Queuing Models of Classification and Connection Delay in Railyards

MARK A. TURNQUIST

Cornell University, Ithaca, New York

MARK S. DASKIN

Northwestern University, Evanston, Illinois

The major components of delay to rail cars in passing through yards are waiting for classification and connection to an appropriate outbound train. This paper proposes queuing models for each of these components which provide expressions for both the mean and variance of delay times. The models are then used in an example application to draw inferences regarding the effectiveness of alternative strategies for dispatching trains between yards.

1. INTRODUCTION

Due to the nature of railroad operations, rail cars spend a great deal of their time in classification yards. According to data gathered by REEBIE ASSOCIATES[17] the average loaded rail car requires 8.8 days to move from its loading point to its destination. Of this time, 6.8 days (77%) are spent in yards. Thus, finding ways to reduce average yard time is an important means to improve the quality of service railroads are able to provide.

In addition, classification yards are the major source of schedule reliability problems in rail car movements.[6,7,18,20] The importance of trip time reliability as an attribute of rail service quality implies that we must also be concerned with the variability in yard delays, as well as with average delays.

An understanding of classification yard operations is also important to a thorough analysis of railroad costs. Such a study, in fact, provided the motivation for the work reported here. The overall project is described elsewhere[5]; this paper concentrates specifically on models of delays encountered by rail cars in passing through classification yards.

Transportation Science
Vol. 16, No. 2, May 1982

0041-1655/82/1602-0207 $01.25

While in a railyard, a car undergoes four basic operations: (1) inbound inspection, (2) classification, (3) assembly into an outbound train, and (4) outbound inspection. It is quite natural to consider these operations as a series of queues through which the rail car passes. Significant prior work on the application of queuing models to railroad yards has been done by HEIN,[11] SIDDIQEE[19] and PETERSEN.[15,16] Certainly the most comprehensive of these studies is the work by Petersen. His efforts served as a major starting point for the work described here.

The models described below view yard operations from the perspective of individual cars, rather than from the perspective of trains. This allows a significant improvement in the ability of the models to analyze policy options. We begin by describing briefly the sequence of operations performed on a rail car as it passes through a classification yard. For more detailed discussions, the interested reader is referred to FOLK[6] or BECKMAN ET AL.[2]

Inbound trains pull into a receiving yard area, where road locomotives and the caboose are taken off the train and the cars are inspected. The essential elements of a queuing process are certainly evident here. However, as Petersen[15] has pointed out, this inspection operation is not a major bottleneck because additional inspectors can often be assigned quite easily. Also, this inspection can generally be carried out while the inbound cars are, for practical purposes, in queue waiting for classification. Hence, this inspection step is not a focus of our modeling effort.

Once inspected, the cars are ready to be classified. Classification yards fall into two general categories: hump (or gravity) yards and flat yards. In a hump yard, a switch engine gets behind the string of cars, and pushes the cars, at a speed of about 2 mph, over a hump elevated 10 to 30 feet above the level of the classification tracks. As a car rolls down the hump, switches are set (either automatically or manually) to direct the car onto its proper classification track. In a flat yard, the classifying is done by a switch engine pushing and pulling cars into and out of the set of classification tracks, rather than letting them roll down a hill onto the right track. Waiting to be classified is a major source of delay to rail cars passing through a yard,[7,18] and thus classification delay is one main focus of attention in this study.

When the rail cars have arrived on the classification tracks, they are ready to be assembled into outbound trains. To assemble an outbound train a yard engine comes into the classification tracks, picks up the cars on one or more tracks, and assembles these strings of cars, in a specified order, on the outbound departure tracks. The nature of train assembly as a queuing process is complicated by the combined effects of switch engine availability and the schedule of outbound trains. In many cases, the principal source of delay to cars following classification is not the limited

availability of switch engines to assemble trains, but the schedule of outbound trains on which the cars are to depart.[7,18] In such cases, the "service process" to be modeled is the schedule of departures and not the physical assembly of outbound trains.

It should be noted of course that changes in the outbound schedule may be constrained by the availability of yard engines to make up outbound trains. Thus, the emphasis on the schedule does not imply that the potential effects of limited switch engine availability can be ignored completely. However, since the principal source of delay is typically the schedule, our analysis will focus there, and we will refer to this as a "connection" delay, rather than an assembly delay.

Once the train is assembled on the departure tracks, road locomotives and a caboose are coupled on, an outbound inspection is performed and the train is ready to depart. The outbound inspection and dispatching delays will not be analyzed in detail here. The focus of the analysis is on classification and connection delays, which constitute the bulk of the time a rail car spends in a yard.

The objective of this paper is to utilize queuing models to draw inferences regarding strategies for rail service improvement. It is important to consider the variability in these delays as well as their average values, since variability in yard time is a major contributor to rail service reliability problems. Strategies that might lead to service reliability improvements are of particular interest.

The remainder of the paper is organized as follows. In Section 2 we develop a batch arrival model of classification delays. Since the results of the model remain in the transform domain, we develop "worst" case and "best" case bounds on the mean and variance of classification delay in Sections 2.1 and 2.2, respectively. In Section 2.3 we analyze the sensitivity of classification delays to the variance of train length and the variance of car service times. The variance of train length is shown to have a much more significant effect on delay than does the variance of service times. Section 2.4 analyzes the underlying assumption of Poisson train arrivals. In Section 3 we present a batch service model of connection delays. Section 4 discusses the estimation of total delay based on the classification and connection delay models of Sections 2 and 3, respectively. In particular, we discuss empirical results on the covariance of classification and connection delays—the component of total delay which is not explicitly modeled. Section 5 presents an example of the use of the models to analyze delays to cars at two sequential yards as a function of the dispatching policy at the first yard. Two policies are compared: in the first, irregular length trains are dispatched at constant intervals from the first yard; in the second, constant length trains are dispatched at irregular intervals. The first policy results in smaller connection delays at the first

yard, but in larger classification delays at the second yard. The models allow us to determine the policy which minimizes either the mean delay or the variance of delay as a function of the number of cars per day traveling between the yards in question and of the total utilization level of the second yard. Section 6 contains conclusions and some thoughts on appropriate uses of the models presented.

2. CLASSIFICATION DELAY

THERE ARE a number of different queuing models which could be suggested for the classification operation. Petersen[15] suggests several possible models, including $M/M/S$, $M/D/S$, and $M/G/1$.

It should be noted that Petersen considers the basic units of arrival to the system to be trains, not individual cars, and thus he derives parameters for service time to classify an entire inbound train. While this simplifies representation of some elements of the system, it leads to some confusion about whether or not the output process of the classification queue is really the input to the connection/assembly queue. Because the model is conceptually more clear if these are considered to be serial queues, we have adopted the perspective that the models should be based on individual rail cars at each stage. This also aids the analysis of delay variability.

