Transfer Optimization in a Transit Network

JAMES H. BOOKBINDER
University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

ALAIN DÉSILETS
GIRO, Inc., Montréal, Québec, Canada

Transfer optimization attempts to minimize the overall inconvenience to passengers who must transfer between lines in a transit network. Bus trips are scheduled to depart from their terminal so as to minimize some objective function measuring that inconvenience. In this paper, the transit network is assumed to be given, and the scheduled headway is treated as fixed on each line. We denote by \( t_i \) the departure time of the first bus on line \( i \). \( \{t_i\} \) are termed “offset times,” and constitute the decision variables of our model. To take into account stochastic travel times of buses, our treatment of transfer optimization employs a simulation procedure in combination with an optimization model. That model turns out to be a relaxation of the Quadratic Assignment Problem. It can incorporate a wide range of objective functions (measures of overall passenger disutility) and a variety of policies for holding buses at a transfer point. In the case where buses are not held at all, we show, for a number of different objective functions and transit networks, the negative consequences of optimizing transfers with a deterministic bus-travel-times assumption, if these travel times are in fact random variables. Suggestions are then made for future research.

INTRODUCTION

Any transit user, who has missed a transfer by 30 seconds on a rainy day, knows the importance of transfers in evaluating the level of service provided by a particular transit system. Transfers have a number of attributes which make them particularly inconvenient, such as the discomfort of boarding a new bus, and the negative perception of waiting for the arrival of that transfer vehicle. The latter can be particularly frustrating for the user. In recent years, much attention has thus been devoted to finding ways of reducing some of these inconvenient aspects (Keule,[8] Klemt and Stemme,[9] Abkowitz et al.,[10] Systan,[13] Rivers[12]).

There are basically two ways to reduce waiting time for transfers in a transit network. The first one, decreasing the headway on certain routes, involves an increase in the number of buses and drivers required to operate these routes. The second alternative, transfer coordination, includes two possibilities: timed transfer (Systan[13]) and transfer optimization (Désilets,[5] Klemt and Stemme,[8] Keule,[8] Andresson,[2] Rapp and Gehner[11]).

Timed transfer is a scheduling strategy whereby particular trips are scheduled to meet at certain transfer points, called focal points, within given time windows called contact periods. Layover is usually added at the focal points, to ensure that contact occurs even if some of the trips are late. This addition of idle time implies higher operating costs and thus limits the number of focal points to a few, typically less than ten (Systan[13]). Timed transfer is therefore inappropriate for a large transit network with decentralized transfers, where transfers occur virtually at any intersection between two routes (see Fig. 1). This is the motivation for transfer optimization, the subject of the present paper. In transfer optimization, trips are not required actually to meet at transfer points. Instead, their departures from the terminal are scheduled in order to minimize some objective measuring the overall inconvenience of transfers in the network. The main advantage of transfer
TRANSFER POINT

BUS ROUTE

Fig. 1. A transit network with decentralized transfers.

optimization over timed transfers is that since it is not specifically aiming at making buses meet, there is no need to add layover to ensure contact. Therefore the number of vehicles required to operate a route with a certain headway may not be affected by transfer optimization.

There are two points to address in the optimization of transfers: (a) defining an objective function which accurately reflects the overall inconvenience of transfers in a network under a given timetable, and (b) developing an algorithm to find a timetable which minimizes this objective. The first issue has not been studied extensively. Most papers on the optimization of transfers use total waiting time, calculated under the assumption of deterministic bus travel times, as their objective function. Particular attention will be devoted to improving this aspect in the present paper.

The second, algorithmic issue has gained some attention in recent years. Rapp and Gehner\textsuperscript{[11]} first developed a computerized system for transfer optimization, with total waiting time as its objective function. The system employs an iterative improvement heuristic which starts with an arbitrary timetable and changes the scheduling of one line at a time until no further improvement can be obtained (By line, we mean the portion of a route running from one terminal to the other. For example a route along the North-South axis would have two lines running respectively from North to South and from South to North). A similar system was also developed by Andreasson\textsuperscript{[21]} More recently, Keudel\textsuperscript{[10]} described a system which applies a traffic light synchronization algorithm to the problem of optimizing transfers. Finally, Klem\textsuperscript{[9]} and Stemme\textsuperscript{[9]} give a first formal integer programming definition of the transfer optimization problem, and present a constructive heuristic which schedules one line at a time until they have all been scheduled.

Although randomness of bus travel times is a major cause of inconvenience for transfers, much of the literature does not deal very well with this issue. The preceding references either assume travel times to be deterministic or do not explain how randomness of travel times is taken into account. However, some theoretical work has been done on estimating mean waiting time of transfers, in the case of a single transfer point between two lines with random travel times. Grega and Theede\textsuperscript{[6]} consider an analytic approach to this and compute a closed form solution for a simple particular case. Abkowitz et al.\textsuperscript{[13]} used simulation to compare the mean waiting time for transfers when various policies are used for holding buses at the transfer point.

The main contribution of the present paper will be to show the importance of taking randomness of travel times into account when optimizing transfers, and how it can be done by combining a simulation procedure such as the one of Abkowitz et al.\textsuperscript{[13]} with an optimization model such as the one of Klem\textsuperscript{[9]} and Stemme.\textsuperscript{[9]} The remainder of this paper is organized as follows. Section 1 defines a new transfer optimization model while Section 2 uses it to study the benefits of transfer optimization under random travel times. Finally, conclusions and possibilities for future research are discussed in Section 3.

