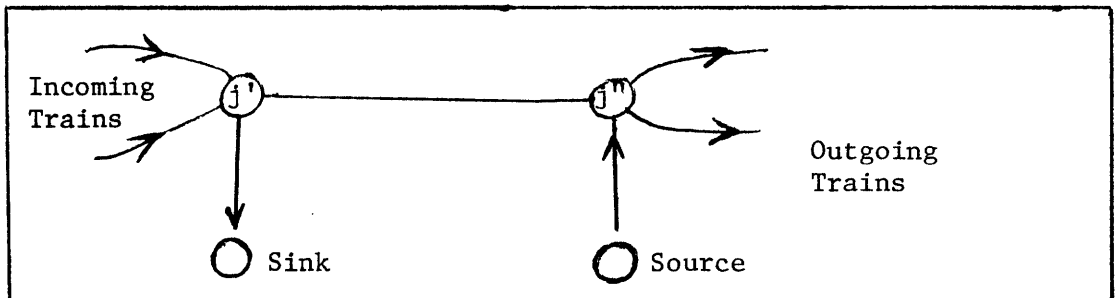


Figure 5 Network Representation of Processing Delay

The node j' acts as a sink for all traffic staying at yard j (all traffic classes with final destination j) while node j" acts as a source of outbound traffic from yard j. We may decompose the traffic volume of a given train [i,j] as follows

$$x_{ij} = \sum_{\substack{p\le i \\ j\le q}} x_{ij}^{pq} = \sum_{p<i} x_{ij}^{pj} + \sum_{\substack{j<q \\ p<i}} x_{ij}^{pq}$$

The first summation (call it u^j) is the traffic staying at yard j that will exit the network directly through the sink at j'. The second term (call it u^{j+}) passes over the connecting intra-yard arc (j',j"). It is then possible to attach the classification and processing costs of yard j to this arc (j',j") by specifying an arc cost function of the form

$$\phi_{j'j''}(u) = \begin{cases} 0 \text{ for } u=0 \\ W_j + v_j.u \text{ for } u>0. \end{cases}$$

Then pure trains, in the sense of trains with $u^{j+} = 0$, will incur no processing delays at yard j. If many different trains [i,j] flow into yard j it is possible to extend this approach by splitting node j into many nodes (equal to the maximum number of trains flowing into the yard) all of which are connected to a sink j' and an outbound node j". This will naturally increase the size of the network considerably if many trains are considered.

An alternative to capturing the cost functions of equations (10)-(12) exactly is to seek simpler cost functions. One such choice is

$$\phi j(y) = W_j \cdot (\sum_{i \in I_j} y_{ij})$$

which captures some measure of both train length and frequency for all trains incoming to yard j. Forms of different cost functions should be considered in more detail in further research. It is important to note, however, that ideally, processing times for different trains coming into a yard should be calculated separately. In particular the fixed cost W_j refers to a single train from a given origin i thus the correct delay term due to fixed setup times at a yard j is

$$[\sum_{i \in I_j} \delta(y_{ij})]W_j \cdot$$

Finally we wish to note that our general model can be considerably enriched by the addition of rather simple constraints. Maximum train lengths (or frequencies according to the interpretation of y_{ij}'s) could be enforced by adding the constraints

$$y_{ij} \leq \bar{y}_{ij}$$

for a given upper bound $\bar{y}_{ij}$ on a route [i,j]. Such constraints could arise, for example, from capacity restrictions on the makeup end of yard i due to limitations of departure tracks or crew availability.

We may also take the allocation of motive power into account more explicitly by specifying engine availabilities at each yard at the beginning of the planning horizon and imposing flow constraints on the variables y_{ij}. In this way in addition to the flow of traffic classes (commodities), we would also have engine flows. Alternatively we may let the model decide the required number of engines at each yard by supplying the costs of providing an engine at yard i. Naturally the modelling of engine flows should reflect the railroad's primary concerns: If the fleetsize is limited and likely to act as a bottleneck factor explicit constraints on the total number of engines used may be added. At any rate, we feel that the engine allocation problem should be approached from an aggregate point of view at this level. Thus we do not wish to incorporate detailed scheduling on spacetime network. The work of Florian et. al. [9], however, has addressed the detailed engine scheduling problem algorithmically with highly encouraging results.