As a result, we must recognize that rail cars arrive at the yard in batches (trains). This dictates the use of a batch arrival queuing model to analyze the classification delays. If the batches arrive as a Poisson process, we can utilize a result developed by BURKE,[3] based on the earlier work of GAVER.[10] This model is denoted $M^X/G/1$, where X is a random variable corresponding to train length. The distribution of waiting time has the Laplace-Stieltjes transform $F(z)$:

$$F(z) = ((1-\rho)z[1-B(z)])/L_1[1-\beta(z)]\{z-\lambda[1-B(z)]\} \quad (1)$$

where

λ = mean arrival rate of trains
$\rho = \lambda L_1/\mu$ = traffic intensity
L_1 = mean train length (cars)
μ = mean classification service rate (cars/unit time)
$\beta(z)$ = Laplace-Stieltjes transform of the service time distribution
$B(z) = \sum_{j=0}^{\infty} C_j[\beta(z)]^j$
C_j = probability that train is of length j.

The average wait time for classification is $E(T_1)$:

$$E(T_1) = (1/2\mu)[(1/1-\rho)((L_2/L_1) + \rho\mu^2\sigma^2) - 1] \quad (2)$$

where

L_2 = second moment (about the origin) of train length

σ^2 = variance of service time distribution.

The variance in classification delay is $V(T_1)$:

$$V(T_1) = (\rho/(1-\rho)^2)[\rho\gamma_1^2 + (1-\rho)\gamma_2] + \xi_2 - \xi_1^2 \qquad (3)$$

where γ_1 and γ_2 are the first and second moments about the origin of the distribution whose transform is:

$$G(z) = \mu[1 - B(z)]/L_1 z \qquad (4)$$

and ξ_1 and ξ_2 are the first two moments of the distribution whose transform is:

$$W(z) = (1 - B(z))/L_1[1 - \beta(z)]. \qquad (5)$$

Equations 2–5 can be used to predict the mean and variance of classification delay, given information on the mean arrival rate of inbound trains, the train length distribution, and the classification service time distribution. This model assumes a single server (classification facility). This assumption is certainly appropriate in hump yards. In large flat yards, however, it is common to have multiple switch engines doing classification work. In such situations, more complicated multiple-server models might be considered. However, these multiple servers do not operate independently, as assumed by multiple-server queuing models, and thus detailed treatment of the service characteristics is difficult. In light of this complexity, a more fruitful course is likely to be construction of an "effective" single-server service-time distribution for the multiple engines operating together. Equations 2–5 could then be used with this effective service time distribution to predict yard performance.

The distributions of train length and service time will be specific to a particular application. In the interest of developing some general insight into the nature of classification delays, we will focus on two cases which are likely to provide upper and lower bounds on real situations.

2.1. A "Worst Case" Bound

The first case of interest results from assuming that train lengths are geometrically distributed with mean L_1, and that classification service times are exponentially distributed with mean μ^{-1}. This represents a case of extreme variability in both train length and service times, and hence provides a "worst case" analysis.

The geometric distribution of train lengths implies that:

$$C_j = (1-\alpha)\alpha^{j-1}, \qquad j = 1, 2, \cdots \qquad (6)$$

for some value of α, $0 < \alpha < 1$. The mean train length is:

$$L_1 = 1/(1 - \alpha). \tag{7}$$

Thus, if average train length is given, the value of α can be found by manipulation of (7), yielding:

$$1 - \alpha = 1/L_1 \tag{8a}$$

or

$$\alpha = L_1 - 1/L_1. \tag{8b}$$

The second moment of train length is:

$$L_2 = (1 + \alpha)/(1 - \alpha)^2. \tag{9}$$

If service times are exponentially distributed with mean μ^{-1}, the variance in service times is μ^{-2}, and we may write the expression for mean classification delay as follows:

$$\begin{aligned} E(T_1) &= [1/2\mu(1 - \rho)][((1 + \alpha)/(1 - \alpha)) - 1 + 2\rho] \\ &= (L_1 - 1 + \rho)/\mu(1 - \rho). \end{aligned} \tag{10}$$

If $\alpha = 0$, so that trains were all one car long ($L_1 = 1$), (10) obviously reduces to the standard result for mean waiting time in an $M/M/1$ system. Thus, Equation 10 represents a straightforward generalization to the case where arrivals are in batches.

To obtain an expression for the variance of classification delays, we note that if service times are exponentially distributed with mean μ^{-1},

$$\beta(z) = \mu/(\mu + z). \tag{11}$$

Combining (6) and (11), we have:

$$\begin{aligned} B(z) &= \textstyle\sum_{j=1}^{\infty} (1 - \alpha)\alpha^{j-1}[\mu/(\mu + z)]^j \\ &= (1 - \alpha)/\alpha \textstyle\sum_{j=1}^{\infty} [\alpha(\mu/(\mu + z))]^j \\ &= (1 - \alpha)\mu/(\mu(1 - \alpha) + z). \end{aligned} \tag{12}$$

Substituting (12) and (8a) into (4) produces the result:

$$\begin{aligned} G(z) &= (\mu/L_1)(1 - [(1 - \alpha)\mu/(\mu - \alpha\mu + z)])/z) \\ &= \mu/L_1(\mu(1 - \alpha) + z) \\ &= (\mu/L_1)/(\mu/L_1 + z). \end{aligned} \tag{13}$$

This is the transform of an exponential distribution which may be written as follows:

$$f(t) = G^{-1}(z) = (\mu/L_1)e^{-\mu t/L_1}, \qquad t \geq 0. \tag{14}$$

The first two moments of this distribution are:

$$\gamma_1 = L_1/\mu \tag{15}$$

$$\gamma_2 = 2L_1^2/\mu^2. \tag{16}$$

Substituting (11) and (12) into (5), we obtain:

$$W(z) = (\mu + z)/L_1[\mu(1 - \alpha) + z]. \tag{17}$$

The first two moments of the distribution whose transform is given by (17) may be evaluated by noting that:

$$\xi_1 = -W'(0) \tag{18}$$

$$\xi_2 = W''(0). \tag{19}$$

The first two derivatives of $W(z)$ can be evaluated quite easily, producing the following results:

$$\xi_1 = (L_1 - 1)/\mu \tag{20}$$

$$\xi_2 = (2L_1(L_1 - 1))/\mu^2. \tag{21}$$

Substituting (15), (16), (20) and (21) into (3), we obtain the variance of classification delay:

$$V(T_1) = (1/\mu^2)[(L_1/(1 - \rho))^2 - 1]. \tag{22}$$

2.2. A "Best Case" Bound

A second case of interest is obtained by assuming constant train lengths and deterministic service times. This case represents extreme regularity and provides a "best case" analysis. Deterministic train lengths imply $L_2 = L_1^2$, and deterministic service times imply $\sigma^2 = 0$. Thus, the expression for the mean classification delay is:

$$E(T_1) = (L_1 - 1 + \rho)/2\mu(1 - \rho). \tag{23}$$

Note that this is exactly one-half as large as the delay in the "worst case." This result is analogous to the well-known comparison between the $M/M/1$ and $M/D/1$ systems.