1. A MODEL FOR TRANSFER OPTIMIZATION

1.1. The Objective Function

In this section we define the mean disutility function which we will use to evaluate the inconvenience under random bus travel times, of a transfer connection going from a trip \( F \) running on a feeder line \( L_F \), to a receiving line \( L_R \). Let \( C \) be the trip on \( L_R \) whose scheduled time of arrival at the transfer point with \( L_F \) is closest to that of \( F \), and let \( N \) be the trip following \( C \) on \( L_R \) (\( F, C \) and \( N \) will be respectively called the Feeder, Critical and Next trips). Let the independent random variables \( t_F, t_C \) and \( t_N \) be the arrival times of these three trips at the transfer point, and assume they respectively take their values in the finite intervals: \([a_F, a_F + \delta_F], [a_C, a_C + \delta_C], [a_N, a_N + \delta_N]\). These intervals will be called arrival windows and their length \( \delta \) the arrival spread. Their lower limit \( a \) will be referred to as earliest possible arrival time. Note that \( a \) is earlier than the scheduled time, the expected time of arrival at the transfer point.
We further assume: (1) \( \delta_R \leq (1/2) h_R \) and (2) \( \delta_F \leq (1/2) h_R \) where \( h_R = a_N - a_C \) is the scheduled headway on the receiving line \( L_R \). Conditions (1) and (2) are easily satisfied for a route in the off-peak period. It will later be seen that this is the time-of-day most promising for transfer optimization. Furthermore, all the numerical examples of this paper are consistent with (1) and (2).

Conditions (1) and (2) ensure that \( N \) cannot overtake \( C \) and that passengers transferring from \( F \) to \( L_R \) will always board either \( C \) or \( N \). The latter implies the waiting time for the transfer will depend only on \( t_C \), \( t_N \), and \( t_F \), and allows us to write the mean disutility function as: \( D = E[g(w(t_C, t_N, t_F))]. \) The components of this function, in more detail, are as follows.

The wait function \( w(t_C, t_N, t_F) \) specifies the waiting time for the transfer connection associated with given values of the arrival times \( t_C, t_N \) and \( t_F \). It can be used to model various policies for holding buses at the transfer point. For example in the case of a no-hold policy (i.e., buses depart as soon as they arrive at the transfer point), the wait function would be:

\[
w(t_C, t_N, t_F) = \begin{cases} 
  t_C - t_F & \text{if } t_C \geq t_F \\
  t_N - t_F & \text{otherwise}
\end{cases}
\]

This is equivalent to the wait function defined by Grega and Theede[6] and implicitly assumed by Rapp and Gehrner,[11] Klem and Stemme[9] and Andresson.[2] As a second example of a holding policy, consider a case with scheduled departure times \( d_C \) and \( d_N \) for \( C \) and \( N \). These trips will depart immediately if they are late, but at their scheduled departure time otherwise. The corresponding wait function is:

\[
w(t_C, t_N, t_F) = \begin{cases} 
  d_C - t_F & \text{if } t_C \leq d_C, t_F \leq d_C \\
  t_C - t_F & \text{if } t_C > d_C, t_F \geq d_C \\
  d_N - t_F & \text{if } t_F > t_C, t_F > d_C, t_N \leq d_N \\
  t_N - t_F & \text{otherwise}
\end{cases}
\]

Note that these two wait functions satisfy the property (3): \( w(t_C, t_N, t_F) = w(t_C + z, t_N + z, t_F + z) \) for all real \( z \), and thus waiting time only depends on the relative position of the arrival time of \( F, C, \) and \( N \). Most wait functions arising in a real world context would be expected to satisfy that property. Other examples of holding policies with their corresponding wait functions are presented in Désilets.[5]

The disutility function \( g(w) \) gives the desirability of a waiting time of \( w \), as perceived by the user. A common choice of \( g \) in the literature is \( g(w) = w \) (Rapp and Gehrner,[11] Andresson,[2] Abkowitz et al.,[1] Grega and Theede[6]). In this case, mean disutility is just mean waiting time. In spite of its attractiveness, mean waiting time can be a misleading measure of transfer inconvenience, because it penalizes long and short waiting times with the same weight, and does not take into account reliability of the transfer connection. For example, it would not differentiate between a situation where waiting time is 0 or 20 minutes with a 50% probability each, from a situation with a guaranteed waiting time of 10 minutes. Users may have a preference for the second alternative whose waiting time is reliable and never extreme.

A disutility function which penalizes long waiting times with more weight than short ones is \( g(w) = w^2 \). Another example is \( g(w) = 1 \) if \( w > w^* \), and 0 otherwise, where \( w^* \) is a waiting time considered to be comfortable. Mean disutility in this latter case is the probability that waiting time is greater than \( w^* \). A disutility function which takes reliability into account is the variance of waiting time \( \text{Var}(w) \). It cannot be expressed directly as \( D = E(g(w)) \), but when viewed as the difference \( \text{Var}(w) = E(w^2) - (E(w))^2 \), both \( E(w^2) \) and \( E(w) \) have the required form.