5. THE COMBINATORIAL-SEARCH MODEL OF THOMET: REVIEW AND COMPARISON

In this section we will review two major attempts in providing an optimization model of rail systems and compare them with our model.

5.1 The Combinatorial-Search model of Thomet

The work of Thomet [15,16] has stimulated our formulation and deserves some elaboration. He has developed a heuristic method for optimizing a model which accounts for both routing and classification costs with cost functions essentially as given in equations (5) and (11). We will only give a simplified summary of his work to elucidate the nature of his solution technique.

The basic component of Thomet's algorithm is a cancellation procedure. For a given pair of nodes (i,j) consider a direct train travelling from i to j and an alternative sequence of direct trains travelling over the same physical route from i to j, but with stops at intermediate yards. A typical case is shown in Figure 5. The dotted line represents the direct train and [k,l] is a typical intermediate train on the route from i to j.

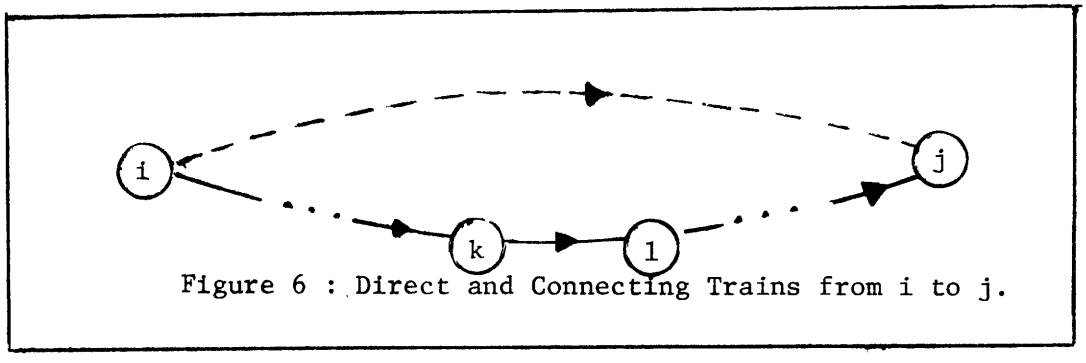


Figure 6 : Direct and Connecting Trains from i to j.

suppose train [i,j] is cancelled and all traffic previously assigned to it is shifted onto a sequence of intermediate trains such as [k,l]. We wish to evaluate the impact of this traffic diversion on total costs. The benefits

resulting from the cancellation include a reduction in train-miles (to use one aggregate measure of variable routing costs) and a possible decrease in motive power requirements and other costs of running a train from i to j. The costs involve possible addition of motive power over the routes of some intermediate trains [k,l] and, more importantly, classification and processing costs at intermediate yards. An intermediate train [k,l] has to be regrouped at yard 1 and all traffic going beyond yard 1 (say to j) should be classified. Such shifting of cars from [i,j] onto [k,l] may thus change the classification status 0_{kl} of the train [k,l] and incur delay costs as in (11). By adding these costs and benefits we may find a quantity

$$\gamma_{kl}^{ij} = \text{net costs of diverting traffic from train [i,j] to train [k,l].}$$

We may ask what is the best sequence of intermediate trains onto which the traffic of train [i,j] may be shifted. We will show by means of small example that this best sequence (best in the sense of providing the maximum savings in the cancellation of [i,j] may be found by solving a shortest path problem. Note that the total cost of diverting traffic from [i,j] is the sum of γ_{kl}^{ij} 's for all trains [k,l] which now carry the traffic of train [i,j].

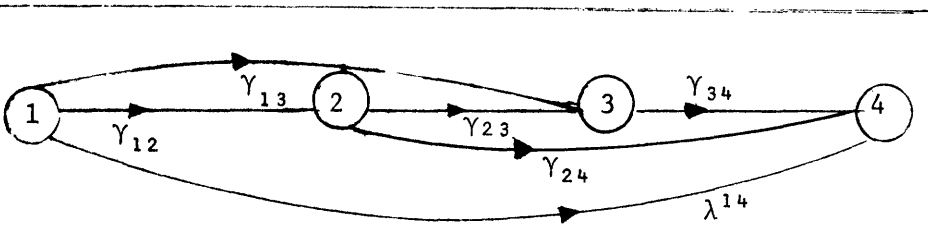


Figure 7 : Cancellation Costs for Train [1,4]

