STRATEGIES FOR EFFECTIVE RAIL TRACK CAPACITY USAGE

Final Report

METRANS Project

January 16, 2010

Investigators:

Maged Dessouky, Ph.D

Fernando Ordóñez, Ph.D

Robert Leachman, Ph.D

Graduate Student: Pavankumar Murali

University of Southern California

Daniel J. Epstein Dept. of Industrial and Systems Engineering

Los Angeles, CA 90089-0193

And

University of California, Berkeley

Department of Industrial Engineering and Operations Research

Berkeley, CA 94720-1777

METRANS
Transportation Center
USC CSULB

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, and California Department of Transportation in the interest of information exchange. The U.S. Government and California Department of Transportation assume no liability for the contents or use thereof. The contents do not necessarily reflect the official views or policies of the State of California or the Department of Transportation. This report does not constitute a standard, specification, or regulation.

Abstract

In the United States, railways are the major means to trans-continentally move goods from ports to the various inland destinations. Due to mergers and abandonment of rail lines, there has been a reduction in the track capacity, concentrating rail traffic to fewer lines. In addition to this, booming trade with Pacific Rim nations has seen the annual trade in the Ports of Los Angeles and Long Beach exceed 100 million tons with anticipation to double and possibly triple their cargo by 2020 (Leachman, 2002). The growth in the number of containers has already introduced congestion and threatened the accessibility and capacity of the rail network system in the Los Angeles area. Average transit times have stretched out in many corridors.

There is clearly a need among US freight railroads for better analytical tools to manage their capacity and scheduling. A challenging problem is determining the effect of shipments on a railroad, including estimation of the travel times and delays in the network plus determination of the most efficient method of scheduling these loads. This entails the ability to assign trains to routes based on the statistical expectation of running times in order to balance the railroad traffic, and to reject or defer shipment requests that would overload the network. In the first year of this research, we used mathematical modeling techniques to perform the former, that is, to be able to route and schedule trains on a railway network so as to minimize the travel time delays. In the second year we developed a simulation-based delay estimation methodology that would be able to estimate the travel-time delay over any given single-track or double-track rail network. These estimates can be used to reject or defer shipments that could cause congestion in the network.

Efficient solutions to the above problems are necessary to obtain an effective capacity planning and scheduling system for train networks. As a whole, this research represents an original effort in developing the first quantitative model to accept, defer or reject shipments on a railroad, with decisions based on an accurate representation of the delays these shipments cause on the railroad and the possibility of real-time rerouting trains to alternative tracks.

Contents

List of Figures

List of Tables

Disclosure

The project was funded in entirety under this contract to California Department of Transportation.

Acknowledgements

We would like to thank METRANS for funding this research.

1 Introduction

Until recently, almost the entire history of US freight railroads has been one of excess capacity. Until deregulation in 1980, management focus was on both cost reduction and service quality. Subsequent to deregulation, management focus on service quality dropped and focus on margin improvement increased. Several waves of mergers and abandonment of rail lines have eliminated much track capacity and other fixed assets, concentrating rail traffic on fewer lines. While rates of return have improved, at present US railroads do not earn their cost of capital (Winston, 2005).

Juxtaposed with the declining track capacity is the explosive growth in Asian imports and steady growth in rail movement of coal, chemicals, grain and other bulk commodities. As global trade continues to increase, cargo traffic at the nation's ports continues to increase at dramatic levels. For example, the Ports of Los Angeles and Long Beach (San Pedro Bay Ports) are among the busiest ports in America. Booming trade with Pacific Rim nations has seen the annual trade in the two ports exceed 100 million tons with anticipation to double and possibly triple their cargo by 2020 (Leachman, 2002). The total volume that these ports handle is evenly divided between transcontinental and local shipments. Furthermore, a large portion of the local shipments are re-packaged and/or sorted at local warehouse facilities for re-shipment across the continent. Railways form the major means to transcontinentally move these goods. The growth in the number of containers has already introduced congestion and threatened the accessibility and capacity of the rail network system in the Los Angeles area. As a result, various US rail lines have experienced severe congestion. Average transit times have stretched out in many corridors.