The mean disutility of a connection is the expectation of the disutility function over all values of \( t_F, t_C \) and \( t_N \), and will thus depend on which distribution is used to model these arrival times. Jenkins[7] has surveyed various distributions used in simulating bus operations. However, our assumption that arrival times take their values on a finite interval makes most of these inappropriate. In the present paper, we thus use a variant of the exponential distribution, which we call the Shifted Truncated Exponential (STE). It has the density function:

\[
f_{a,b}(t) = \begin{cases} 
  (K \lambda) \exp[-\lambda(t - a)] & \text{if } t \in [a, a + \delta] \\
  0 & \text{otherwise}
\end{cases}
\]

where: \( \lambda = -\ln(1 - 0.95)/\delta \approx 3/\delta \) is the parameter of an exponential random variable having 95% of its distribution within an interval of length \( \delta \), and \( K = 1/0.95 \) is a factor ensuring that \( P(t \in [a, a + \delta]) = 1 \). Conceptually, the STE corresponds to an exponential distribution with 95% of its values falling in the interval \([0, \delta]\), which has been translated by an amount \( a \), and whose right tail has been truncated at \( a + \delta \) and redistributed proportionately on the interval \([a, a + \delta]\).

An advantage of the STE distribution is that a bus, say on the feeder line, can be late by a greater amount than it can be early: \( a_F + \delta_F - E(t_F) >
One can easily verify that the STE density function satisfies property (4): \( f_{a, \delta}(t) = f_{a, + \delta}(t + x) \) for all \( x \in R \). This property states that shifting the time at which a trip leaves its terminal (and thus its earliest possible arrival time \( a \)) by a time \( x \), amounts to shifting the density of arrival times by the same time \( x \). We can now write the mean disutility function as:

\[
D = E\left[ g\left( w(t_C, t_N, t_P) \right) \right]
\]

\[
= \int_{t_C = -\infty}^{\infty} \int_{t_N = -\infty}^{\infty} \int_{t_P = -\infty}^{\infty} g\left( w(t_C, t_N, t_P) \right) \times f_{a, \delta}(t_C) f_{a+C, \delta}(t_N) f_{a, \delta}(t_P) \, dt_C \, dt_N \, dt_P
\]

which is completely determined by the value of the parameters \( a, a+C, \delta \) and \( h, \delta \). An interesting property of the mean disutility function is that if the wait and density functions satisfy properties (3) and (4), its value does not depend on the actual values of \( a \) and \( a+C \) but on their positions relative to one another. This is formalized by:

**Theorem 1.** Let a disutility function \( g \), a wait function \( w \) satisfying property (3), a family of distributions \( f_{a, \delta} \) satisfying property (4) and values of the parameters \( h, \delta \) be given. For any values \( a, a+C \) of the earliest arrival times of \( F \) and \( C \), let \( D(a, a+C) \) be the corresponding mean disutility. Then for any real number \( z \):

\[
D(a + z, a+C + z) = D(a, a+C).
\]

**Proof.** For the given values of the parameters \( h, \delta \) and \( a \), and any values \( x \) and \( y \) let:

\[
A_{x, y}(t_C, t_N, t_P) = g\left( w(t_C, t_N, t_P) \right) \times f_{x, \delta}(t_C) f_{x+h, \delta}(t_N) f_{y, \delta}(t_P).
\]

Properties (3) and (4) imply that for any real number \( z \) we have:

\[
A_{x+z, y+z}(t_C + z, t_N + z, t_P + z) = A_{x, y}(t_C, t_N, t_P).
\]

Thus for any \( z \), we can write:

\[
D(a + z, a+C + z) = \int_{t_C = -\infty}^{\infty} \int_{t_N = -\infty}^{\infty} \int_{t_P = -\infty}^{\infty} x A_{a, a+C, \delta}(t_C, t_N, t_P) \, dt_C \, dt_N \, dt_P
\]

Finding an analytic form for \( D \) can be difficult and lead to very complex expressions, even with simple wait, disutility and density functions (Désilets[6] Grega and Thedeen[6]). Considering the great variety of choices for \( D \) and the fact that any of these can be of interest, a simulation approach seems more appropriate. With this approach, the values of different mean disutility functions can be estimated simply by changing a few lines of code. Désilets[6] describes such a simulation program, which uses a control variate technique (LAW and KELTON[10]) to rapidly obtain accurate estimates of \( D \).

### 1.2. The Integer Programming Model

The mean disutility function defined in the preceding section measures the inconvenience of a single transfer connection. But for the purpose of transfer optimization, we need an objective function which measures the aggregate inconvenience of all transfers occurring in a transit network under a certain timetable \( T \). We define such a function in the following way. Let the possible transfer connections occurring under \( T \) be indexed by the set \( \{1, 2, \ldots, M\} \). For each such connection \( k \), let \( D_k(T) \) be its mean disutility under \( T \) as defined in Section 1.1. Let \( n_k \) denote its transfer flow, i.e. the number of users making the connection, or some other factor reflecting the importance of transfer connection \( k \). Then \( \mu(T) = C(T) / n \) is the mean disutility for a passenger transferring during the operation of \( T \), where \( C(T) = \sum_{k=1}^{M} n_k D_k(T) \) and \( n = \sum_{k=1}^{M} n_k \). Minimizing \( C(T) \), will be equivalent to minimizing \( \mu(T) \), assuming that transfer flows are not affected by the timetable (Note that changing the timetable may well change the attractiveness of certain transfer connections, and thus affect transfer flows, but this simplifying assumption does allow a more tractable objective function).