The variance of classification delays may be determined by noting that for service times that are deterministic and equal to μ^{-1}:

$$\beta(z) = e^{-z/\mu} \tag{24a}$$

$$B(z) = e^{-L_1 z/\mu} \tag{24b}$$

$$G(z) = (\mu/L_1)[(1 - e^{-L_1 z/\mu})/z] \tag{25}$$

and

$$W(z) = (1 - e^{-L_1 z/\mu})/L_1(1 - e^{-z/\mu}). \tag{26}$$

TABLE I

Summary of Classification Delay Bounds

"Worst Case"—Geometric Train Length, Exponential Classification Service Times	
$E(T_1) = (L_1 - 1 + \rho)/\mu(1 - \rho)$	(Eq. 10)
$V(T_1) = (1/\mu^2)[(L_1/(1 - \rho))^2 - 1]$	(Eq. 22)
"Best Case"—Deterministic Train Length; Deterministic Classification Service Times	
$E(T_1) = (L_1 - 1 + \rho)/2\mu(1 - \rho)$	(Eq. 23)
$V(T_1) = (1/12\mu^2)[((1 + 2\rho)/(1 - \rho)^2)L_1^2 - 1]$	(Eq. 31)

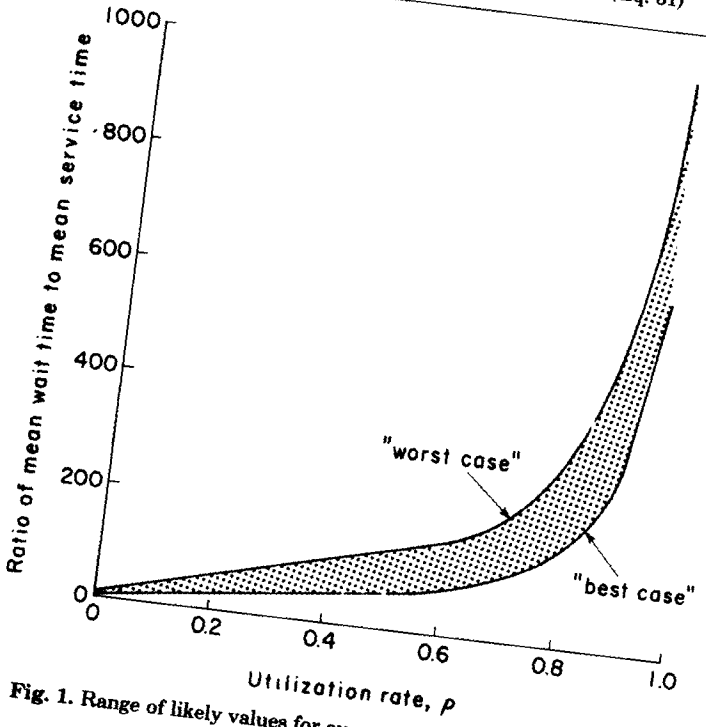

Fig. 1. Range of likely values for average wait time as a function of ρ.

To obtain γ_1, γ_2, ξ_1 and ξ_2, we differentiate $G(z)$ and $W(z)$. The first two derivatives of $G(z)$ can be evaluated using l'Hopital's rule to yield:

$$\gamma_1 = L_1/2\mu \tag{27}$$

$$\gamma_2 = L_1^2/3\mu^2. \tag{28}$$

In a similar fashion, evaluation of the first two derivatives of $W(z)$ yields:

$$\xi_1 = (L_1 - 1)/2\mu \tag{29}$$

$$\xi_2 = (L_1 - 1)(2L_1 - 1)/6\mu^2. \tag{30}$$

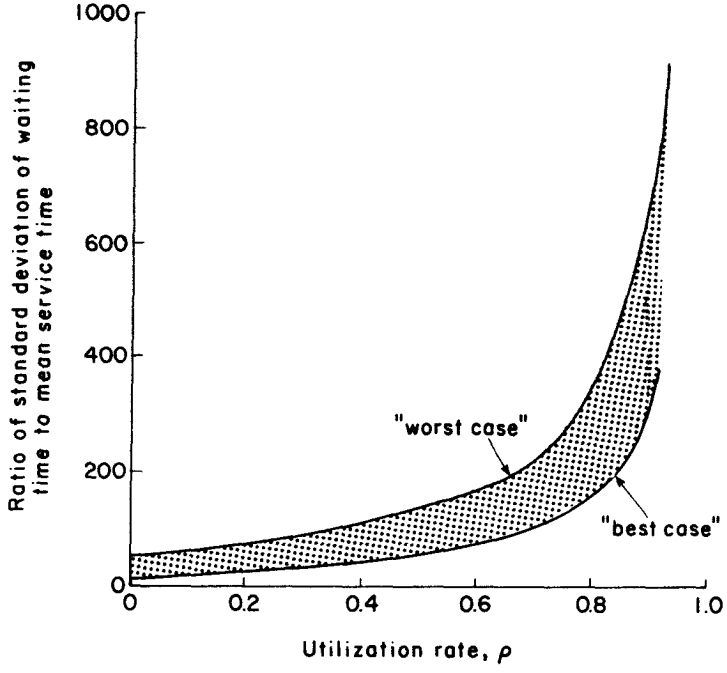

Fig. 2. Range of possible values for the standard deviation of waiting time as a function of ρ.

Substituting (27)-(30) into (3), we obtain the result for the variance of classification time:

$$V(T_1) = (1/12\mu^2)[((1 + 2\rho)/(1 - \rho)^2)L_1^2 - 1]. \qquad (31)$$

The variance given by (31) will always be less than the variance given by (22). The magnitude of the difference depends upon ρ and L_1, but the variance of the "worst case" is between 4 and 12 times as large as that of the "best case." Table I summarizes the bounds on the mean and variance of classification delays obtained in this and the previous section.

Equations 10 and 23 can be used to provide likely bounds on observed mean classification delays, and Equations 22 and 31 provide a similar reference for variance. Figure 1 shows the range of values for mean delay as a function of the utilization level, ρ. Figure 2 provides comparable information for the standard deviation of delay. Standard deviation is graphed rather than variance, in order to indicate units comparable with mean delay. Note that the difference in the bounds for mean delay is a factor of 2, and the difference is even larger for the standard deviation of delays. In both Figures 1 and 2, mean train length is assumed to be 66 cars, the value reported by the ASSOCIATION OF AMERICAN RAILROADS[1] as the average for all U.S. Class I railroads in 1979.

Since real situations are likely to be somewhere between the bounds represented by the "worst case" and "best case" analyses, Figures 1 and 2 provide useful information on the range within which observed values should fall. In order to gain a better understanding of how close to one bound or the other a particular situation might be, it is useful to examine the sensitivity of both mean delay and the variance of delay to changes in the variability of train length and service times.

2.3. Sensitivity Analysis

Let us first examine the expression for mean classification delay. Holding the mean train length constant, increases in the variance of train lengths are represented as increases in L_2. A useful way to describe the effect of changes in L_2 on $E(T_1)$ is by defining the elasticity of mean wait time with respect to train length variability. Denote this quantity δ_T; it is defined as follows:

$$\delta_T = (\partial E(T_1)/\partial L_2) \cdot (L_2/E(T_1)). \tag{32}$$

The elasticity denotes the percentage change in $E(T_1)$ which would result from a 1% change in L_2.

