TRANSPORTATION SCIENCE
Vol. 42, No. 4, November 2008, pp. 466–477
ISSN 0041-1655 | EISSN 1526-5447 | 08 | 4204 | 0466

informs

DOI 10.1287/trsc.1080.0245

A Rolling Horizon Model to Optimize Aerodynamic Efficiency of Intermodal Freight Trains with Uncertainty

Yung-Cheng Lai
Department of Civil Engineering, National Taiwan University, Taipei, Taiwan 10617, yclai@ntu.edu.tw

Yanfeng Ouyang, Christopher P. L. Barkan
Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801 {yfouyang@uiuc.edu, cbarkan@uiuc.edu}

Aerodynamic efficiency of intermodal freight trains can be significantly improved by minimizing the adjusted gap lengths between adjacent loads. This paper first develops a static model to optimize load placement on a sequence of intermodal trains that have scheduled departure times. This model applies when full information on all trains and loads is available. Then, we develop a dynamic model to account for the more realistic situation in which there is incomplete or uncertain information on future trains and incoming loads. This paper develops methodology to balance between (i) the advantage from optimizing multiple trains together and (ii) the risk of making suboptimal decisions due to incomplete future information. We propose a rolling horizon scheme to address this challenge, where exponentially decreasing weights are assigned to the objective functions of future trains. Numerical results based on empirical data show significant aerodynamic efficiency benefits from these optimization models.

Key words: rail transportation; intermodal freight; energy efficiency; emissions; uncertainty
History: Received: November 2006; revision received: October 2007; accepted: May 2008.

1. Introduction

Intermodal (IM) freight transportation has experienced rapid growth (Gallamore 1998) and recently replaced coal as the largest source of revenue for U.S. railroads (AAR 2004–2006). However, IM trains are generally the least fuel efficient trains due to the physical constraints imposed by load configuration, placement, and railcar design (Engdahl, Gielow, and Paul 1987). This is particularly ironic given that IM trains are typically the fastest freight trains operated in North America. However, there are considerable opportunities to reduce the aerodynamic penalties when IM trains are loaded (Lai and Barkan 2005; Lai, Barkan, and Önal 2008).

At IM terminals, containers and trailers (i.e., IM loads) are assigned to available well, spine, or flat cars (TTX 1999; UP 2005). The assignment of loads is largely a manual process even though computer software (Optimization Alternatives 2005) is often used by terminal managers to assist the process. Railroads provide incentives to terminal managers for maximizing slot utilization (i.e., using all feasible slots available where a "slot" is defined as the space to accommodate IM loads in an IM railcar). However, these incentives do not take into account the size of the slot compared to the size of the load. Perfect slot utilization is not intended to, nor does it, ensure that IM cars are loaded to maximize the energy efficiency of IM trains. For example, two trains may have identical slot utilization but different loading patterns and subsequent differences in train resistance (Lai and Barkan 2005). Consequently, there is a gap between slot utilization and energy efficiency.

Past efforts on IM train loading assignment have focused on maximizing the utilization of trailer hitches (Feo and Gonzales-Velrade 1995), optimizing the flow of flat cars in the network (Powell and Carvalho 1998), and minimizing excess handling time and optimizing the mass distribution of the train (Corry and Kozan 2006). Each of these studies focused on certain types of IM loads or railcars. We are unaware of any previous studies other than our prior work (Lai, Barkan, and Önal 2008) that have considered optimization of the energy efficiency of IM train loading patterns.

Lai and Barkan (2005) demonstrated that aerodynamic characteristics significantly affect IM train fuel efficiency; a train can be more efficiently operated if loads and slots are carefully matched during the process of load assignments. To help terminal managers assemble more fuel efficient trains, Lai, Barkan, and Önal (2008) developed an aerodynamic loading

assignment model (ALAM) in which the objective was to maximize aerodynamic efficiency (by minimizing the adjusted gap length) of the outgoing IM freight train given any particular static combination of loads and railcar types. Their analysis of one major railroad IM route revealed the potential to reduce fuel consumption by 15 million gallons per year, with a corresponding cost savings opportunity of $20 million.

Lai, Barkan, and Önal's earlier model (2008) is based on current terminal practices and considers optimization of the loading pattern of a single train at a time. However, if advance information on outgoing trains and loads is available, a better loading plan will often be possible by simultaneously considering more than one train. The larger pool of loads and railcars will enable better matching but may also introduce greater uncertainty about the composition of the future load pool. In this paper, we extend ALAM to optimize the aerodynamic efficiency for a system of multiple trains. First, the benefit from optimizing multiple trains and loads is evaluated, assuming full static information on trains and loads. We then consider the more realistic case with incomplete future information. A dynamic load assignment model with a rolling horizon scheme is developed for continuous terminal operations; the model balances the advantage from optimizing multiple trains together against the risk of making suboptimal decisions due to incomplete future information.

This research is particularly timely in light of recent increases in fuel prices, their impact on industry operating costs, and the desire to conserve energy and reduce greenhouse gas emissions. Class I railroads spent more than $6.2 billion on fuel in 2005, making fuel cost their second largest operating expense (AAR 2006), and fuel costs continue to increase. North American railroad fuel cost doubled from 2002 to 2005, and since 1999 it is up by nearly a factor of three (BNSF 2004a). Therefore, our research addresses an important economic and environmental topic in rail transport and also makes a methodological contribution by introducing rolling horizon operations for IM loading efficiency to the literature.

2. Methodology

2.1. Loading Assignment at IM Terminals

The rail IM business in North American is quite different from the general freight business, and intermediate stops are no longer the norm for most IM trains, the subject of our research. Railroads try to avoid intermediate switches and stops because the IM business is highly time sensitive. For example, approximately 80% of the IM trains on the BNSF Transcon route (Chicago–LA) have no intermediate operations; most of the other 20% have no more than two intermediate stops, and these are generally close to the final destination, so there is little container shifting occurring en route (Avriel et al. 1998; Giemsch and Jellinghaus 2004; Utterback 2006). Therefore, the initial loading pattern for most trains will be the principal factor affecting their aerodynamic performance for all or most of the trip.

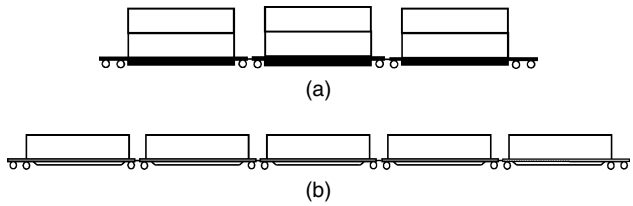

Figure 1 (a) A Three-Unit Well Car with Six Slots and (b) A Five-Unit Spine Car with Five Slots

Note. From Lai, Barkan, and Önal (2008).