We can now define the problem of finding a timetable \( T \) which minimizes \( C(T) \), using the integer programming model proposed by Kletl and
Stemme.\cite{Stemme69} This model assumes a set of lines $i = 1, \ldots, N$, each with a finite set $T_i$ of feasible timetables. Assuming constant and given headway $h_i$ ($i = 1, \ldots, N$), a timetable for each line $i$ is totally determined by the departure from the terminal, of the first bus on this line during the time period under consideration (Following Rapp and Gehner,\cite{Rapp11} we call this departure time the offset time of line $i$). One may thus think of $T_i$ as the finite set of all feasible offset times for line $i$.

We can then construct a timetable $T$ for the whole network by choosing offset times $t_i \in T_i$ ($i = 1, \ldots, N$) for each line; we will therefore represent a timetable by its $N$-tuple of offset times $T = (t_1, \ldots, t_N)$. For each pair $\{i, j\}$ of lines let $A_{ij} = \{k: \text{connection } k \text{ goes from line } i \text{ to line } j\}$. For every $r \in T_i$ and $s \in T_j$, let $C_{ijrs} = \sum_{k \in A_{ij}} h_k D_k(r, s)$, where $D_k(r, s)$ is the mean disutility of connection $k$, given the choices $r$ and $s$ of offset times for lines $i$ and $j$ respectively. Note that if lines $i$ and $j$ do not intersect, $A_{ij}$ is empty and thus $C_{ijrs} = 0$.

We can then express our optimization problem (TOP) as:

$$\text{minimize } C(T) = \sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij} t_i t_j \quad \text{(TOP)}$$

subject to $t_i \in T_i$ ($i = 1, \ldots, N$).

This is equivalent to the 0-1 quadratic optimization formulation (TOP') proposed by Klemt and Stemme:\cite{Stemme69}:

$$\text{minimize } C(T) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{r \in T_i} \sum_{s \in T_j} C_{ijrs} x_{ir} x_{js}$$

subject to $\sum_{r \in T_i} x_{ir} = 1$ ($i = 1, \ldots, N$) \quad \text{(TOP')}$

$x_{ir} \in \{0, 1\}$ ($i = 1, \ldots, N; t \in T_i$).

The binary variable $x_{ir}$ takes the value 1 if and only if offset time $s$ is chosen for link $i$; equality constraints ensure that one and only one of the feasible offset times is chosen for every line.

The use of an integer program, a deterministic optimization tool, may seem to contradict our goal of taking randomness of travel times into account. In fact, the cost coefficients $C_{ijrs}$ incorporate all the impact of random travel times on mean disutility. The stochastic nature of travel times does not affect the method by which we obtain the optimal timetable from the finite set of feasible timetables, but only the way in which we cost them.

The optimization problem (TOP) can be viewed as a relaxed Quadratic Assignment Problem (Désilets\cite{Desilets81}) and is NP-complete (Désilets,\cite{Desilets81} Cameron\cite{Cameron49}). There is thus little hope of optimally solving it for real-life instances. To obtain heuristic solutions, we used the same iterative improvement procedure as Rapp and Gehner,\cite{Rapp11} which can be described as follows:

0. Chose an arbitrary feasible timetable $T = (t_1, \ldots, t_N)$.

1. Repeat
   1.1 $T_0 := T$  
   1.2 For $i := 1$ to $N$ do
      1.2.1 Let $t^* \in T_i$ be such that changing the offset time of line $i$ from its current value to $t^*$ would produce the most decrease in $C(T)$.
      1.2.2 Let $T_i := (t_1, \ldots, t_{i-1}, t^*, t_{i+1}, \ldots, t_N)$
   Until ($T_0 = T$).

This procedure starts with an initial solution and looks for improvements by changing the offset time of one link at a time, until no further improvement can be obtained. In Désilets,\cite{Desilets81} this heuristic is generalized to a $K$-interchange algorithm which looks for improvements by simultaneously changing the offset times of $K > 1$ lines. Computational tests showed however that values $K > 1$, while increasing the computation time, do not guarantee a better solution.

2. TRANSFER OPTIMIZATION UNDER RANDOM BUS TRAVEL TIMES

2.1. Optimization of a Single Connection

In the next two sections, we use the tools developed in Section 1, to study the optimization of transfers under random travel times. The present section looks at the single connection case, whereas Section 2.2 looks at the optimization of many connections in a network. Recall that the decision variables of our transfer optimization model are the offset times, which in turn determine the values of $r_{FC}$ (the relative earliest arrival times of the feeder and critical buses) for the various connections occurring in the network. Therefore it is of interest to study how $D$, the mean disutility of a single connection, behaves as we change the value of $r_{FC}$. We will do so by studying mean disutility curves which plot $D$ as a function of $r_{FC}$. Our analysis will be limited to cases where there is no holding of buses at the transfer point.

Figure 2 shows mean disutility curves for the four mean disutility functions: $D = E(w)$, $D = E(w^2)$, $D = Var(w)$ and $D = P(w > w^*)$. These
were visually fitted on a scatter graph obtained by running the simulation program of Désilets[6] for values $r_{FC} = -10, -9, \ldots, 9, 10$ minutes. All curves assumed STÉ arrivals and values of the simulation parameters: $\delta_F = \delta_R = \delta = 4$ minutes and $h_R = 20$ minutes (for simplicity, we assumed the same arrival spread $\delta$ on both lines). Although the curves pertain to specific values of $r_{FC}, \delta_F, \delta_R$ and $h_R$, their shapes reflect the general behavior of mean disutility for other choices of these parameters.