Differentiating the expression in (2) and substituting into (32), we obtain the result:

$$\delta_T = (L_2/L_1)/(L_2/L_1 + \rho\mu^2\sigma^2 - (1 - \rho)). \tag{33}$$

We note first that this quantity is positive. As we would expect, increases in train length variability always increase mean delay. Secondly, note that $\delta_T \geq 1$ when $\rho(\mu^2\sigma^2 + 1) \leq 1$. Thus, when service is completely regular ($\sigma^2 = 0$), a given percentage change in L_2 produces a larger percentage change in $E(T_1)$. As the service times become less regular, as the mean service rate increases, or as the utilization level increases, $E(T_1)$ becomes somewhat less sensitive to variation in train length.

Further insight can be obtained by considering the likely magnitudes of L_1, L_2, ρ, μ and σ^2 in practical situations. The variance of service time is likely to be less than μ^{-2} (exponential service). Thus, the second term in the denominator of (33) is likely to be between 0 and 1, and will tend to offset the $1 - \rho$ term. Typical average train lengths are between 60 and 70 cars, with a standard deviation in the range of 15–30 cars. Thus, an approximate range for the ratio L_2/L_1 is 60–80. It is clear that the term L_2/L_1 is much larger than either of the other terms and hence, $\delta_T \approx 1$.

In a similar fashion, the elasticity of mean delay with respect to the variance of service times can be developed. Define this elasticity, δ_S, as follows:

$$\delta_S = (\partial E(T_1)/\partial \sigma^2)(\sigma^2/E(T_1)). \tag{34}$$

Differentiating Equation 2 and substituting, we obtain the result:

$$\delta_S = \rho\mu^2\sigma^2/(L_2/L_1 + \rho\mu^2\sigma^2 - (1 - \rho)). \tag{35}$$

Because $L_2/L_1 \gg \rho\mu^2\sigma^2 - (1 - \rho)$ for most practical situations, δ_S will be very small. An order of magnitude estimate would be 10^{-2}. Thus, we can see that mean delay for classification is *much* more sensitive to train length variation than to service time variation. This has implications for potential strategies to improve yard performance, an example of which is discussed in Section 5.

Examination of the sensitivity of the variance in classification delay to variations in train length and service time is more difficult, but some similar insights may be obtained. Consider the case of Erlang-k service times with geometric train lengths. Obviously the "worst case" analysis of the previous section is a special case ($k = 1$), and by examining the sensitivity of variance in classification delays to changes in k (with the mean service time remaining constant) we can develop an understanding of the effect of service time variability.

The Laplace transform of an Erlang-k service time distribution may be written as follows:

$$\beta(z) = (m/(m + z))^k \tag{36}$$

where $m/k = \mu$, the average service rate. The variance in service time is then k/m^2. If we maintain a constant average service time, this variance will be proportional to k^{-1}; thus, increases in k will test the effects of reduced variability of service times.

If train lengths are geometrically distributed, we can write $B(z)$ as follows:

$$\begin{aligned} B(z) &= \textstyle\sum_{j=1}^{\infty} (1 - \alpha)\alpha^{j-1}(m/(m + z))^{kj} \\ &= (1 - \alpha)m^k/((m + z)^k - \alpha m^k). \end{aligned} \tag{37}$$

The transforms of the distributions whose first two moments are required for evaluation of the variance are then:

$$G(z) = (\mu/L_1)(((m + z)^k - m^k)/(z[(m + z)^k - \alpha m^k])) \tag{38}$$

and

$$W(z) = (m + z)^k/(L_1[(m + z)^k - \alpha m^k]). \tag{39}$$

We can evaluate the first two moments of these distributions by evaluating derivatives. The evaluation of the first two derivatives of $G(z)$ requires application of l'Hopital's rule. In the interest of space, these derivations are omitted, and only the results will be presented. The first two moments are as follows:

$$\gamma_1 = (1 - \alpha + (1 + \alpha)k)/2k\mu(1 - \alpha) \tag{40}$$

$$\gamma_2 = (k^2(1 + 4\alpha + \alpha^2) + 3k(1 - \alpha^2) + 2(1 - \alpha)^2)/3k^2\mu^2(1 - \alpha)^2. \tag{41}$$

Evaluation of the derivatives of $W(z)$ is somewhat easier, and yields the following results:

$$\xi_1 = (L_1 - 1)/\mu \tag{42}$$

$$\xi_2 = ((L_1 - 1)/k\mu^2)[L_1 k(1 + \alpha) + 1]. \tag{43}$$

Note that ξ_1 does not depend on k. Clearly, for $k = 1$, (40)–(43) reduce to the expressions in (15), (16), (20) and (21). However, by substituting (40)–(43) into (3), the variance of delay can be evaluated for various values of k.

Insight on the effect of service time variability may be obtained by looking at the ratio of the variance in classification delay for Erlang-k service to the variance in delay when service is exponential. In general, this ratio depends upon α (or L_1, the mean train length), ρ and k, and will decrease as k increases (implying increasing regularity of service). However, the dependence on ρ is very weak, and for mean train lengths in the neighborhood of 60 cars, the sensitivity of this ratio to k is also very limited. To see this, we can find the limiting values of γ_1, γ_2, ξ_1 and ξ_2 as $k \to \infty$. The limiting values are as follows:

$$\lim_{k\to\infty}\gamma_1 = (1 + \alpha)/2\mu(1 - \alpha) \tag{44}$$

$$\lim_{k\to\infty}\gamma_2 = (1 + 4\alpha + \alpha^2)/3\mu^2(1 - \alpha)^2 \tag{45}$$

$$\lim_{k\to\infty}\xi_1 = (L_1 - 1)/\mu \tag{46}$$

$$\lim_{k\to\infty}\xi_2 = L_1(L_1 - 1)(1 + \alpha)/\mu^2. \tag{47}$$

Using these limiting values, we find that the limiting value of the variance ratio when $L_1 = 60$ is approximately 0.99.

We can conclude from this that the variance in classification delay is only weakly related to the variance in service times. Clearly then, the substantial difference in variance between the "worst case" and the "best case" analyses in Sections 2.1 and 2.2 is due largely to the variability of train length. This result is very similar to the result for mean delay, and its implications for strategies to improve yard performance will be discussed more fully in Section 5.

2.4. Train Arrivals

All of the results in this section are based on an underlying assumption that train arrivals at the yard may be considered to be a Poisson process. Therefore, it is worthwhile to examine this assumption more carefully, both from a theoretical perspective and with empirical data.