Example Let us consider the train route graph of Figure 2. Suppose a direct train is scheduled between yards 1 and 4 that bypasses yard 2 and 3. Consider cancelling this train, that is, [i,j] = [1,4] in this case. The intermediate trains between yards 1 and 4 are [1,2] , [2,3] , [3,4] , [1,3], and [2,4]. Thus one shifting pattern is [1,3] , [3,4], that is we divert the load of train [1,4] onto train [1,3] and classify this latter train's cars at yard 3 to transfer the added cars to train [3,4]. Thus essentially the cars of [i,j] make a "connection" at yard 3. This shifting scheme involves two intermediate trains. A sequence which uses 3 trains is [1,2] , [2,3] , [3,4]. What is the best sequence to use?

Suppose we have calculated the costs γ_{k1}^{14} of diverting from [i,j] onto [k,1] for all intermediate trains [k,1] ($1 \leq k < 1 \leq 4$). Let us attach this value as a cost to the arc (k,1) on the graph. Figure 7 shows these arc costs with the superscript 1,4 omitted. To arc (1,4) we attach the cost λ^{14} which is simply the cost of running the direct train [1,4].

Suppose now we find the shortest path from 1 to 4 with respect to these costs. The links on the shortest route will specify an optimal "decomposition" of the train [1,4] into intermediate trains. Suppose, for example, that the shortest route is (1,2) , (2,4). Then train [1,4] will be cancelled and its traffic will be placed on trains [1,2] and [2,4] consecutively. If we let the length of this shortest path be $\bar{\gamma}^{14} = \gamma_{12}^{14} + \gamma_{24}^{14}$, then the savings realized by cancelling train [1,4] is $\delta^{14} = \lambda^{14} - \bar{\gamma}^{14}$. Note if the shortest route were given by the link (1,4) then we would save nothing by cancelling [1,4].

In general, for a train [i,j] we consider its physical route through the network G_p given by the sequence of yards $i_1=i$, i_2, $i_3,\ldots,i_m = j$.

Usually this route is the shortest distance route from i to j on G_p. We assume this route goes through at least one intermediate yard so that cancellation is possible ($m \geq 3$). Then we

(i) construct another network G(i,j) on the node set $\{i_1,\ldots,i_m\}$ with arcs of the form $(k,l) = (i_r,i_s)$ for $1 \leq r < s \leq m$,

(ii) compute the costs γ_{kl}^{ij} for all such arcs, and

(iii) find $\bar{\gamma}^{ij}$ - the length of the shortest route from i to j on G(i,j). If λ^{ij} is the cost of running [i,j] directly the savings due to cancellation are defined as

$$s^{ij} = \lambda^{ij} - \bar{\gamma}^{ij} .$$

Thomet's algorithm may now be described as follows: Initiate the algorithm by assigning direct trains to all OD pairs with nonzero requirements (all(p,q) with $r^{pq} > 0$). This is called the Minimum Transit Time Policy (MTT) and involves no classification work (initially all $\theta_{pq} = 0$). At a given iteration of the cancellation step compute the savings s^{ij} of all trains [i,j] as described above and cancel the train with maximum savings. Transfer the traffic of this train to the intermediate trains and update train parameters accordingly. To find the minimum cost policy this strategy is continued (one cancellation at a time) until no positive savings can be found - that is, until $s^{ij} < 0$ for all remaining trains [i,j].

If we view running trains between certain yard pairs as opening (or locating) facilities, and OD demands as demand sites; Thomet's approach may be likened to drop heuristics in facilities location problems [7]. One starts with all trains and successively cancels trains, one at a time, which

yield a savings according to a myopic criterion of change in total costs.

We may immediately point out several limitations of Thomet's approach:

a) As Thomet himself realizes this sequential cancellation procedure does not guarentee optimality upon its termination. Rather it provides a local optimum in the sense that there will be no "one-move"(cancellation of a single train) which could further reduce costs.

b) In the algorithm described above the physical routing of traffic is never changed. All traffic follows the shortest route between its origin and destination on the physical graph G_p. What changes is the loading pattern of the traffic on the intermediate trains on the shortest route itinerary. However as the small example of Figure 4 indicates, it may be advantageous to divert traffic away from its shortest route. In practice rail freight is known to be routed through such circuitous paths, occasionally on the basis of a reduction in the railroad operating costs. Diversion of traffic from the shortest route will become necessary if flow on certain routes is limited by capacity constraints. This limitation of routing alternatives to the shortest path may result in further suboptimality in Thomet's model.