At IM terminals, containers, and trailers of a variety of lengths are assigned to available well, spine, or flat cars by terminal managers (BNSF 2004b; UP 2005). IM loads, i.e., trailers or containers, vary in length from 20 to 57 feet. There is considerable variety in the design and capacity of IM railcars; they have different numbers of units and slots and thus loading capabilities. An IM railcar can have one or more units permanently attached to one another (via articulation or drawbar). A unit is a frame supported by at least two trucks, providing support for one or more platforms (a.k.a. slots). For example, Figure 1(a) shows an articulated three-unit well car, and Figure 1(b) is a five-unit spine car. A platform (or slot) is a specific container/trailer loading location. As a result, each well car unit has two slots because of its accommodation of two containers, one stacked on the other (a.k.a. "double stacking"), and each spine car unit has one slot (Figure 1).

There are also a number of loading rules developed for safety purposes and various feasible and infeasible combinations of IM load and car configurations. Because IM cars in a train are not generally switched in and out at terminals (i.e., cars will not be uncoupled from one train and coupled to another), managers primarily control the assignment of loads but not the configuration of the equipment (i.e., railcars) in a train. Consequently, we treat the train configuration as given.

Aerodynamic drag is a major component of train resistance, particularly at high speeds (Hay 1982; AREMA 2001; Lai and Barkan 2005). In the 1980s, the Association of American Railroads (AAR) sponsored research on wind tunnel testing of rail equipment, including large-scale IM car models (Gielow and Furlong 1988). The results were used to develop the Aerodynamic Subroutine of the Train Energy Model (TEM) (Drish 1992). These experiments showed that

Table 1 Adjusted Factor for Each Gap in the Train

k	Drag area (ft^2)	Adjusted factor
1 (locomotive)	31.618	1.5449
2	28.801	1.4073
3	26.700	1.3046
4	25.133	1.2280
5	23.963	1.1709
6	23.091	1.1283
7	22.440	1.0964
8	21.954	1.0727
9	21.591	1.0550
10	21.320	1.0418
100	20.466	1.0000

Note. From Lai, Barkan, and Önal (2008).

gap length between IM loads and position-in-train were the two important factors affecting train aerodynamics (Engdahl 1987). Larger gaps result in a higher aerodynamic coefficient and greater resistance, and the front of the train experiences the greatest aerodynamic resistance due to headwind impact. Therefore, to incorporate both important aerodynamic factors, the model chooses to minimize the summation of total adjusted gap lengths (i.e., gap lengths multiplied by adjusted factors). The adjusted factors (accounting for the position-in-train effect) are computed by dividing the drag area (representing the aerodynamic resistance in ft^2) of a given unit by the drag area of the 100th unit; the result is listed in Table 1.

2.2. Static Aerodynamic Efficiency Model

Placing loads with shorter gaps in the frontal position generates less aerodynamic resistance; therefore, the objective function of the aerodynamic model is to minimize the total adjusted gap length of all trains considered in the decision horizon. The following notation is used in the algebraic model: i is an index referring to the type and size of the load (namely, 40′ container, 48′ trailer, 53′ trailer, etc.); C_L is the subset of i for container loads; and T_L is the subset of i for trailer loads. We group loads of the same type together with an index, j ($j = 1, 2, 3, \ldots, J_i$); J_i is the number of loads of a specific type and size i ($i =$ C40, T48, T53, etc.). For instance, $J_{T48} = 10$ means that there are ten 48′ trailers in the storage area. Let t denote the index for outgoing trains ($t = 1, 2, \ldots, T$). The symbol k defines the position of each unit in the train ($k = 1, 2, 3, \ldots, N$), where $k = 1$ corresponds to the first IM unit of the train. The slot position in each unit is denoted by p, where $p = 1$ represents the upper (top) platform in a well car unit or the single platform in a spine car or flat car unit and $p = 2$ represents the lower (bottom) platform in a well-car unit (Figure 2). The following symbols represent the parameters used in the model: A_k is the adjusted factor of the kth gap shown in Table 1, where $A_1 > A_2 > \cdots > A_N$; U_{tk} is the length of the kth unit of train t; δ_{tk} indicates the type of the kth unit in train t, where $\delta_{tk} = 1$ when the unit is a well car unit, and $\delta_{tk} = 0$ otherwise; L_i is the length of the ith type load; Q_{tkp} is the length limit of position p in the kth unit of train t; w_{ij} is the weight of the jth load of type i; W_{tk} is the weight limit of the kth unit of train t; and R_{itpk} is a four dimensional matrix for loading capabilities of each slot, where $R_{itpk} = 1$ if the ith type of load can be assigned to position p in unit k of train t, or it equals 0 otherwise. Finally, Φ represents an arbitrarily large number introduced for modeling purposes as explained in the model description below.

Two sets of binary decision variables are included in the model. The first variable is denoted by y_{ijtpk} such that

$$y_{ijtpk} = \begin{cases} 1, & \text{if } j\text{th load of type } i \text{ is assigned to} \\ & \text{position } p \text{ in the } k\text{th unit of train } t, \\ 0, & \text{otherwise.} \end{cases}$$

The second binary variable, denoted by x_{tk}, determines whether the top slot in a well unit can be used; namely,

$$x_{tk} = \begin{cases} 1, & \text{if the top slot of the } k\text{th unit in train } t \\ & \text{can be used,} \\ 0, & \text{otherwise.} \end{cases}.$$

According to the loading rules, the top slot can be used when the bottom slot is filled by containers whose total length is at least 40′ (AAR 2004).

The loading problem is formulated as a mixed integer program (MIP) that minimizes fuel consumption (i.e., the total adjusted gap length) of T

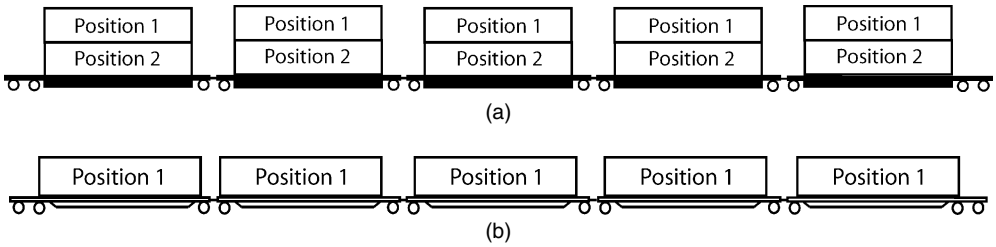

Figure 2 The Available Slots in (a) Five-Unit Well Car and (b) Five-Unit Spine Car
Note. From Lai, Barkan, and Önal (2008).