Because trains arriving at a particular yard often come from many different origin points, there is some theoretical basis on which to expect

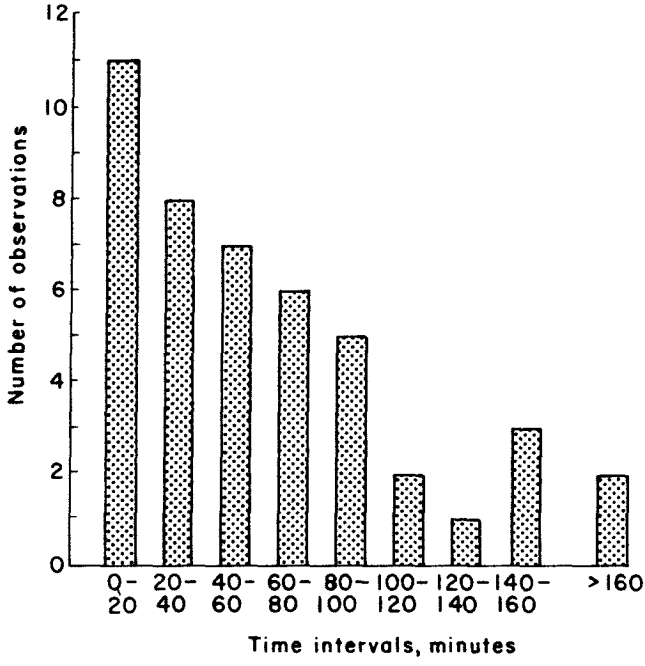

Fig. 3. Histogram of observed train interarrival times.

that the pooled input process is approximately Poisson. Cox[4] shows that the superposition of several renewal processes tends rather rapidly to a limiting Poisson process as the number of individual processes increases. This is true even if the distributions governing inter-event times in the individual processes are quite different from exponential. Thus, since in general there will be several origin points dispatching trains destined for a particular yard, and these individual processes are superposed as the arrival process to the yard in question, we might expect the Poisson model to be quite acceptable, even though individual processes may be scheduled or semischeduled.

Data gathered at a major Midwestern yard facility support this conclusion empirically. Figure 3 shows the histogram of interarrival times for 45 samples collected. Neither a Chi-square test nor a Kolmogorov-Smirnov test against an exponential distribution (corrected to account for the fact that the parameter is estimated from the data; see [12]) rejects the hypothesis that these data are samples from an exponential distribution at any reasonable confidence level. The Chi-square test statistic is 0.9588 with 4 degrees of freedom and the K-S test statistic is 0.12. Thus, while the empirical data set is a small one, it appears that the Poisson model for train arrivals is appropriate.

3. CONNECTION DELAYS

ONCE CARS have been classified, they must wait for dispatch on an appropriate outbound train. Operationally, we can think of this process as being one in which cars arrive on the classification tracks, either singly or in small groups (cuts), and wait for the designated outbound train to be "called." At this point, all the cars for this train are assembled, and when made up, the train departs. In terms of a queuing model, we may think of this as a batch-service system in which the "server" is the outbound train. Service for a batch of cars begins when the appropriate outbound train is called for assembly, and the service time is the time between successive outbound trains on which a given cut of cars may be dispatched. The delay time for connection with the outbound train is then the waiting time in queue derived from such a queuing model.

It should be noted that this perspective on modeling the system places principal emphasis on the outbound train schedule as the source of delay for cars following classification. Delays in assembly due to insufficient numbers of switch engines and crews are not considered directly. This effect is only represented indirectly, in terms of late departures of outbound trains, for example. The emphasis on schedule is in keeping with the findings of several previous researchers,[7,18] and has been recognized by a rail industry task force on reliability studies.[8]

The average delay for a simple batch-service queue of this type can be derived easily. Let us assume that individual cars arrive randomly in time (i.e. as a Poisson process) from the classification operation, and that the outbound train takes all cars available at the time it is assembled. Because the classification queue is not Markovian, the assumption that its output is a Poisson process is not strictly true. However, exact characterization of the interdeparture times is a very difficult problem, and is unlikely to produce important additional insights. In order to preserve computational tractability, we will simply assume that a Poisson approximation is sufficiently accurate for our purposes. The second assumption means that train length constraints on the outbound trains are ignored, for the time being. We will return to this issue following the basic derivation.

Define a random variable, H, with probability density function, $g(h)$, $0 \leq h < \infty$, as the time interval between successive outbound trains for a given block of cars. If cars arrive randomly in time on the classification tracks, the expected delay is:

$$E(T_2) = E(H)/2 + V(H)/2E(H) \tag{48}$$

where $V(H)$ is the variance in the time interval between successive departures. Equation 48 is analogous to a result widely used in studies of urban mass transit systems, expressing the mean waiting time of passen-

gers at a transit stop. Derivations of the same result in that context can be found in WELDING[21] or OSUNA AND NEWELL.[13]

Note that if departures are completely regular ($V(H) = 0$), the second term vanishes, and the expected delay is one-half the interval between trains (e.g., 12 hours for trains dispatched once per day). On the other hand, if dispatches occur very irregularly, the second term indicates that expected delay to cars will increase.

An underlying assumption in Equation 48 is that outbound train length is unlimited, or in queuing terms, that the batch size is infinite. In practical terms, this assumption is not really true, since there are limits to the length of train which can be dispatched. Such limits can be the result of mainline track configuration, power availability, etc. More sophisticated batch-service queuing models can be constructed to reflect these constraints, but solution is extremely difficult in all but the simplest limiting cases.[14] Thus, we have chosen to work with the simpler infinite batch size model to retain analytic tractability, even though it must be viewed as an approximation.

The variance of waiting times can also be derived quite readily. The details are given by FRIEDMAN.[9] The result is:

$$V(T_2) = (E(H^3)/3E(H)) - [E(T_2)]^2. \tag{49}$$

In many cases, the distribution of times between successive train departures, $g(h)$, will be symmetric (or nearly so). When $g(h)$ is symmetric the skewness is zero, and the third moment may be written as:

$$E(H^3) = 3E(H)V(H) + [E(H)]^3 \tag{50}$$

The variance in wait time is then:

$$\begin{aligned} V(T_2) &= V(H) + [E(H)]^2/3 - [E(T_2)]^2 \\ &= V(H) + [E(H)]^2/3 - [E(H)/2 + V(H)/2E(H)]^2 \\ &= [E(H)]^2/12 + V(H)/2 - [V(H)/2E(H)]^2. \end{aligned} \tag{51}$$

The expression in Equation 51 reflects the dependence of $V(T_2)$ on the mean and variance of the headway distribution between successive outbound trains. A primary implication of Equations 48 and 51 is that more regular dispatch of outbound trains will reduce both the mean and the variance of connection delays in the yards.

4. PREDICTION OF TOTAL DELAY

THE BULK of the total delay to rail cars in passing through a classification yard can be represented as the sum of waits for classification and

connection. Thus, if we define D as the total delay, we have:

$$E(D) = E(T_1) + E(T_2), \tag{52}$$

and

$$V(D) = V(T_1) + V(T_2) + 2\,\mathrm{cov}(T_1, T_2). \tag{53}$$

To test whether or not the covariance term in (53) is significant, a random sample of 115 car records from one major yard was used to estimate correlations between wait time for classification and connection delay. The sample correlation coefficient was 0.31, with an approximate 95% confidence interval on the true correlation of (0.14, 0.48). Based on this one set of data, the hypothesis of significant correlation cannot be rejected.