c) The diversion of traffic from one train to another is always performed in bulk form, that is, the entire traffic load of a train is shifted. One may easily envisage cases where only partial diversion of traffic if necessary. For example, consider the situation of Figure 4 with the demand from i to j changed to 60 from 20. Then we may retain 40 cars on the train [i,j] and only divert the additional 20 to the route going through yard k. This possibility is allowed by our routing/makeup model

(P), but not by Thomet. We shall dwell on the point raised in (c) above slightly further: At any given point of Thomet's algorithm the value of a typical variable x_{ij}^{pq} is limited to two choices - 0 and r^{pq}. Indeed we may recast our planning model into Thomet's form by the following device. Let z_{ij}^{pq} be a new variable defined as $z_{ij}^{pq} = x_{ij}^{pq}/r^{pq}$. If we then substitute the quantity $r^{pq}\ z_{ij}^{pq}$ for x_{ij}^{pq} in equations (1) - (4) of (P) we realize that for each (p,q) the variables z_{ij}^{pq} obey flow conservation equations for a flow value of 1. Restricting the variables to be integral will now mimic Thomet's procedure. Note that we will have

$$z_{ij}^{pq} = \begin{cases} 1 & \text{if all } r^{pq} \text{ cars of (p,q) demand is on train [i,j]} \\ 0 & \text{otherwise.} \end{cases}$$

5.2 The Railcar Network Model

An important example of optimizing rail models is the work developed at Queen's University at Kingston and the Canadian Institute of Ground Transport [13]. This study has evolved over the past five years into a comprehensive model of over-the-road and yard activities which has been validated against real data. As a result, it has claims to being the most comprehensive rail planning model based on optimization in the existing literature.

The object to the model is to route freight on the rail network to meet demand at minimal total delay (in car-hours). The approach is to derive delay functions for component rail operations as a function of the flow of freight handled by each operation. If all the delay functions are convex, the routing problems may be viewed as a minimum cost network problem with

convex costs for which a number of algorithms are available (see Sections V and VI of [3]). Algorithmically, the model is thus identical to the work in traffic assignment [8] where the delay functions (also called service functions in road traffic assignment literature) reflect over-the-road congestion delays. The derivation of the delay functions, however, is very different as it is based on average waiting times derived from queueing theory.

The Railcar Network model minimizes an expression for total delay comprised of the following components:

(i) Inbound Inspection time (W_I): The inspection work is performed on incoming trains in the receiving end of the yard.

(ii) Classification time (W_c): This involves the queueing delay before the yard's sorting facilities.

(iii) Train Assembly Time (W_A): The waiting time of cars to be assembled into a train for departure (sometimes called accumulation delay).

(iv) Outbound Inspection Time (W_D): This inspection is carried out before the departure of an outbound train.

(v) Over-The-Road Time (W_o): This reflects the required time for traversing a physical a physical link with congestion effects of meets and passes incorporated.

We note that factors 1-4 refer to yard activities and only 5 represents the main-haul delay. Inspection times W_I and W_D are taken to be constant (per train) while W_o is derived from an analytical expression incorporating train interference effects of limited track capacity. Finally queueing delays W_c

and W_A are derived from usual waiting time formulas in queueing theory for a variety of arrival/service characteristics.

To cast the results into the form of a network minimization problem, we may assign total delay functions to two types of arcs in the network as shown schematically in Figure 8. Yard i is represented by two nodes corresponding to yard entry and exit from the sorting facilities. The arc joining these two nodes that carry the total flow through yard i ($X_i = \sum_{h \in I_i} x_{hi}$) is assigned the delay function $W_I + W_c(X_i)$ (as a function of the flow X_i on the arc). Similarly the flow x_{ij} of cars from yard i to yard j is assigned the delay $W_A(x_{ij}) + W_D + W_o(x_{ij}, x_{ji})$. The resulting nonlinear network minimization algorithm (assumed to have convex arc costs) is solved by a primal-dual algorithm based on linearizing the above cost functions successively. Other node splitting techniques are used to represent bypass trains and yards where some traffic is set off or picked up, but no classification has taken place.