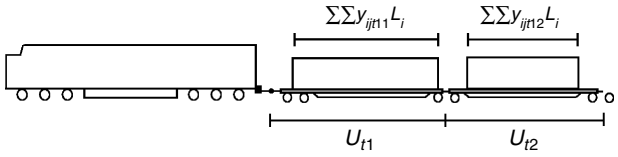

Figure 3 Locomotive and First Two IM Units in a Train

Note. From Lai, Barkan, and Önal (2008).

outgoing trains. For train t, the objective function (total adjusted gap length) is

$$z_t = 0.5 \times \left\{ A_1 \left(U_{t1} - \sum_i \sum_j y_{ijt11} L_i \right) + \sum_{k=1}^{N-1} A_{k+1} \left[\left(U_{tk} - \sum_i \sum_j y_{ijt1k} L_i \right) + \left(U_{tk+1} - \sum_i \sum_j y_{ijt1k+1} L_i \right) \right] \right\}. \quad (1)$$

This objective function is composed of two parts. The first part, representing the gap length between the locomotive and the first load (Figure 3), is the difference between the length of the first unit (U_{t1}) and the length of the load in position 1 of the first unit ($\sum\sum y_{ijt11}L_i$), which is then divided by two. Multiplying the gap length by the adjusted factor A_1 results in the first adjusted gap length. Each of the subsequent gaps is half of the difference in length between the current unit and the load $(U_{tk} - \sum\sum y_{ijt1k}L_i)/2$ plus half of the length difference between the next unit and the load $(U_{tk+1} - \sum\sum y_{ijt1k+1}L_i)/2$, multiplied by the appropriate adjusted factor, A_k. Thus the second part of the objective function computes the sum of the subsequent adjusted gap lengths. Note that we only take into account the loads in position 1 of all units in the train. This is reasonable because they are the only loads in spine or flat cars; for well cars, the upper level gaps have a more significant aerodynamic effect than the lower level gaps (Furlong 1988; Storm 2005; Airflow Science 2006). A schematic representation is given in Figure 3.

The complete mathematical program for all T trains is as follows:

$$\min \sum_{t=1}^{T} z_t \quad (2)$$

$$\text{subject to: } \sum_t \sum_p \sum_k y_{ijtpk} R_{itpk} \le 1 \quad \forall i, j \quad (3)$$

$$y_{ijtpk} \le R_{itpk} \quad \forall i, j, t, p, k \quad (4)$$

$$40 - \sum_{i \in C_L} \sum_j y_{ijt2k} L_i \le \Phi(1 - x_{tk}) \quad \forall t, k \text{ (such that } \delta_{tk} = 1) \quad (5)$$

$$\sum_{i \in C_L} \sum_j y_{ijt1k} \le x_{tk} \quad \forall t, k \text{ (such that } \delta_{tk} = 1) \quad (6)$$

$$\sum_{i \in T_L} \sum_j y_{ijt2k} \le 2 \times \left(1 - \sum_{i \in C_L} \sum_j y_{ijt1k} \right) \quad \forall t, k \text{ (such that } \delta_{tk} = 1) \quad (7)$$

$$\sum_i \sum_j \sum_p y_{ijtpk} w_{ij} \le W_{tk} \quad \forall t, k \quad (8)$$

$$\sum_i \sum_j y_{ijtpk} L_i \le Q_{tkp} \quad \forall t, k, p \quad (9)$$

$$y_{ijtpk}, x_k = 0, 1. \quad (10)$$

Minimizing total adjusted gap length creates the most efficient train configuration, but the loading assignment must conform to the loading capability of each unit as well as length and weight constraints. Constraints (3) and (4) ensure that each load can be assigned to no more than one slot and must obey the loading assignment rules (R_{itpk}). Constraints (5) and (6) together state that if the bottom slot (position 2) in a well car unit ($\delta_{tk} = 1$) is not filled with containers greater than 40′ (in which case Equation (5) requires that x_{tk} =0), then no load can be assigned to the top slot (position 1) for the same unit; i.e., $\sum\sum y_{ijt1k} = 0$ and therefore $y_{ijt1k} = 0$ for all i, j. Note that constraint (6) allows a bottom load without a top load ($y_{ijt1k} = 0$). Constraint (7) ensures that containers cannot stack on top of trailers in the well car units; the parameter 2 is used for the possible scenario of two trailers in one well car unit. Constraint (8) is the weight limit that is imposed for each car unit in order to reflect its total carrying capacity (W_{tk}). Constraint (9) is the length limit imposed for each slot to guarantee that the total length of loads in a given slot does not exceed the length of that slot (Q_{tkp}). Note that the trivial solution, namely, $y_{ijt1k} = 0$ and $x_{tk} = 0$, satisfies all the constraints of the model. However, this would result in the largest total adjusted gap because all gaps would be at their maximum value. This case is ruled out because of the minimization of the total gap length. Thus the model prefers not to leave a load behind if a suitable slot is available.

The above optimization formulation (2)–(10) reminds us of certain network flow problems (e.g., assignment problem) that can be solved efficiently. However, the existence of certain constraints (e.g., the weight constraints) makes the problem NP-hard. Optionally we can develop relaxation or decomposition based heuristics for our model, but earlier research (Lai, Barkan, and Önal 2008) has shown that existing commercial MIP solvers (i.e., CPLEX) can solve this problem within reasonable time (this is also found to be true in our numerical experiments). Therefore, in this study we choose to use CPLEX to solve the problem instances.

The current terminal operational practice and a previous paper by Lai, Barkan, and Önal (2008) consider

loading plans for the current outgoing train only. This scenario is a special case of the general model developed here in which $T = 1$. The model thus optimizes the aerodynamic efficiency of one outgoing train for a given set of loads. However, some degree of advance information about outgoing trains and loads is often available (Anderson 2006). This provides an opportunity to achieve even more aerodynamically efficient loading patterns by optimizing more trains and loads together.

2.3. Dynamic Aerodynamic Efficiency Model

Obviously, optimizing multiple trains simultaneously will lead to more efficient loading plans if complete information on all trains and loads is available at the time of optimization (i.e., static current information). In practice, however, information about some loads may not be immediately available (i.e., future information). Under some circumstances, including the loading pattern of a later train in the optimization will reduce the efficiency of the immediate outgoing train. For example, suppose the two trains compete for the same "suitable" load, and the later train gets this load in the optimization (with the objective of minimizing the total adjusted gaps in both trains). It is possible, however, that after the dispatch of the immediate train, another suitable load with the same characteristics becomes available. As a result, the earlier optimal solution (before knowing the future load information) turns out to be suboptimal (overall). Therefore, uncertainty about future loads introduces some degree of risk in optimizing multiple trains; i.e., the overall optimum for multiple trains will not be achieved. In a dynamic setting, there is a trade-off between the benefit of optimizing multiple trains simultaneously and the risk of making wrong decisions for the uncertain future.

To address this trade-off, we propose a dynamic loading approach with rolling horizons, where loading decisions with "smoothed" objectives are updated over time as new information becomes available. Carrying out this approach poses three questions: (1) when to optimize loading patterns for one (or more) train; (2) how many trains to optimize each time and how to optimize them; and (3) how many trains to load after each optimization.

The first and third questions are relatively simple to answer. In principle, it is always better to postpone an optimization decision to the last moment possible (before loading a departing train) because it maximizes the available information, thereby reducing uncertainty. Therefore, to the extent practicable, train loading should be delayed until just before its departure. For the same reason, it is always better to load only the next outgoing train based on the optimal loading pattern even though multiple trains may be optimized together. Hence, we should always load the minimum number of trains, assuming that each optimization process can be conducted efficiently to update the optimal loading patterns in time. The only remaining question is how many trains should be optimized each time and how to optimize them.