Note that the sample correlation is positive, indicating that cars which spend a long time waiting for classification are likely to also spend a long time waiting for outbound connections. To some extent, this probably reflects prioritization of cars within the yard. High priority traffic (e.g. high-value merchandise, TOFC/COFC, etc.) moves through both phases of the yard more rapidly than low priority cars (e.g. low-value bulk commodities). However, a second interpretation is also possible. The positive correlation may indicate that strategies which tend to reduce long delays in one part of the yard also reduce delays in the other part. If this is true, a "double" effect can be generated, making such strategies very effective.

Because the true nature of this covariance is not well understood, and because the empirical data on which the estimated correlation was based are limited, the covariance term will not be included in further analysis. Since the estimated covariance is positive, this omission means that predictions of effects on variance reduction will be conservative. This appears to be an appropriate assumption, pending further empirical analysis of other yard operations.

5. EFFECTS OF TRAIN DISPATCHING STRATEGIES

THE MODELS of classification and connection delays developed in Sections 2 and 3 may be used to indicate the effects of many possible strategies for reducing the mean and variance of delay time in railyards. In this section, we examine one class of strategies, focusing on the regularity of dispatching of outbound trains. Recall that the sensitivity analysis of the classification delay model indicated that both the mean delay and the variance of delays were relatively sensitive to inbound train length variation. This implies that making train lengths relatively constant is a potential means for reducing classification delays. On the other hand, the connection delay model indicates that delays are sensitive to the regularity of outbound dispatches. This raises an important issue. If outbound trains

at one yard are dispatched regularly so as to reduce connection delays there, these trains will tend to be of variable length, since the arrival of cars on the classification tracks is a stochastic process. Thus, the classification delays at the destination yards for these trains will be increased because of the train length variability. An alternative strategy would be to make outbound train lengths relatively constant so as to reduce classification delays at the destination yards. However, this implies that intervals between successive outbound trains will be more irregular, increasing the connection delay at the origin yard.

We will examine two extreme cases of a dispatching strategy for outbound trains. In one case, trains are dispatched regularly, every H hours, with whatever traffic is available. In the other case, trains are dispatched whenever L cars are available, regardless of time. Consider a prototypical situation involving two yards, A (the dispatching yard) and B (the terminating yard).

In the first case, since departures from A are at regular intervals, the expected connection delay is:

$$E(T_{2A}) = H/2. \tag{54}$$

The variance of the delay can be determined using Equation 51, with $V(H) = 0$. This yields the result:

$$V(T_{2A}) = H^2/12. \tag{55}$$

Since we assume Poisson arrivals of cars on the classification tracks, the number of cars available at A at a fixed time after the departure of the previous train is a Poisson random variable. If the arrival rate of cars per hour for yard B is r, the parameter of this random variable is rH. At yard B arriving trains then have a length which is Poisson distributed, with mean rH and second moment $rH(rH + 1)$.

The sensitivity analysis in Section 2 indicated that the service time distribution plays only a small role in determining the mean and variance of classification delay. For convenience, let us assume that service is deterministic, so $\sigma^2 = 0$. The mean delay for classification at yard B is then found using Equation 2:

$$E(T_{1B}) = (rH + \rho_B)/2\mu_B(1 - \rho_B) \tag{56}$$

where the subscript B has been used to denote parameters for yard B.

To evaluate the variance of classification delay at B, we note that:

$$C_j = ((rH)^j/j!)e^{-rH}, \quad j = 0, 1, \cdots. \tag{57}$$

If service times are deterministic, we know that:

$$B(z) = e^{-z/\mu_B}. \tag{58}$$

Thus, $B(z)$ is given by:

$$\begin{aligned} B(z) &= \textstyle\sum_{j=0}^{\infty} ((rH)^j/j!)e^{-rH}[e^{-z/\mu_B}]^j \\ &= e^{-rH} \textstyle\sum_{j=0}^{\infty} [rHe^{-z/\mu_B}]^j/j! \\ &= \exp(-rH)\exp(rHe^{-z/\mu_B}) \\ &= \exp[rH(e^{-z/\mu_B} - 1)]. \end{aligned} \tag{59}$$

$G(z)$ is then:

$$G(z) = (\mu_B/rH)(1 - \exp[rH(e^{-z/\mu_B} - 1)])/z. \tag{60}$$

The first two moments, γ_1 and γ_2, can be evaluated from the derivatives of $G(z)$ evaluated at $z = 0$. This yields the following results:

$$\gamma_1 = (rH + 1)/2\mu_B \tag{61}$$

$$\gamma_2 = ((rH)^2 + 3rH + 1)/3\mu_B{}^2. \tag{62}$$

From (58) and (59), we can also determine $W(z)$:

$$W(z) = (1 - \exp[rH(e^{-z/\mu_B} - 1)])/(rH(1 - e^{-z/\mu_B})). \tag{63}$$

Evaluating the derivatives of $W(z)$ yields the following results for ξ_1 and ξ_2:

$$\xi_1 = rH/2\mu_B \tag{64}$$

$$\xi_2 = rH(3 + 2rH)/6\mu_B{}^2. \tag{65}$$

The variance of classification delay at B can then be evaluated using Equation 3. The result is:

$$V(T_{1B}) = ((2\rho_B + 1)(rH)^2 + 6rH + \rho_B(4 - \rho_B))/12\mu_B{}^2(1 - \rho_B)^2. \tag{66}$$

An alternative strategy at yard A is to dispatch a train whenever a fixed number of cars are available. Trains will then all be of equal length, but will depart irregularly. If arrivals of cars for B on the outbound tracks of yard A are again assumed to follow a Poisson process the time required to accumulate L cars will have an Erlang distribution with parameters L and r. The mean time between trains will be L/r and the variance will be L/r^2. Average connection delay is then:

$$E(T_{2A}) = L/2r + (L/r^2)/(2L/r) = (L + 1)/2r. \tag{67}$$

The third moment about the origin of an Erlang (L, r) random variable is $L(L + 1)(L + 2)/r^3$. Using Equation 49, the variance of connection delay is then:

$$V(T_{2A}) = (L + 1)(L + 5)/12r^2. \tag{68}$$

TABLE II

Relations Used in Dispatching Policy Comparison

Regular Dispatches		Constant Train Lengths	
Dispatching Delay (Yard A):		Dispatching Delay (Yard A):	
$E(T_{2A}) = H/2$	(Eq. 54)	$E(T_{2A}) = (L+1)/2r$	(Eq 67)
$V(T_{2A}) = H^2/12$	(Eq 55)	$V(T_{2A}) = (L+1)(L+5)/12r^2$	(Eq. 68)
Classification Delay (Yard B)[a]		Classification Delay (Yard B)[b]	
$E(T_{1B}) = (rH + \rho_B)/2\mu_B(1-\rho_B)$	(Eq. 56)	$E(T_{1B}) = (L-1+\rho_B)/2\mu_B(1-\rho_B)$	(Eq. 23)
$V(T_{1B}) = ((2\rho_B+1)(rH)^2 + 6rH + \rho_B(4-\rho_B))/$ $12\mu_B^2(1-\rho_B)^2$	(Eq. 66)	$V(T_{1B}) = (1/12\mu_B^2)[((1+2\rho_B)/(1-\rho_B)^2)/$ $L^2-1]$	(Eq. 31)

[a] Assuming deterministic classification service times and Poisson distributed train lengths.
[b] Assuming deterministic classification service times and constant train lengths

At yard B, inbound trains are of constant length, and if we again assume deterministic service times, the mean and variance of classification delay are given by Equations 23 and 31.