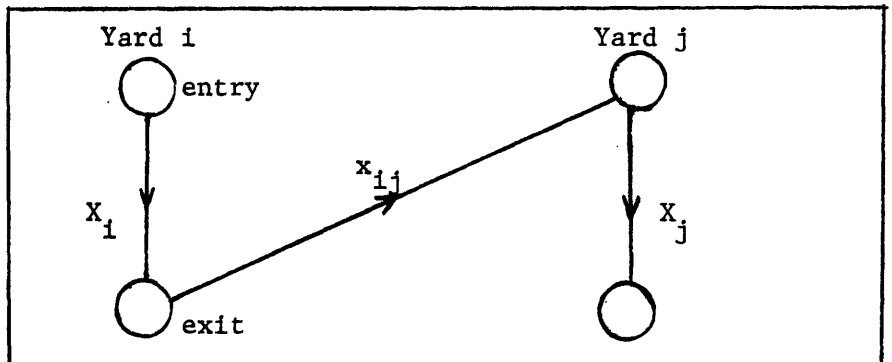


Figure 8 : Network Representation of Delays

As mentioned before, we regard the Railcar model as a major contribution. The close contact of its developers with Canadian railroads and their efforts in the way of model validation make this model a valuable planning tool. We wish, however to bring out some basic differences between their work and our approach in this report:

Our model makes a basic distinction between the flow of cars and the flow of trains on a given service [i,j]. The effect of running a train is determined by integer variables y_{ij}, in contradistinction to the continuous flow variables x_{ij} for the cars. The Railcar model, instead, derives train flows from car flows by using an average train length of $\bar{\ell}$ cars/train. Thus a flow of x_{ij} cars over a route automatically results in $x_{ij}/\bar{\ell}$ trains. Moreover the number of trains affects only the delay time, no attention is paid to the cost of scheduling an additional train. In Section 4, we discussed how these costs may directly influence routing and makeup decisions, possibly resulting in train cancellations if such costs are high. Moreover, the variables y_{ij} shed some light on the issue of optimal train lengths which the Railcar model takes to be given. Our model attempts to capture some of the economies of the routing process more directly to allow us to decide, for example, whether or not a certain train should run or a particular traffic class should be incorporated into a train's makeup. The same distinction arises with respect to classification costs. A component of our classification costs refers to the train - via the fixed charge term $W_j\theta_{ij}$ in equation (10) - and not to the traffic throughput. In fact, this term attempts a first approximation to the effect of train composition on yard costs. Naturally knowing the detailed composition of any train incoming to a yard should

enable us to estimate the extent of the classification work it requires. In a deterministic routing model, this composition is completely specified and so, ideally, we should not have to use average waiting time formulas. On the other hand, handling detailed data on train composition would complicate the model beyond all hope. While the Railcar model chooses a purely stochastic approach to evaluating waiting times, we have have opted for using a delay term for which the effect of train cancellation or composition is more apparent. Finally we regard some of the delay terms in the Railcar model as being highly dynamic in nature and thus not particularly suitable for static waiting time analysis. In particular, the assembly delay term W_D will depend on the yard dispatching policy. Such policies, which belong to the operational level, should be studied separately incorporating the time element into direct account.

To conclude this section we wish to situate the models described above in the hierarchical framework of Section 2.

We believe the Railcar model to relate to decision-making on the strategic level. This model can specify the flow of traffic on the network. At this level, one must use highly aggregate measures of yard and over-the-road delays to allow for their impact on the traffic flow configuration without going into much detail. Our model, however, belongs to the tactical level: We deal with issues of traffic routing and train scheduling more directly. This reflects our desire to have firmer control on the question of trains to run and their composition. Thomet's model shows the same concern. Indeed one may be well-advised to use the Railcar model and our model sequentially. The former will provide an aggregate picture of traffic flows and supply a 'base-level' of yard and line activities. Thereupon

we may pass onto the routing/makeup model to evaluate certain more detailed decisions (for example, the use of direct versus local trains, and long versus short trains).

VI. SPECIALIZATIONS OF THE GENERAL MODEL AND CONCLUSIONS

In this section we shall explore some specializations of our general model and describe possible scenarios where such specialized versions may be of interest. Our main guide in specialization is to obtain problems which are algorithmically tractable. Thus this section may also be viewed as a preliminary discussion of the algorithmic issues involved in solving the routing/makeup model.