We further propose an exponential smoothing approach under the rolling horizon framework, in which future trains are considered simultaneously with the current train. Before loading the tth train, suppose we have known information on $n_k(t)$ unassigned loads, and these loads can fill a maximum number of $T(t)+1$ sequential trains (i.e., trains $t, t+1, \ldots, t+T(t)$), where $T(t)+1$ is the number of future trains considered in an assignment. We define the time horizon to be from the departure time of train t to that of train $t + T(t)$. The loading decision for train t will be directly relevant to the trains departing in this horizon. Meanwhile, these trains are also directly influenced by the future loads incoming within this horizon; assume there are $n_u(t)$ such future loads. We optimize the following weighted average of objective functions:

$$\begin{aligned} \min \quad & \sum_{s=t}^{t+T(t)} \alpha_{t,s} z_s \\ \text{s.t.} \quad & (3)\text{–}(10). \end{aligned} \tag{11}$$

In (11), parameter $\alpha_{t,s}$ is a nonnegative weight assigned to a future train s, for $t \le s \le t + T(t)$. The vector of weights, $\tilde{\alpha}(t) := (\alpha_{t,t}, \alpha_{t,t+1}, \ldots, \alpha_{t,t+T(t)})$, specifies how future trains are included in the loading decision. For example, $\tilde{\alpha}(t) = (1, 0, 0, \ldots, 0)$ corresponds to the trivial case where we optimize and load the departing train t only, whereas $\tilde{\alpha}(t) = (1, 1, 0, \ldots, 0)$ corresponds to optimizing two trains t, $t+1$ together and loading train t only. Ideally, we want to define $\tilde{\alpha}(t)$ in such a way that the objective in (11) is a weighted average of short-term (currently departing train) and long-term (future trains) objectives. To achieve this, we propose to use exponentially decreasing weights:

$$\tilde{\alpha}(t) = (1, \alpha_t, \alpha_t^2, \ldots, \alpha_t^{T(t)}), \tag{12}$$

where α_t is a scalar such that $0 \le \alpha_t \le 1$.

Then (11) becomes

$$\begin{aligned} &\min \sum_{s=t}^{t+T(t)} (\alpha_t)^{s-t} z_s \\ &= \min \Bigg[(\alpha_t)^{T(t)} (z_t + \cdots + z_{t+T(t)}) \\ &\qquad + \sum_{r=0}^{T(t)-1} (1-\alpha_t)(\alpha_t)^r (z_t + \cdots + z_{t+r}) \Bigg], \end{aligned} \tag{13}$$

which is a weighted average of z_t, $(z_t + z_{t+1}), \ldots$, and $\sum_{s=t}^{t+T(t)} z_s$. If most load information is already known and there are few unknown loads on the horizon (i.e., $n_u(t) \ll n_k(t)$), we should choose $\alpha_t \approx 1$, such that $\tilde{\alpha}(t) \approx (1, 1, \ldots, 1)$ and $\sum_{s=t}^{t+T(t)} \alpha_{t,s} z_s \approx \sum_{s=t}^{t+T(t)} z_s$, to exploit the efficiency from optimizing multiple trains together. In the limit, this scenario converges to the static optimization case where full future information is available. On the other hand, if we expect a large number of unknown loads on the horizon (i.e., $n_u(t) \gg n_k(t)$), we should choose $\alpha_t \approx 0$, such that $\tilde{\alpha}(t) \approx (1, 0, 0, \ldots, 0)$ and $\sum_{s=t}^{t+\tau} \alpha_{t,s} z_s \approx z_t$, to avoid the penalty due to future uncertainty.

The weight scalar α_t can vary over time and across the train index t. Its appropriate value can be calibrated from historical data over repeated experiments or simulations for any existing IM facility. When empirical data are not available, a reasonable value should be estimated. Note from (11) and (13) that the value of α_t controls the balancing between the short-term objective (regarding immediate train departure) and long-term importance (future trains to be loaded). It reflects the relative significance of static information versus dynamic information, which is closely related to the concept of "degree of dynamism" (DOD) introduced in Lund, Madsen, and Rygaard (1998) and Larsen (2001)—the proportion of dynamic information at the time of decision. We propose an "adjusted DOD" defined as follows:

$$\text{DOD} = n_u(t) \big/ (n_k(t) + n_u(t)) \quad \forall t, \tag{14}$$

and we propose using α_t as the following:

$$\alpha_t = 1 - \text{DOD} = n_k(t)/(n_k(t) + n_u(t)) \quad \forall t. \tag{15}$$

For every optimization, the value of DOD is determined based on not only the numbers of known loads, n_k, but also the estimated number of unknown loads, n_u. For example, if n_k is large enough for three consecutive outgoing trains, then $n_u$will be the number of estimated future incoming loads from now until the decision time for the third outgoing train. Therefore, a uniform time for the arrival of loads and departure of trains would result in a constant DOD, whereas a nonuniform arrival and departure time would lead to adaptive DODs.

2.4. Model Extensions and Operating Rules

2.4.1. Level of Service. The model thus far treats all loads as equally important by assuming each load can be placed on any of the trains. This assumption is reasonable given the frequent service on many IM routes; however, sometimes there may be certain loads with a higher priority than others. Railroads may promise their customers that loads making the cutoff time will be loaded onto one of the next several trains (cutoff is the time an IM load physically has to be in the terminal ramp for it to catch the outbound train going to its destination). These specific operational practices can be accommodated by adding level of service (LOS) constraints during data preprocessing. For example, if we know there are 30 United Parcel Service (UPS) trailers that must make it onto the first outgoing train, we add constraint $\sum_{i \in \text{UPS}} \sum_j \sum_p \sum_k y_{ij1pk} \geq 30$ to the original model.

More generally, the LOS constraints can be enforced as follows:

$$\sum_t \sum_p \sum_k d_t y_{ijtpk} \leq S \quad \forall i, j, \tag{16}$$

where d_t is the departure time of train t, and S is the service level in hours. This constraint ensures that every load will be assigned within the service level. The constraints can also be modified to impose this rule on specific loads by changing i to "$i \in$ *specific loads*." Similarly, if S is defined in terms of the number of train departures (i.e., the load can be delayed by at most S train departures) and load ij makes to the cutoff time of the train, which is denoted by G_{ij}, the following constraint ensures load ij being assigned in the next S trains.

$$\sum_t \sum_p \sum_k y_{ijtpk} t \leq G_{ij} + S \quad \forall i, j. \tag{17}$$

2.4.2. Blocking and Loading Before Cutoff Time. As mentioned in §2.1, the rail IM business in North America is different from the general freight business in the sense that railroads try to avoid intermediate switches and stops because this business is highly time sensitive. Therefore, the initial loading pattern for most trains will be the principal factor affecting their aerodynamic performance for all or most of the trip.