The four curves have interesting features in common. They all attain their minimum around the value $r_{FC} = -\delta = -4$. With the exception of $\text{Var}(w)$, which is maximum at $r_{FC} = 0$, all curves attain their maximum around the value $r_{FC} = \delta = 4$. This suggests the following rule of thumb. It is a good idea, in minimizing mean disutility of a transfer connection under a no-hold policy, to schedule the feeder bus to arrive as close as possible to the critical bus while virtually ensuring that it will catch it. This principle seems valid independently of the mean disutility function used. Conversely, it is a bad idea to schedule $F$ to arrive as far from $N$ as possible, yet with no chance of making the connection with $C$. These rules are of course very intuitive and come as no surprise.

Figure 2 also reveals that the value of $D$ at $r_{FC} = 0$ is at best halfway between the maximum and minimum value of $D$ (in the cases $D = E(w)$ and $D = P(w > w^*)$), and at worst equal to the maximum value of $D$ (in the case $D = \text{var}(w)$). But under the assumption of deterministic travel times (i.e., $\delta_F = \delta_R = 0$), $r_{FC} = 0$ is optimum because it ensures waiting times of zero. This suggests that, if travel times in the actual transit network are random, optimizing a transfer connection under the assumption of deterministic travel times may lead to a poor connection, from the point of view of many disutility functions.

Figure 3 compares disutility curves in a case where arrival times are random, and a case where they are assumed deterministic and equal to the earliest possible arrival times of the random case. The two cases seem to differ significantly only in the interval $[-\delta, \delta] = [-4, 4]$, where $F$ has a probability $p$, $0 < p < 1$, of missing $C$. Therefore for a single transfer connection, if travel times in the actual network are random, the deterministic-travel-time assumption is misleading only for $r_{FC} \in [-\delta, \delta]$. Unfortunately, in optimizing transfers with the deterministic assumption, one is aiming at values around $r_{FC} = 0$.

Figure 4 compares mean disutility curves for $\delta = 8$ and $\delta = 4$. We find that as $\delta$ increases, the mean disutility curve becomes flatter. The minimum value of $D$ increases because for larger $\delta$, $F$ cannot get as close to $C$ and still ensure that it will catch it. Similarly, the maximum value of $D$ is smaller because $F$ cannot get as far from $N$, yet still have no chance of making the connection with $C$. These observations imply that the difference between the value of $D$ at the optimum, and at a randomly picked $r_{FC}$, will be smaller for larger values of $\delta$. Therefore the potential benefits of transfer optimization become smaller as travel times become more and more random.

Figure 5 illustrates the effect of a doubling in headway. With the exception of $D = P(w > w^*)$, increasing the headway drastically increases the range between the minimum and maximum values of $D$. Therefore the potential benefits of transfer optimization are greater if the headway on the receiving line is large. Comparing the corresponding...
Fig. 3. Mean disutility curves under random (solid line) and deterministic (dashed line) travel times.

Fig. 4. Effect of doubling the arrival spread on each line: mean disutility curves under small (solid line) and large (dashed line) arrival spreads.

For example, a value of $r_{FC}$ which ensures that $F$ will barely catch $C$, corresponds to a situation where $C$ is sure to barely miss $F$.

Figure 6 compares the combined mean disutility of transfers from $F$ to $L_R$ and from $C$ to $L_F$, in a directional and a non-directional case. In the directional case, we assumed there was one passenger transferring from $F$ to $L_R$ and none from $C$ to $L_F$, and in the non-directional case, we assumed one passenger was making each of the connections. Both cases assume equal headway on the two lines. For $D = E(w)$, mean disutility in the non-directional situation stays almost constant at 10 minutes. This is due to the fact that independently of the values of $t_F, t_C, t_N$, the sum of waiting times for the connections $F$ to $L_R$ and $C$ to $L_F$ will
approximately equal the headway of 20 minutes. This is formalized by:

**Theorem 2.** Let $w_{FR}$ and $w_{CF}$ be the waiting times under a no-hold policy, for the transfer connection from $F$ to $L_R$ and from $C$ to $L_F$ respectively. Let $h$ be the headway both on $L_F$ and $L_R$ and suppose the density functions of the arrival times of all buses satisfy property (4). Then:

$$h \leq E(w_{FR}) + E(w_{CF}) \leq h + K,$$

where

$$K = \max\{E(t_P) - a_P, E(t_C) - a_C\}.$$

This is proved in the Appendix, where we also prove:

**Corollary 2.1.** Let $\mu$ be the combined expected waiting time of transfers from $F$ to $L_R$ and from $C$ to $L_F$. If the transfer flows in both directions are equal, and the headway on each line is $h$ then:

$$h/2 \leq \mu \leq h/2 + K/2.$$

This corollary implies that when transfer flows and headways are equal, then $K/2$ is the maximum gain in combined mean waiting time that should be expected from a change in $r_{FC}$. But if arrival spreads are reasonably small, $K/2$ may be negligible. For example, if $\delta_P = \delta_R = 4$ min and if the distribution of arrival times is symmetric, then $K/2 = (0.5)(4)/2 = 1$ min. Thus, mean waiting time (traditionally used as the objective function
of transfer optimization (Rapp and Gehner, Andreasson, Klemt and Stemme, Keudel) may not be appropriate when headways are almost equal and the directionality of transfers is low. Such may be true of many situations where transfer optimization might be used, such as off-peak or weekend periods.