Let us recall the basic cost components of the general model. We considered over-the-road costs in terms of dicrete variables y_{ij} which specify how many trains should run on a given route or service [i,j]. We also had yard costs which may involve fixed charge terms as well as a convex function of traffic throughput of the yard. Note also that once the y_{ij} variables are set, the set of available trains is known and the problem reduces to an assignment of traffic classes (commodities) to the trains. In general this will be a capacitated multicommodity flow problem. The state of the art for solving such problems is reviewed in [3]. Let us now consider some specializations of this model in increasing order of complexity.

a) A Traffic Assignment Problem

Suppose we are given a set of operative services (trains) and that we regard each train route as having unlimited capacity. This assumption of unlimited capacity is justified in a scenario where a sufficient number of trains operate on a given service with a fixed frequency so that we may assign as much traffic as we wish to that route. Mathematically this corresponds to a large value of α_{ij} in equation (2) so that the constraint

would no longer be binding. In this case only traffic routing and yard costs will matter. Suppose we use convex cost functions to describe these costs in terms of link flow and node throughput. The result will be a traffic assignment problem with convex congestion costs: we have to route the traffic over the available routes on the network (operative services) in such a way as to minimize total costs. Note that while the solution technique will be similar to algorithms for the traffic equilibrium problem, our network is in terms of feasible routes and not physical arcs.

Let us now take the simplest possible cost functions i.e. functions linear in chain flow and node throughput (i.e. $v_j=0$ in (7)). Then in terms of yard delay, for example, each traffic class passing through a node j will suffer a constant delay W_j. The resulting problem may be solved by a shortest path method which will specify the optimal sequence of train connections for each traffic class much in the way described in Section 5. While this observation is a simple one, it is still of importance in modelling. Indeed the model proposed by Truskolaski [17] seems to have features very similar to this simplified case, however he does not appear to use the shortest path method.

b) A Combined Service Scheduling and Traffic Assignment Problem

Let us now maintain the assumption of unlimited capacity but attach costs to providing regular service on a given route [i,j]. Once again this corresponds to running trains sufficiently frequently on a given route. In this case, we may restrict variables y_{ij} to be 0,1. Thus let

$$y_{ij} = \begin{cases} 1 & \text{if service [i,j] is operative} \\ 0 & \text{otherwise.} \end{cases}$$

and let α_{ij} be a large number in equation (2). The result, will of course be an uncapacitated network design problem. Once a choice of the y_{ij} variables

is made - that is a particular route configuration is set - the problem reduces to that of part (a) with traffic equilibrium or shortest route subproblems. The objective function involves the sum of routing costs and the costs of scheduling operating services. Alternatively we may only retain the routing costs in the objective function and incorporate constraints restricting our choice of services. One such constraint, frequently called a budget constraint reads

$$\sum_{(i,j)\in A} d_{ij}\, y_{ij} \leq D$$

where d_{ij} equals the cost of providing service on route [i,j] and D is the total available budget. A simpler constraint simply limits the total number of services we may choose to operate and may be written as

$$\sum_{(i,j)\in A} y_{ij} \leq p .$$

This problem is similar to the p-median problem. This last constraint will be important when the total number of services is limited due to crew or motive power limitation. For recent algorithmic work on the network design problem, we refer the reader to the papers [9] and [12].

c. The Routing/Makeup Model

This model is already discussed in Section 3 and 4. Rather than considering the general model again, let us focus on the case where each traffic class undergoes a constant delay W_j if it visits yard j - that is to say if it is loaded on a train with destination j. In that case we obtain a simple version of the routing problem where we are interested in minimizing yard delay for traffic while taking account of the additional

costs of sending traffic direct (in order to bypass intermediate yard stops). Even this simplified version will not be easy to solve. For a given assignment of y_{ij}, variables, we obtain capacitated multicommodity flow subproblems. We propose this last model for more careful algorithmic consideration. This model can deal with a variety of "loading problems" in transportation studies which also aim to determine optimal itineraries (stop-schedules) for the carriers. As a result, an efficient algorithm for this problem will be a valuable contribution. Richardson [14] has used Benders Decomposition on an airport routing problem which shares some features with our model. His work may serve as a useful point of departure.

Since problems of this type will be of a large scale in any realistic study (a small railroad might involve 20-40 yards and save 200-300 possible train routes) it is also important to evaluate the efficiency of heuristic solution techniques. In our review of Thomet's work, we have already seen one class of heuristics for this problem. Naturally other heuristic strategies may be proposed and should be duly assessed. The theoretical evaluation of heuristics may profitably pursue the approach of [5].