Blocking is usually defined on a car basis prior to loading assignment. Therefore, for the small number of cases in which the IM train does make intermediate stops, we can set y_{ijtpk} to zero if load ij is not supposed to be on the kth unit of train t prior to the optimization. This would possibly expedite the solution process by reducing the size of the decision space.

Occasionally, the actual time required to load the whole train is longer than the time period from cutoff to departure; therefore, terminal managers have to start loading trains before the cutoff time. Some site-specific loading rules can be developed to ensure aerodynamic efficiency under this circumstance. For yards handling only containers or mixed IM loads, managers should try to load the shorter containers ($\geq 40'$) in the lower positions of well cars. This is because the aerodynamics of longer containers

atop shorter containers are better than the opposite (shorter atop longer). Thus, holding longer containers for upper level slots will generally be a good option. For yards handling only trailers, managers should assign loads that best match the available slots, beginning at the front of the train, to guarantee an aerodynamic loading pattern. With above rules, managers can still apply the proposed load-assignment models with available loads and slots at cutoff time and can approach the system optimum.

3. Model Implementation and Case Study

In the following sections, we first apply the static model to evaluate aerodynamic efficiency from optimal loading at the system level, assuming static information of trains and loads. Then we use the dynamic model to analyze continuous terminal operations when information is dynamic. The dynamic model is implemented for two different cases: (1) a terminal with uniform arrival rate of incoming loads and (2) a terminal with nonuniform arrival rate of incoming loads.

3.1. Static Case with Perfect Information

In the static case, we conducted an analysis of 16 trains with 90-minute departure intervals ranging from 84 to 122 units (mean = 104) and 4,224 loads (both domestic and international IM loads) in a 24-hour window. Trains originated from a major IM terminal in one 24-hour period. The numbers and types of available loads for each train were obtained from data provided by the railroad. Five possible scenarios were conducted to evaluate the benefit of optimizing more trains together. They are to optimize 1, 2, 4, 8, and 16 trains at a time, assuming perfect information for all trains and loads. CPLEX 10.0 incorporated with GAMS (Brooke et al. 1998) was used to solve the model in reasonable time. For example, the 16-train scenario (with 45,477 variables and 8,605 effective constraints after data preprocessing) was solved to optimality within 0.922 seconds by a 2.26 GHz CPU with 1.5 GB RAM.

Considering only one train at a time is consistent with current terminal practice. However, with perfect information, the more trains that are optimized at a time, the better the aerodynamic efficiency (Figure 4), although the marginal benefit declines considerably beyond four trains. Terminal managers' goal in load assignment is to maximize slot utilization; therefore, they are largely indifferent to alternative loading patterns as long as they achieve 100% slot utilization and comply with applicable loading rules. Consequently, with 100% slot utilization, average terminal operating practice will be some intermediate value between the minimal and maximal total adjusted gap length. We assumed that the mean of the total adjusted gap length for the minimum and maximum cases represents average terminal loading performance in scenarios with 100% slot utilization (Lai, Barkan, and Önal 2008) and is therefore equal to 31, 822′ (Figure 4).

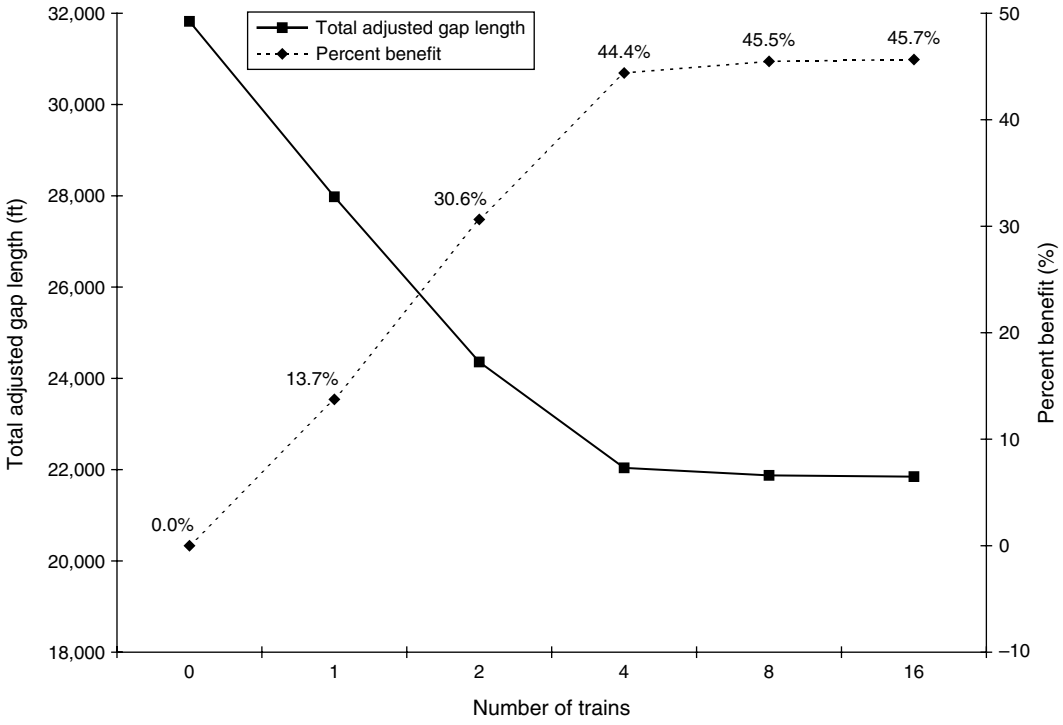

Figure 4 Effect of the Number of Trains Optimized Simultaneously on the Adjusted Gap Length

Compared with this baseline performance, in which load assignment is based on maximizing slot utilization only, the benefit of optimizing aerodynamic efficiency of IM trains ranges from 14% to 46%. Because scenarios 2–5 (optimizing multiple trains) are more beneficial than scenario 1, the fuel savings are also more significant.

Placing loads on different trains is generally feasible due to the frequent service of IM operations. Lai, Barkan, and Önal (2008) estimated that optimizing one train at a time can save 15 million gallons of fuel per year compared to current operations; optimizing multiple trains together shows the potential to further improve savings by an additional 30% (Figure 4). In other words, if there is no flexibility in placing loads on different trains, the potential 30% savings is lost.

In practice, loads often arrive at and trains depart from terminals quickly, with little lead time. Therefore, it is rarely the case that reliable load information will be available for more than three trains at any time. Consequently, in §§3.2 and 3.3 we implement the rolling horizon framework developed in this study for continuous terminal operations.

3.2. Rolling Horizon Operations with Uniform Arrival Rate

To implement the rolling horizon scheme, we need to know the number of loads initially available and the arrival pattern of additional loads between consecutive train departures. According to the cutoff time and load information, we assume that 690 loads (for three trains) are known at the beginning of the 24-hour period, with a constant rate of incoming loads. The constant rate is computed as the sum of all the loads for the 16 trains divided by 16. The resultant rate is approximately 230 loads per 90 minutes. In other words, because the train departure interval is also 90 minutes there will be 230 new loads incoming before the next optimization and assignment. The detailed distribution of load numbers (by type) is shown in Table 2.