Table II summarizes the key equations needed to compare a regular dispatch policy with a constant train length policy at yard A. To maintain comparability between the two cases we will insist that the total number of trains operated is the same. This implies that the constant train length in the second case must equal the average train length in the first case, or $L = rH$. This identity can be used to relate Equations 54 and 55 to Equations 67 and 68.

We are now able to compare the two dispatching strategies, by examining the net changes in mean delay and variance of delay summed over both yards, as the dispatching policy changes. A policy of constant train length will result in larger mean connection delays at A and lower mean classification delays at B. There will be a net decrease in mean delay, relative to regular dispatches, if the savings at B outweigh the increases at A. Using Equations 23, 54, 56 and 67, this condition may be written as follows:

$$\rho_B > (\mu_B - r)/\mu_B = \rho^*_{\text{MEAN}}. \tag{69}$$

A similar statement of the conditions under which the constant train length policy reduces the variance of total delay can be obtained using Equations 31, 55, 66 and 68. The resulting condition is:

$$(6L + 2\rho_B - 1)/\mu_B^2(1-\rho_B)^2 > (6L+5)/r^2. \tag{70}$$

Alternatively, the condition may be written as a quadratic equation in

ρ_B:

$$[(6L+5)\mu_B{}^2]\rho_B{}^2 - 2[(6L+5)\mu_B{}^2 + r^2]\rho_B + [(6L+5)\mu_B{}^2 - (6L+1)r^2] < 0. \tag{71}$$

The roots of the quadratic are:

$$\rho^*_{\text{VAR}} = 1 + (r^2 \pm \mu_B r\sqrt{3(6L+5)(2L+1) + r^2/\mu_B{}^2})/(6L+5)\mu_B{}^2. \tag{72}$$

We are interested only in the smaller root since the larger root will always exceed 1.0. Since we require $r < \mu_B$ for steady-state and since $6L \gg 1$, the smaller root may be approximated by:

$$\begin{aligned}\rho^*_{\text{VAR}} &\approx 1 - r\sqrt{3(6L+5)(2L+1)}/(6L+5)\mu_B \\ &\approx 1 - r(6L+4)/\mu_B(6L+5) \\ &\approx 1 - r/\mu_B = \rho^*_{\text{MEAN}}.\end{aligned} \tag{73}$$

Thus, relation (71) implies that for

$$\rho^*_{\text{VAR}} < \rho_B < 1 \tag{74}$$

the constant train length policy reduces the variance of total delay, where ρ^*_{VAR} is the smaller root in (72) which may be approximated by (73).

Equations 70–74 implicitly assume that the covariance of connection delay at yard A and classification delay at yard B is unaffected by changes in the dispatching policy at yard A. Conditions (69) and (74) depend upon the utilization level at yard $B(\rho_B)$, the service rate at $B(\mu_B)$, the level of traffic between A and $B(r)$ and the average train length (L). For operations planning we may assume μ_B is fixed, since the service rate typically depends largely on the physical layout of yard B and will not be adjustable in the short run. Note also that neither condition depends strongly on L, the mean train length. In fact, condition (69) under which a constant train length policy reduces the mean delay is independent of L. The major variables are thus ρ_B and r.

Figure 4 shows these conditions graphically, for a value of $\mu_B^{-1} = 1$ minute (mean classification time per car at yard B) and a mean train length of $L = 60$. Thus, for example, if $\rho_B = 0.9$ and traffic volumes are greater than 144 cars/day, a constant train length dispatching policy is superior to a regular schedule. For volumes less than 144 cars/day the reverse is true. We note that there is a *very* small region between the two conditions in which a trade-off must be made between a constant train length policy to reduce the mean delay and a regular schedule to reduce the variance. The region becomes larger as the train length decreases and as the level of traffic increases. However, even with a mean train length of 40 cars and 500 cars per day between yards A and B, the range in the

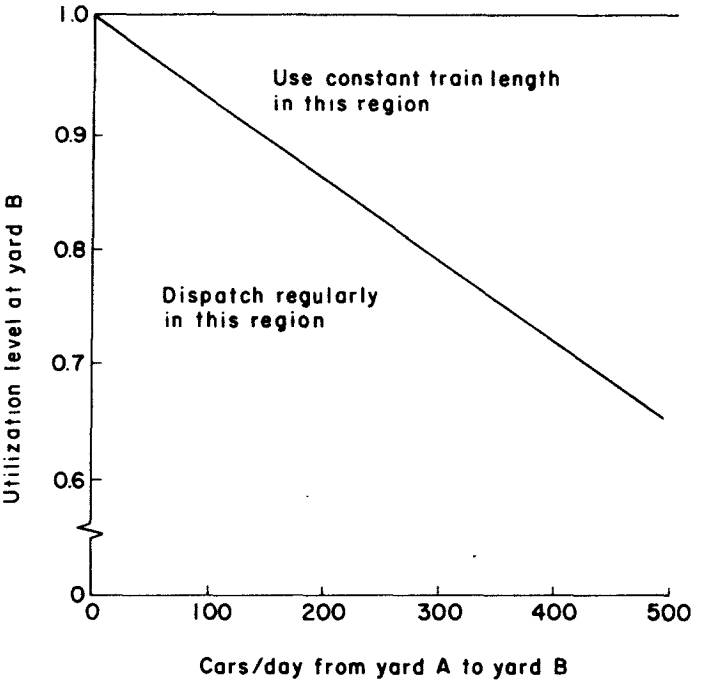

Fig. 4. Regions of effectiveness for alternative dispatching strategies.

utilization level at yard B over which a trade-off must be made is less than 0.002. Alternatively, for an average train length of 40 cars, the trade-off region is always less than 3 cars/day for traffic volumes up to 500 cars/day. This region is too small to show in Figure 4.

It should also be noted that this analysis is based on an assumption of random (Poisson) arrivals of cars on the classification tracks at A. It may be (for some blocks at least) that this assumption is violated quite badly. For example, outbound trains may be scheduled based on expected arrival times of blocks of cars from certain inbound trains. In that case, both the expected connection delay and its variance would be smaller than predicted by the models used here. The reductions in delay from regularizing outbound dispatches would also be overstated. This would tend to shift the curves in Figure 4 down and to the left, making it effective to run constant length trains at somewhat lower values of ρ_B and r.

Finally, the relationship of these results to earlier work by Folk[6] should be noted. He conducted a series of simulation experiments with different dispatching policies on a simple railroad network. His conclusion, in general, was that the best dispatching policy lies somewhere between strict schedule adherence and insistence on constant train length.

This is certainly likely to be true, and detailed knowledge of a particular situation can be used to tailor operating strategies effectively. The major insight offered by the models developed here is to show the general nature of the trade-offs involved in evaluating different operating strategies, and to indicate situations in which various types of strategies are likely to be most effective, as shown in Figure 4. These results can be obtained without detailed simulation of each particular network, and thus the models developed here provide useful screening tools.