The mean disutility curves for \( D = E(W^2) \) and \( D = P(w > w^*) \) are not constant in the non-directional case, but the difference between the maximum and minimum values of \( D \) is smaller than in the directional case. Variance of the waiting time \( \text{Var}(w) \) remains unaffected by the directionality of transfers. This was to be expected, since the \( D = \text{Var}(w) \) curves for a single connection from \( F \) to \( L_C \) and for a single connection from \( C \) to \( L_R \) are the same, and each is symmetric. Since we have \( r_{FC} = -r_{CF} \) (where \( r_{CF} = a_C - a_F \) is the relative earliest arrival time of \( C \) with respect to \( F \)), the variance of the connection from \( F \) to \( L_R \) is always equal to that of the connection from \( C \) to \( L_F \). This means that values of \( r_{FC} \) which produce connections of low variance in one direction also produce connections of low variance in the opposite direction.

### 2.2. Simultaneous Optimization of Many Connections in a Network

In the optimization of a single transfer connection, we were free to choose any value for \( r_{FC} \). In particular, it was always possible to choose that value which minimizes mean disutility \( D \). This is not true when we are trying to simultaneously optimize many connections at various transfer points in a transit network, since in that case, timetables which yield optimal \( r_{FC} \) for some connections may yield nonoptimal \( r_{FC} \) for others. Such trade-offs make optimization of many connections much more complex, and thus some observations made for a single-connection case may no longer hold.

For example, a timetable which minimizes total waiting time under the assumption of deterministic travel times may not yield values of \( r_{FC} \) around zero (which is optimal when travel times are deterministic, but average when they are random) for many of the transfer connections. Since the values of \( D \) under deterministic and random travel times tend to differ only for values of \( r_{FC} \) close to 0, this deterministically optimal timetable might also produce good connections when travel times are random. The negative consequences of assuming deterministic travel times may thus be smaller when optimizing many connections in a large network. Another consequence of these trade-offs is that the difference between the disutilities of the optimal and worst timetables may be smaller than for the optimization of a single connection. This is because in the case of many connections, the optimal and worst timetables do not produce optimal and worst conditions for every connection.

We will now use the integer programming model (TOP) to investigate empirically this more complex situation. More precisely, we will use the model on several instances representing conditions where the gains of transfer optimization would, according to our analysis of the single connection case, be large versus small. We will also run the model on these cases assuming deterministic and then random travel times, and compare the resulting solutions. Again, we will restrict ourselves to a no-hold policy.

All instances studied are based on two networks: WIN and MAN (see Fig. 7). The first is derived from data for the weekday-morning peak of the city of Winnipeg, while the second was created by the authors. Four other networks were obtained from these two by changing some parameters. WEASY and MEASY were respectively obtained from WIN and MAN by increasing headways, decreasing arrival spreads and making transfers more directional. According to our analysis of the single connection case, they represent situations where optimization of transfers could produce the most improvement over a randomly picked timetable. Similarly, WHARD and MHARD were obtained by decreasing headways to the same value, increasing arrival spreads and equalizing transfer flows in opposite directions. Under these conditions, little could apparently be gained by optimizing transfers.

Optimization of transfers for these six networks was done according to the three mean disutility functions: \( E(w) \), \( E(w^2) \) and \( \text{Var}(w) \). The resulting 18 instances were solved heuristically using the iterative improvement procedure described in Section 1.2, with five different randomly generated initial solutions. In each case, the average percent improvement (between the disutilities of the final and initial solutions) was computed for the three mean disutility measures. Table I suggests a number of interesting trends. One observes that headways, arrival spreads and directionality of transfers have a striking influence on the benefits of transfer optimization. This is especially true for the optimization of \( E(w) \), where the mean percentage improvement varied from 2% under the “worst” conditions to 27.6% for the “best” conditions. In the case of \( \text{Var}(w) \), however, significant improvements were obtained in all instances, even though they were greater for certain values of the above
mentioned parameters. Thus, it seems that whereas mean waiting time is sometimes impossible to control, variance of waiting time can always be decreased through transfer optimization.

Another interesting observation is that optimization of $E(w)$ and of $E(w^2)$ are consistent with each other, but not necessarily with the optimization of $\text{Var}(w)$. For all networks, optimizing $E(w^2)$ also produced improvements in $E(w)$ comparable to those obtained by optimizing $E(w)$ itself. Similarly, with the exception of WHARD, optimization of $E(w)$ produced improvements in $E(w^2)$ of the same order as those obtained by optimizing $E(w^2)$ itself. However, optimizing $E(w)$ or $E(w^2)$ sometimes increased and other times decreased $\text{Var}(w)$; optimizing $\text{Var}(w)$ occasionally decreased both $E(w)$ and $E(w^2)$, but sometimes increased the value of one or both of them. This compatibility between the optimization of $E(w)$ and $E(w^2)$ may be explained by the similarity of their mean disutility curves, which is apparent from Figure 2.