In this case, $T(t) = 2$ is true for all trains except the final two. The adjusted DOD within the time window of this system is approximately $230 \times 2/(690 + 230 \times 2) = 0.4$ most of the time, and we simply recommend using $\alpha_t = 1 - 0.4 = 0.6$ for all t. As proposed, we consider up to three trains in each optimization and load only the current outgoing train. The unassigned loads and future trains will be incorporated into the next optimization along with any new loads. Figure 5 shows the result of implementing the rolling horizon scheme for this empirical example. The dotted line is a benchmark representing the objective value for a static assignment case where one train is optimized at a time with 230 available loads (as described in Lai, Barkan, and Önal 2008). The dashed line shows the idealized scenario with full information, where 16 trains are optimized together. This objective value is the best performance possible and serves as a reference point to evaluate the performance of the proposed rolling horizon framework. The solid line in Figure 5 shows the experiments with the rolling horizon scheme where we vary α_t over 0 to 1.

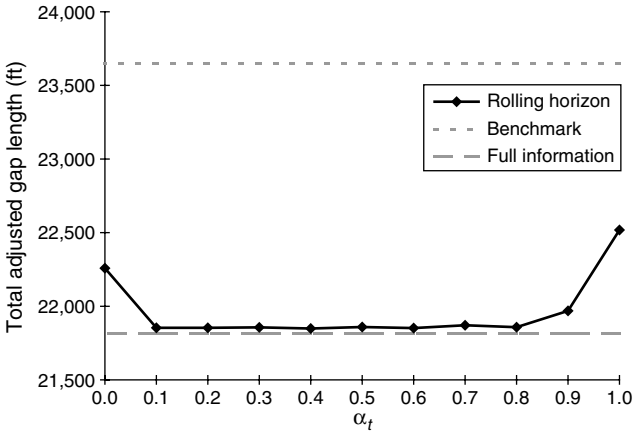

Figure 5 **Rolling Horizon Scheme on Uniform Operation with Different α_t Values**

These numerical results verify our qualitative arguments. When $\alpha_t = 1$, all trains are treated equally in each optimization, and the final objective value is actually about 3% higher than the best possible. This shows that the optimality of the current outgoing train is unnecessarily over-compromised by putting too much emphasis on future trains. On the other hand, when $\alpha_t = 0$, the objective value is also about 2% higher than the best possible, confirming our argument that there are benefits from considering multiple trains together. The exponential smoothing scheme, however, successfully reduces the objective value to within 0.2% of the best possible for any α_t between 0.1 and 0.6. These values are about 7.5% smaller than the benchmark value for the

Table 2 **The Distribution of Number of Loads for All Types of IM Loads**

	Containers					Trailers						
Load type	C20	C40	C45	C48	C53	T20	T28	T40	T45	T48	T53	Total
Initial number of loads	81	225	18	60	255	6	9	6	6	9	15	690
Incremental loads/90 min	27	75	6	20	85	2	3	2	2	3	5	230

one-train-a-time strategy, thus demonstrating that the proposed rolling horizon scheme with exponentially decreasing weights is beneficial.

The numerical experiments also reveal interesting insights into the choice of the weight parameter. Because DOD ≈ 0.4 at all times, we propose that a possible α_t value be $1 - \text{DOD} = 0.6$ and that $\tilde{\alpha}(t) = (1, 0.6, 0.36, \ldots)$, $\forall t$. Figure 5 shows that the optimal objective value is actually insensitive to α_t for a wide range, $0.1 \le \alpha_t \le 0.8$. This finding indicates that an appropriate value of α_t can be calibrated from historical data; if historical data are not available, $1 - \text{DOD}$ would be a good choice. Compared to optimizing one train at a time ($\alpha_t = 0$), rolling horizon operations yield a 2.2% fuel savings in this example, or approximately 160,000 gallons of fuel per year for trains on the single route considered in this analysis.

The computation time and optimality gap by CPLEX 10.0 does vary across optimization instances. For example, when $\alpha_t = 0.6$, 12 out of the 16 instances are solved to within 0.1% relative optimality gap in less than one CPU second. Note that only part of the solution for each instance (regarding the current outgoing train) is finally implemented into the overall solution throughout the horizon. Although the other four instances have a relative optimality gap between 1%–3% after 600 CPU seconds (predetermined upper limit), the overall quality of the rolling horizon solution is close to the known optimum (with full information and zero optimality gap) (Figure 5). This computational performance is also found to be true for the example in the next section.

3.3. Rolling Horizon Operations with Nonuniform Arrival Rate

In §3.2 we implemented the rolling horizon scheme for a terminal with a hypothetical uniform load arrival and train schedule. In this section we consider a class 1 railroad terminal with nonuniform conditions according to a three-month sample of real terminal data. IM loads typically do not arrive at terminals uniformly throughout the 24-hour day cycle (Figure 6). Instead, more loads arrive by day than at night, with the peak at noon at the terminal we studied. A detailed breakdown of the incoming loads by type and intended departure time is presented in Table 3. The number of units in the ten trains ranges from 93 to 122 units (mean = 110). The cutoff time is assumed to be two hours before departure, so the 11 A.M. train can draw from the pool of loads not assigned after the 5 A.M. train, plus those newly arrived from 3 A.M. to 9 A.M.

In this experiment, we examine the effect of varying (1) the constant factor ($0 \le \alpha_t \le 1$) and (2) adaptive α_t. The inventory level is always enough for two trains, i.e., $T(t) = 1$ throughout the decision horizon;

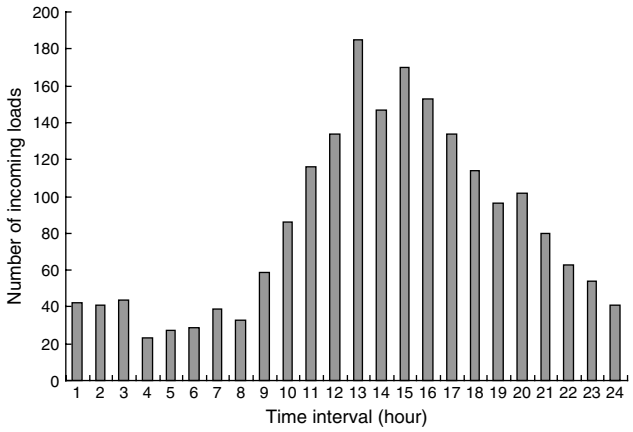

Figure 6 Distribution of Incoming Loads by Time of Day

hence, we consider two future trains in each optimization and load only the current outgoing train. The adjusted DOD for each of the 10 assignments depends on the current inventory level and the next outgoing train's departure time. For instance, the DOD of the 5 A.M. train is $201/(201 + 324) = 0.617$ because it has 324 loads available in the pool and there will be 201 new loads incoming to the terminal before the cutoff time of the next outgoing train (11 A.M. train).