We wish to conclude this report with a few words in the way of recapitulation and some indications for future research:

We approached planning for rail system by providing a hierarchical view of the decision-making process. We then concentrated on a mathematical programming model on the tactical level which we couched in a fairly general form. We described how this model can address a number of issues in train routing and makeup. Subsequently we gave specifications of the general model which may be profitably studied from the algorithmic point of view.

In particular, we derived a network flow problem with integer variables reflecting trainload economies. Solving this model for a large-scale system may pose serious computational difficulties. We feel that further work should concentrate on algorithmic studies of the model on two levels: First, we may look at various decomposition techniques which would render the solution of the general routing/makeup model more tractable. Here, solution techniques based on Benders Decomposition and Lagrangian Relaxation come to mind. Second, we may pursue certain simplified forms of the general problem and attempt to find efficient algorithms for such subproblems. For example, the routing/makeup model may be studied on a line network with simplifying assumptions on yard delays and routing costs. The qualitative insights provided by these simpler problems may serve as basis for developing efficient heuristics for the general problem. Progress along either one of tese dimensions will substantially increase the promise of optimization-based models in aiding the planning process for rail systems.

Acknowledgements

The author wishes to thank Professors R.W. Simpson and T.L. Magnanti for careful readings of the manuscript.

REFERENCES

[1] Allman, "A Network Simulation Approach to the Railroad Freight Train Scheduling and Car Sorting Problem", unpublished, Ph.D dissertation, Northeastern University (1966).

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

[3] Assad, A.A., "Multi-commodity Network Flows - A Survey," to appear in Networks.

[4] Assad, A.A., "Analytical Models in Rail Transportation", Working Paper OR 066-77, Operations Research Center, MIT (Oct.1977).

[5] Cornuejols,G., M. Fischer, and G.L. Nemhauser, "Location of Bank Accounts to Optimize Float: An Analytical Study of Exact and Approximate Algorithms", Management Science, 23, 8, 789-810. (1977).

[6] Dionne, R., and M. Florian, "Exact and Approximate Algorithms for Optimal Network Network Design", Publication 41, Centre de Recherche sur les Transports, University of Montreal (February 1977).

[7] Feldman,E.F., A. Lehrer, and T.L. Ray, "Warehouse Location under Continuous Economies of Scale", Management Science, 2, 9, 670-684,(1966).

[8] Florian, M. (editor), Traffic Equilibrium Methods, Volume 118, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, New York(1976).

[9] Florian, M. et.al.,"The Engine Scheduling Problem in a Railway Network", INFOR, 14, 2, 121-138, (June 1976).

[10] Hax, A.C. and R.J. Armstrong, "A Hierarchical Approach for a Naval Tender Job Shop Design", Technical Report No. 101, Operations Research Center, MIT (August, 1974).

[11] Leblanc, L.J., "Global Solutions for a Nonconvex Nonconcave Rail Network Model", Management Science, 23, 2, 131-139, (October,1976).

[12] Magnanti, T.L., and R.T. Wong, "Accelerating Benders Decomposition for Network Design", Discussion Paper, Center for Operations Research and Econometrics, Cath. Univ. de Louvain, Belgium (1977).

[13] Petersen, E.R., and H.V. Fullerton (editors), "The Railcar Network Model", Canadian Institute of Guided Ground Transport, Queen's Universtiy at Kingston, Ontario, CIGGT Report Number 75-11 (June 1975).

[14] Richardson, R., "An Optimization Approach to Routing Aircraft", Transportation Science, 10, 1, 52-71, (February, 1976).

[15] Thomet,M.A., 'A Combinatorial-Search Approach to the Freight Scheduling Problem', unpublished Ph.D dissertation, Department of Electrical Engineering, Carnegie-Mellon University, (August, 1971).

[16] Thomet, M.A., 'A User-Oriented Freight Railroad Operating Policy', IEEE Trans. On Systems, Man. & Cybernetics, SMC-1,4, 349-356, (Oct.1971).

[17] Truskolaski, A., "Application of Digital Computer to the Optimization of Freight Train Formulation Plans", Bulletin of the International Railway Congress Association, Cybernetics and Electronics on the Railways, 5, 5, 211-226, (May, 1968).

[18] Studies in Railroad Operations and Economics, Volumes 1-5, Department of Civil Engineering, MIT, (June, 1972).