Next, we study the negative consequences of the deterministic travel-time assumption in a multi-connection context. We do so by "optimizing" $E(w)$ and $E(w^2)$ for each of the six networks with five different initial solutions, while assuming deterministic travel times. We then compute the average percent change between these initial and final solutions, when travel times are random. Using the same five starting solutions, we then optimize $E(w)$ and $E(w^2)$ while taking randomness of travel times into account, and again compute the average percent change between initial and final solutions. Table II summarizes the results. In all cases, the improvements obtained by optimizing probabilistically (taking randomness of travel times into account) were greater than those obtained by optimizing deterministically. In fact, optimizing $E(w)$ deterministically even produced, for MHard and WHARD, a final solution worse than the initial one when mean disutility was computed under random travel times.

As mentioned previously, optimizing $E(w)$ or $E(w^2)$ sometimes increases the value of $\text{Var}(w)$, even if randomness of travel times is taken into account. But if randomness of travel times is not
taken into account, this increase always occurs. Table III gives the percentage change in the value of \( \text{Var}(w) \) when \( E(w) \) and \( E(w^2) \) were optimized in a deterministic versus a probabilistic way. In all cases, deterministically optimizing \( E(w) \) and \( E(w^2) \) caused an increase in \( \text{Var}(w) \). Moreover, even in cases where probabilistic optimization also caused an increase of \( \text{Var}(w) \), this increase was smaller than the one caused by deterministic optimization, except in two instances (optimization of \( E(w^2) \) for MAN and MEASY).

3. SUMMARY AND AREAS FOR FUTURE RESEARCH

In this paper, we constructed a model for transfer optimization in a transit network which improves upon existing models in two ways: (a) it takes randomness of bus travel times into account and (b) it can optimize transfers according to various objective functions and under various holding policies. This model constitutes an important contribution both theoretically and practically. From a theoretical point of view, it allows one to investigate the way in which randomness of bus travel times affects transfer optimization. From a practical point of view, it is a more accurate and versatile tool which could be used for transfer optimization in actual transit properties.

Table II

<table>
<thead>
<tr>
<th>Network</th>
<th>( E(w) )</th>
<th>( E(w^2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIN</td>
<td>-14.5</td>
<td>-20.5</td>
</tr>
<tr>
<td>MAN</td>
<td>-0.8</td>
<td>-2.0</td>
</tr>
<tr>
<td>WEASY</td>
<td>-8.4</td>
<td>-23.1</td>
</tr>
<tr>
<td>MEASY</td>
<td>-8.8</td>
<td>-27.6</td>
</tr>
<tr>
<td>WHARD</td>
<td>+2.4</td>
<td>-3.0</td>
</tr>
<tr>
<td>MWHARD</td>
<td>+0.3</td>
<td>-2.0</td>
</tr>
</tbody>
</table>

Percent changes were computed under the assumption of random arrivals, even when optimization was done deterministically. A plus sign represents an increase over the initial solution.

Table III

<table>
<thead>
<tr>
<th>Network</th>
<th>( E(w) )</th>
<th>( E(w^2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIN</td>
<td>+70.5</td>
<td>+33.6</td>
</tr>
<tr>
<td>MAN</td>
<td>+17.8</td>
<td>-9.3</td>
</tr>
<tr>
<td>WHARD</td>
<td>+112.6</td>
<td>-35.5</td>
</tr>
<tr>
<td>MWHARD</td>
<td>+43.1</td>
<td>-42.9</td>
</tr>
<tr>
<td>WEASY</td>
<td>+263.4</td>
<td>-33.7</td>
</tr>
<tr>
<td>MEASY</td>
<td>+46.2</td>
<td>+9.1</td>
</tr>
</tbody>
</table>

We then used the model to study the optimization of transfers under random bus travel times and no-hold policies, according to various objective functions (\( E(w) \), \( E(w^2) \), \( \text{Var}(w) \) and \( P(w > w^*) \)). We confirmed the intuitive statement that transfer optimization under a no-hold policy may produce more improvement when headways are long, buses are on time and transfers are directional. We also observed that whereas it was sometimes difficult to improve the value of \( E(w) \) and \( E(w^2) \) through transfer optimization, significant improvements could always be obtained for \( \text{Var}(w) \).

The preceding suggests that, of the several objectives employed, \( \text{Var}(w) \) is the most sensitive to offset times and is hence more successful at differentiating between solutions. This need not imply, however, that \( \text{Var}(w) \) is a more suitable objective function than the traditional choice \( E(w) \). The perceived disutilities of groups of passengers should probably reflect a combination of expected wait and variance, or of squared expected wait and variance. This may mean that \( E(w^2) \) is the best single proxy as a plausible disutility function. In any case, the analyses of this paper have shown the advantages of a model capable of optimizing transfers according to several objective functions.

Finally, we demonstrated some negative consequences of assuming deterministic bus travel times in optimizing transfers, if these travel times are in fact random. For example, optimizing a single transfer connection under such an assumption yields a mean disutility which is worse than average under many disutility measures, and even maximum under the measure \( D = \text{Var}(w) \). Such negative outcomes also occur to a lesser extent in the simultaneous optimization of many connections. In some exceptional cases, optimizing many connections with the deterministic-travel-time assumption can even produce a final solution worse than the initial one, and in most cases, it produces an improvement lower than that obtained if randomness of travel times is taken into account. These results show the advantages of our probabilistic model over existing deterministic ones (Rapp and Gehner, [11] Andreasson, [2] Klemt and Stemme [5] and Keudel [6]).