Figure 7 shows the result of implementing the rolling horizon scheme to this empirical example for a range of static α_t. The results of applying the rolling horizon scheme to nonuniform operations are largely similar to uniform operations. Either placing too much emphasis on future trains (with $\alpha_t = 1$) or myopically ignoring future trains (with $\alpha_t = 0$) compromises the optimality of the solution. Again, the optimal objective value is relatively insensitive to the value of α_t for a wide range, $0.1 \le \alpha_t \le 0.9$. The static case with applying adaptive DOD (changing α_t over time) yields results very close to $\alpha_t = 0.6$. This again demonstrates that if historical data are not available, $\alpha_t = 1 - \text{DOD}$ would be a good choice. Compared with optimizing one train at a time ($\alpha_t = 0$), using rolling horizon operations yields an 8.6% benefit, or approximately 700,000 gallons of fuel savings per year for the IM trains at this route.

4. Discussion

Optimizing multiple trains together has the potential to further improve the fuel savings compared to optimizing one train at a time. The greater the flexibility in placing loads on different trains, the better the aerodynamic efficiency. However, the marginal benefit drops considerably beyond optimizing four trains at a time (Figure 4). As expected, the marginal benefit is greatest when comparing "no flexibility" (i.e., optimizing one train at a time) to "flexibility in several trains." Beyond that, there are diminishing

Table 3 (a) Number of Incoming Loads by Time and Type; (b) Departure Time of Outgoing Trains

(a)

	Load type					
Time	C40	C45	C48	C53	T40	Total
Initial	144	4	12	168	72	400
1:00	15	0	1	18	8	42
2:00	15	0	1	17	7	40
3:00	16	0	2	18	8	44
4:00	8	0	1	10	4	23
5:00	10	0	1	11	5	27
6:00	10	0	1	12	5	28
7:00	14	0	1	16	7	38
8:00	12	0	1	14	6	33
9:00	21	1	2	17	11	52
10:00	31	1	3	36	15	86
11:00	42	1	3	49	21	116
12:00	48	1	4	56	24	133
13:00	67	2	6	78	33	186
14:00	53	1	4	62	26	146
15:00	61	2	5	71	31	170
16:00	55	2	5	64	28	154
17:00	48	1	4	56	24	133
18:00	41	1	3	48	21	114
19:00	35	1	3	40	17	96
20:00	37	1	3	43	18	102
21:00	29	1	2	34	14	80
22:00	23	1	2	26	11	63
23:00	19	1	2	23	10	55
0:00	15	0	1	17	7	40

(b)

	Depature time
Train 1	1:00
Train 2	5:00
Train 3	11:00
Train 4	13:00
Train 5	15:00
Train 6	16:00
Train 7	18:00
Train 8	20:00
Train 9	21:00
Train 10	0:00

returns because there are not enough new types of load choices available to yield additional benefit in train aerodynamics.

Railroad IM terminals usually use computer software (e.g., OASIS) to assist the loading assignment process; therefore, integration of our proposed model into the software currently being used would not require significant institutional or process change. It also should have little if any impact on operating cost because the general process remains the same; the only difference would be terminal managers' decisions about which load should be assigned to which slot to maximize fuel efficiency. We believe implementation of the proposed rolling horizon optimization scheme can automate the terminal managers' tasks

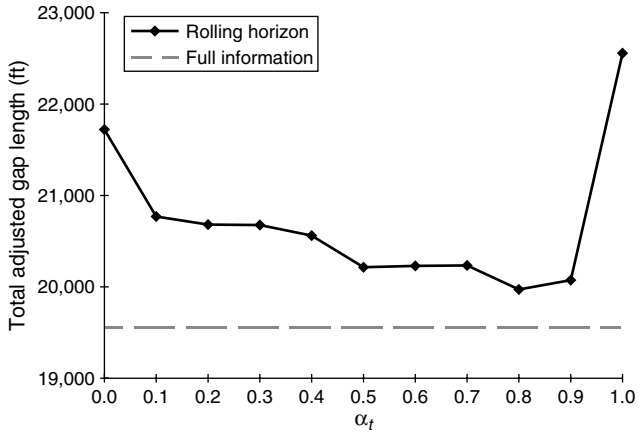

Figure 7 Rolling Horizon Scheme Under Nonuniform Conditions for Different α_t Values

regarding the large variety of loads and railcar types, thereby enabling them to load trains in a more aerodynamically efficient manner.

The experiments described above have thus far treated loads with equal importance, ignoring potential constraints on loads' time priorities. It is worthwhile to examine the effect of LOS constraints on the outcome as described in §2.4.1. Using the same data from §3.3, we now assume that the railroad promises to its customers that all loads arriving before the cutoff time will be loaded onto one of the next three departing trains (i.e., $S = 3$). We compared the results in aerodynamic efficiency (total adjusted gap length) and delay (in units of departed trains and hours) of imposing or not imposing LOS constraints.

The optimization results show that imposing LOS constraints reduces the aerodynamic efficiency of IM trains by 36%. This is expected because LOS constraints reduce the flexibility of assigning loads to different trains. In Figure 8, the vertical axis represents the cumulative number of loads, and the horizontal axis is the delay (in units of departed trains and hours); it demonstrates the cumulative delay of loads (in units of departed trains and hours) with or without the LOS constraints. With the LOS constraints, the delays are more uniform across loads, but the total delay of all loads is actually slightly longer. For example, forcing one trailer to be placed in an IM well car due to the LOS constraints would probably cause two containers to be left behind. The impact on delay would decrease if there were enough equipment dedicated to trailers (e.g., spine or flat cars); however, this will often not be the case in practice. Thus, LOS constraints will generally have a significant impact on both energy efficiency and operational cost.

5. Conclusion

This paper presents static and dynamic aerodynamic efficiency models for the loading of multiple IM