6. CONCLUSIONS

QUEUING MODELS of basic operations in railroad classification yards have been developed. The two component models are a batch-arrival model for analyzing delays to railcars prior to classification, and a batch-service model for analyzing connection delays subsequent to classification but prior to departure on outbound trains. The batch-arrival classification delay model assumes that train arrivals at the yard follow a Poisson process and that the yard operates as a single server queue. The batch-service connection delay model assumes that cars arrive at the outbound tracks according to a Poisson process and that there is no limit on train lengths.

Limiting cases of the classification delay model with respect to inbound train length variability and service time variability can be evaluated to illustrate the range of possible performance likely to be observed in any particular yard. Table I summarizes these results as well as the additional assumptions employed. Investigation of the elasticity of mean delay and variance of delay with respect to both train length variability and service time variability indicates that service time variations are relatively unimportant. The variability of train lengths (batch sizes) is a much more important contributor to both the mean and variance of delays prior to classification.

The model of connection delays indicates that both the average delay and the variance of delays are affected by the distribution of times between successive outbound trains. In particular, irregular departures of outbound trains increase both the mean and the variance of connection delay.

An interesting trade-off is thus presented in developing an operational strategy over a network of yards. A simple example involving two yards is used to illustrate this tradeoff. At yard A cars are classified and wait for connecting trains to yard B. In order to keep connection delay at A to a minimum, regular high-frequency trains should be run to B. However, because the number of cars ready to depart for B in each regular interval is stochastic, the train lengths arriving at B are variable, increasing the classification delay there. Alternatively, trains of constant length could

be dispatched from *A* to *B*, in order to reduce the classification delay at *B*. However, such a policy implies more irregular dispatching at *A*, with attendant increases in connection delay. Analysis of this issue has demonstrated that simple rules-of-thumb can be developed using the queuing models, to determine the conditions under which each strategy is appropriate. In general, these rules depend on the level of traffic between *A* and *B*, and on the utilization level of the classification process at yard *B*.

This analysis is an example application of the queuing models, and demonstrates ways in which they can be used effectively in examining various operating policy decisions for railyards. These models can provide useful guidance in the development of appropriate operating policies without resorting to detailed simulation of a large number of options. Because they are based on a number of simplifying assumptions, however, these models should be viewed primarily as screening tools.

For example, Figure 4 could be used to make an initial judgment on dispatching policy between a particular pair of yards. If the volume of traffic and the utilization level of classification facilities at the destination yard are such that current operation is represented by a point far from the dividing line between the regions in Figure 4, a decision is relatively clear-cut, and more detailed analysis is probably not necessary. However, if current conditions are very close to the dividing line, the best operating decision could be sensitive to the simplifying assumptions made in these models, and more detailed examination of options may be required.

Because several simplifying assumptions are required to make the analysis tractable, analytic queuing models of railyards (or of other types of transportation terminals, for that matter) are most effective when used in preliminary analysis. They can aid in identifying situations and alternatives for which decisions are relatively clear, and provide analytic support for decisions in those cases. They also can identify situations where the decisions are much less obvious, and provide guidance for more detailed analysis, either by simulation or other means.

ACKNOWLEDGMENTS

THE WORK reported here was supported in part by the U.S. Department of Transportation, Office of University Research, through contract DOT-OS-70061. The authors also wish to thank Robert Gentzel and John Gray for assistance in data collection.

REFERENCES

1. ASSOCIATION OF AMERICAN RAILROADS, *Yearbook of Railroad Facts*, Washington, D.C., 1980.
2. M. BECKMANN, C. B. MCGUIRE AND C. B. WINSTEN, *Studies in the Economics of Transportation*, Yale University Press, New Haven, 1956.

3. P. J. BURKE, "Delays in Single-Server Queues with Batch Input," *Opns. Res.* **23,** 830–833 (1975).
4. D. R. COX, *Renewal Theory*, Methuen, London, 1967.
5. A. F. DAUGHETY AND M. A. TURNQUIST, *Development of Hybrid Cost Functions from Engineering and Statistical Techniques: The Case of Rail*, Report DOT/RSPA/DPB-50/79/31, U.S. Department of Transportation, Office of University Research, 1979.
6. J. F. FOLK, *Models for Investigating Rail Trip Time Reliability*, Studies in Railroad Operations and Economics, Vol. 5, MIT Department of Civil Engineering, 1972.
7. J. R. FOLK, *Some Analyses of Railroad Data*, Studies in Railroad Operations and Economics, Vol. 6, MIT Department of Civil Engineering, 1972.
8. *Freight Car Utilization and Railroad Reliability: Case Studies*, Report R-283, Association of American Railroads, 1977.
9. R. R. FRIEDMAN, *Statistical Models of the Mean and Standard Deviation of Passenger Wait Time in Urban Bus Transit*, M.S. thesis, Northwestern University Transportation Center, 1976.
10. D. P. GAVER, "Imbedded Markov Chain Analysis of a Waiting Line Process in Continuous Time," *Ann. Math. Statist.* **30,** 698–720 (1959).
11. O. HEIN, "A Two-Stage Queuing Model for a Marshalling Yard," *Rail Int.* **3,** 249–259 (1972).
12. H. W. LILLIEFORS, "On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown," *J. Am. Statist. Assoc.* **64,** 387–389 (1969).
13. E. E. OSUNA AND G. F. NEWELL, "Control Strategies for an Idealized Public Transportation System," *Trans. Sci.* **6,** 52–72 (1972).
14. E. R. PETERSEN, "Bulk Service Queues: With Applications to Train Assembly Times," Working Paper 71-2, Queen's University School of Business, Kingston, Ontario, 1971.
15. E. R. PETERSEN, "Railyard Modeling; Part I. Prediction of Put-Through Time," *Trans. Sci.* **11,** 37–49 (1977).
16. E. R. PETERSEN, "Railyard Modeling; Part II. The Effect of Yard Facilities on Congestion," *Trans. Sci.* **11,** 50–59 (1977).
17. REEBIE ASSOCIATES, *Toward an Effective Demurrage System*, prepared for U.S. Department of Transportation, Federal Railroad Administration, 1972.
18. REID, R. M., J. D. O'DOHERTY, J. M. SUSSMAN AND A. S. LANG, *The Impact of Classification Yard Performance on Rail Trip Time Reliability*, Studies in Railroad Operations and Economics, Vol. 4, MIT Department of Civil Engineering, 1972.
19. M. W. SIDDIQEE, "Investigation of Sorting and Train Formation Schemes," *Proceedings of Fifth International Symposium on Traffic Flow Theory and Transportation*, pp. 377–387, University of California, Berkeley, 1971.
20. J. M. SUSSMAN, C. D. MARTLAND AND A. S. LANG, *Reliability in Railroad Operations: Executive Summary*, Studies in Railroad Operations and Economincs, Vol. 9, MIT Department of Civil Engineering, 1972.
21. P. I. WELDING, "The Instability of Close Interval Service," *Opnl. Res. Quart.* **8,** 133–148 (1957).

Received 1980