There are several areas of research still open. All calculations performed in Section 2 assumed a no-hold policy, even though other holding policies can be accommodated by our model. It would be worth investigating whether alternative policies lead to better transfer optimization.

We remark that a transit vehicle is often considered "on time" if it leaves a given check point within the interval (0 minute early, 5 minutes...
late). Our model, with its bounded-arrival-time distribution, could be used to model transfer impacts when buses run within their allowable window. This could help quantify the benefits of setting and complying with different on-time standards.

There are also several ways in which the model could be enhanced. For example, the present model assumes constant headway along a given line. This allows a timetable to be described by only two pieces of data, the headway and the offset time. Alternatively, one might assume that a timetable is described by the offset time, and a predefined sequence of headways between consecutive trips on the line. Also, the only control variables of the present model are the offset times, the departure times of all buses being determined by these variables. Instead, one could allow the departure time of a trip $i$ on a line to vary within a small time interval $[d_i(t), D_i(t)]$, determined by the chosen offset time $t$. In the iterative improvement procedure one could then change the departure times, not only of the first trip on each line, but of every trip. This greater flexibility of the model might lead to better transfers, but would of course make it less tractable.

From an algorithmic point of view, one might look for more efficient heuristics for solving (TOP). A possible approach is Simulated Annealing, which has been successfully applied to the Quadratic Assignment Problem (QAP) by Burkard and Rendl.[31] Simulated annealing is a non-greedy search procedure, based on analogy with the transition to thermodynamic equilibrium in physics. Similarities between the QAP and (TOP) suggest this method may also be effective for our problem. The procedures used to implement our iterative improvement algorithm could easily be modified to implement such an annealing approach.

**APPENDIX: PROOF OF THEOREM 2 AND COROLLARY 2.1**

Let $O$ be the next trip on $L_R$ after $F$, and $t_O$ its arrival time at the transfer point. We first prove the following:

**Lemma.** (i) $K \leq E(t_O - t_F | t_C \geq t_F) \leq h + K$ and (ii) $K \leq E(t_N - t_C | t_C < t_F) \leq h + K$.

**Proof.** We only prove (i), the proof of (ii) being similar. By the linearity of expectations we have:

$$E(t_O - t_F | t_C \geq t_F) = E(t_O | t_C \geq t_F) - E(t_F | t_C \geq t_F) = E(t_O) - E(t_F | t_C \geq t_F).$$

But

$$E(t_O) = \int_{t_- - \infty}^{t_+ \infty} t f_{a_p, h, i_p}(t) \, dt$$

$$= \int_{t_- - \infty}^{t_+ \infty} t f_{a_p, i_p}(t - h) \, dt$$

$$= \int_{t_- - \infty}^{t_+ \infty} (t' + h) f_{a_p, i_p}(t') \, dt'$$

(by property (4)) (where $t' = t - h$)

$$= E(t_F + h) = E(t_F) + h$$

since $f_{a_p, i_p}(t')$ is the density function of $t_F$.

Thus we have

$$E(t_O - t_F | t_C \geq t_F) = E(t_F) + h - E(t_F | t_C \geq t_F).$$

But since $t_F$ is always greater than $a_p$, we have

$$E(t_F) + h - E(t_F | t_C \geq t_F) < E(t_F) + h - a_c \leq K + h.$$

Finally, since $E(t_F | t_C \geq t_F) \leq E(t_F)$ (in the first term, $t_F$ is constrained to values smaller than $t_C$), we also obtain

$$E(t_O) - E(t_F | t_C \geq t_F) \geq E(t_O) - E(t_F) = h.$$

**Proof of Theorem 2**

We have, with $p = P(t_C \geq t_F)$,

$$E(w_{FC} + w_{CF})$$

$$= E(w_{FC} + w_{CF} | t_C \geq t_F) p$$

$$+ E(w_{FC} + w_{CF} | t_C < t_F) (1 - p)$$

$$= E[(t_C - t_F) + (t_O - t_C | t_C \geq t_F) p$$

$$+ E[(t_N - t_F) + (t_F - t_C | t_C \geq t_F) (1 - p)$$

$$= E[t_O - t_F | t_C \geq t_F] p$$

$$+ E[(t_N - t_C) | t_C \geq t_F] (1 - p)$$

and thus, by the above lemma,

$$h \leq E(w_{FC} + w_{CF}) \leq h + K.$$

**Proof of Corollary 2.1**

Let $n$ be the number of passengers making the connection in each direction. The mean waiting time for both directions combined is then

$$\mu = E[(nw_{FC} + nw_{CF})/2n] = E[(w_{FC} + w_{CF})/2]$$

$$= [E(w_{FC} + w_{CF})]/2.$$
But by Theorem 2, \( h/2 \leq (E(w_{PC} + w_{CF})/2 \leq (h + K)/2 \) which completes the proof.

ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada, through Operating Grant OGP5292 and a Graduate Scholarship. We are also grateful to Professor Jean-Marc Rousseau for his valuable input throughout this research. Thanks are due the anonymous referees and Associate Editor for their insightful comments, several of which have been incorporated almost directly into the manuscript.

REFERENCES


(Received, July 1989; revisions received November 1989, April 1991; accepted November 1990)