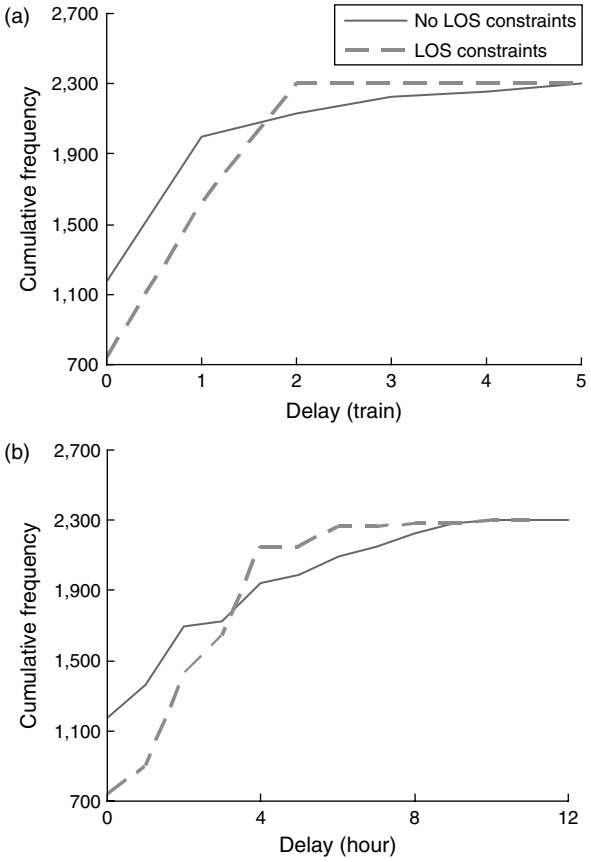

Figure 8 Delay With or Without LOS Constraints in Terms of (a) Trains and (b) Hours

trains. It also develops a rolling horizon scheme for continuous train terminal operations. For the static case, when full information is available, the system optimum can be reached by optimizing as many trains as possible. In practice, however, terminals operate in a dynamic environment where not all information on incoming loads and trains is available. Attempting to optimize the loading of too many trains in this environment will reduce the ability to achieve the most efficient loading configuration. Therefore, a rolling horizon scheme with decreasing weight assigned to each train is proposed to counterbalance the effect of uncertainty. Numerical results show that the realistic rolling horizon scheme significantly reduces the adjusted gap length as compared to the current practice, leading to a substantial benefit from the aerodynamic efficiency of IM trains. Correspondingly larger savings in fuel, emissions, and expense are possible if the methodology described in this paper could be applied to all North American IM trains.

Acknowledgments

The authors are grateful to Mark Stehly, Larry Milhon, and Paul Gabler of the BNSF Railway for their support and help on this project. The first author was supported by a research grant from the BNSF Railway and a CN Research Fellowship in Railroad Engineering at the University of Illinois.

References

Airflow Science. 2006. *Aerodynamic Subroutine Version 4.0*. Association of American Railroads, Washington, D.C.

American Railway Engineering and Maintenance-of-Way Association (AREMA). 2001. *Manual for Railway Engineering, Chapter* 16, *Part* 2, *Train Performance*. American Railway Engineering and Maintenance-of-Way Association, Landover, MD.

Anderson, F. 2006. Personal Communication. BNSF Railway Company, Fort Worth, TX.

Association of American Railroads (AAR). 2004. *Loading Capabilities Guide*. Association of American Railroads, Washington, D.C.

Association of American Railroads (AAR). 2004–2006. *Railroad Facts*. Association of American Railroads, Washington, D.C.

Avriel, M., M. Penn, N. Shpirer, S. Witteboon. 1998. Stowage planning for container ships to reduce the number of shifts. *Ann. Oper. Res.* **76** 55–71.

BNSF Railway Company. 2004a. Every drop counts in goal to improve fuel efficiency. *BNSF News*. BNSF Railway Company, Fort Worth, TX.

BNSF Railway Company. 2004b. *Intermodal Loading Guide*. BNSF Railway Company, Fort Worth, TX.

Brooke, A., D. Kendrick, A. Meeraus, R. Raman. 1998. *GAMS—A User's Guide*. GAMS Development Corporation, Washington, D.C.

Corry, P. G., E. Kozan. 2006. An assignment model for dynamic load planning of intermodal trains. *Comput. Oper. Res.* **33** 1–17.

Drish, W. F. 1992. Train Energy Model Version 2.0 Technical Manual. Publication SD-485, Association of American Railroads, Chicago.

Engdahl, R. 1987. Full-Scale Rail Car Testing to Determine The Effect of Position-in-Train on Aerodynamic Resistance. Publication SD-705, Association of American Railroads, Washington, D.C.

Engdahl, R., R. L. Gielow, J. C. Paul. 1987. Train resistance—Aerodynamics Volume I of II intermodal car application. *Proc. Railroad Energy Technology Conf. II*, Association of American Railroads, Chicago.

Engdahl, R., R. L. Gielow, J. C. Paul. 1987. Train resistance—Aerodynamics Volume II of II open top car application. *Proc. Railroad Energy Tech. Conf. II*, Association of American Railroads, Chicago.

Feo, T. A., J. L. Gonzalez-Velrade. 1995. The intermodal trailer assignment problem. *Transportation Sci.* **29** 330–341.

Furlong, C. F. 1988. Aerodynamic Subroutine Users Guide. Publication SD-683, Association of American Railroads, Washington, D.C.

Gallamore, R. E. 1998. State of the art of intermodal freight transport in the United States. *Intermodal Freight Transport in Europe and the United States*, Chapter 2. Eno Transportation Foundation, Inc., Lansdowne, VA, 17–31.

Gielow, M. A., C. F. Furlong. 1988. *Results of Wind Tunnel and Full-Scale Tests Conducted from* 1983 *to* 1987 *in Support of The Association of American Railroad's Train Energy Program*. Publication SD-685, Association of American Railroads, Washington, D.C.

Giemsch, P., A. Jellinghaus. 2004. Optimization models for the containership stowage problem. D. Ahr, R. Fahrion, M. Oswald, G. Reinelt, eds. *Operations Research Proceedings* 2003. Springer, Berlin, 347–354.

Hay, W. W. 1982. *Railroad Engineering*, 2nd ed. John Wiley & Sons, Inc., New York.

Lai, Y.-C., C. P. L. Barkan. 2005. Options for improving the energy efficiency of intermodal freight trains. *Transportation Res. Record* 1916 47–55.

Lai, Y.-C., C. P. L. Barkan, H. Önal. 2008. Optimizing the aerodynamic efficiency of intermodal freight trains. *Transportation Res. Part E* **44**(5) 820–834. doi:10.1016/j.tre.2007.05.011.

Larsen, A. 2001. The dynamic vehicle routing problem, Ph.D. thesis, Institute of Mathematical Modeling, Technical University of Denmark, Lyngby.

Lund, K., O. B. G. Madsen, J. M. Rygaard. 1998. *Vehicle Routing Problems with Varying Degrees of Dynamism*. Technical report, Institute of Mathematical Modeling, Technical University of Denmark, Lyngby.

Optimization Alternatives Ltd. Inc. 2005. *Optimization Alternatives' Strategic Intermodal Scheduler* (*OASIS*). Retrieved September 1, 2007, http://www.oax.com.

Powell, W. B., T. A. Carvalho. 1998. Real-time optimization of containers and flatcars for intermodal operations. *Transportation Sci.* **32** 110–126.

Storm, B. 2005. Personal Communication. NASA Ames Research Center, Moffett Field, CA.

TTX Company. 1999. *Equipment Guide*. TTX Company, Chicago.

Union Pacific Railroad Corporation (UP). 2005. *Intermodal Loading Guide*, Union Pacific Railroad Corporation. Retrieved June 1, 2007, http://www.uprr.com/customers/dam-prev/attachments/intgenloadguide.pdf.

Utterback, M. 2006. Personal communication. BNSF Railway Company, Chicago.