In the United States, passenger trains have utmost priority and follow timetables. Freight trains have a lower priority, and are sent on routes and scheduled to minimize their impact on passenger trains. There is considerable variability associated with train departures from their stations due to uncertainty in loading time and crew handling. However, the variabilities are more pronounced in the case of freight trains. It is well-recognized that the impact on the American economy from rail melt-downs is unfavorable and severe. Energy shortages, unemployment, inflation and decreased gross domestic product are consequences. The first serious melt-down in post-WWII western US rail operations occurred on the Southern Pacific railroad in the late 1970s in Texas and West Coast states, brought on by deferred maintenance in its locomotive fleet during the 1975-76 recession and the subsequent traffic boom. After the Union Pacific (UP) absorbed the Southern Pacific in 1996, there have been numerous very serious melt-downs in UP operations, continuing right up to the present. So far, the Burlington Northern Santa Fe (BNSF) has escaped serious melt-downs like those that have plagued the UP, but BNSF's main lines are precariously close to melt-down conditions during the late summer peak season.

In the eastern US, after the Conrail break-up about five years ago, both railroads created, CSX and NS, experienced serious melt-downs that took years to eliminate. While NS has largely recovered, CSX still is not fully fluid. Both railroads have lines that are attempting to cope with traffic levels close to melt-down conditions. Because the railroads historically never had to be concerned about a lack of capacity, no routine execution systems are in place to refuse shipments that might overload their networks. While some strategic actions have been taken of late by rail managements (e.g., surcharges and embargoes, strategic re-routing of train movements onto alternative lines or alternative times of week), in our opinion, capabilities of the industry in this regard are still very weak.

There is clearly a need among US freight railroads for better analytical tools to manage their capacity and scheduling. A challenging problem is determining the effect of shipments on a railroad, comprising estimation of the travel times and delays in the network plus determination of the most efficient method of scheduling these loads; each a complicated problem in its own right. Thus, to avoid melt-downs, several analytical capabilities are required, to wit:

1. The ability to translate shipment requests into line volumes, compare these volumes to established thresholds, and to reject or defer shipment requests that would overload the network.

2. The ability to assign trains alternative routes based on statistical expectation of running times in order to balance the railroad traffic.

Efficient solutions to the above problems are necessary to obtain an effective capacity planning and scheduling system for train networks. As a whole, the proposed research represents an original effort in developing the first quantitative model to accept, defer or reject shipments on a railroad, with decisions based on an accurate representation of the delays these shipments cause on the railroad and the possibility of real-time rerouting trains to alternative tracks. The former is important since in many urban areas, like Southern California, there are several different rail routes. For example, there are three distinct rail lines from the Colton crossing area to the Downtown area (two served by UP and one by BNSF). The capability to balance the freight rail traffic along the three routes has the potential to significantly reduce train delays in the area.

Our focus in the second year of this research project was on developing a delay estimation technique that models delay as a function of the train mix and the network topology. These delay estimates can be used to route and schedule trains over a large complex network. One way to reduce the complexity of routing and scheduling could be to "aggregate" suitable sections of a network in our analysis. An estimate of the expected travel time delay with traffic in the aggregated section would need to be fed to a routing and scheduling model so that trains can estimate, well in advance, the delay they would experience along each possible route to their destination. This delay estimation technique reflects on how the delay in the aggregated section of the network varies with each additional train. Once a delay estimation equation has been generated by our technique, it can be used to estimate delay and capacity of a network section with physical attributes within the range of those of the generic networks used in the experiments. That is, individual simulations need not be run for every aggregated network section.

Various kinds of estimation techniques can be used to study how the delay varies with a change in the traffic conditions and/or the railway network topology. The difference between the actual running time and the free running time is termed as *travel time delay* or, simply, *delay*. The free running time of a train over a network is defined as the time the train takes to traverse the network, when traveling at its maximum allowable speed and not experiencing delays due to other train(s). The actual running time is the time a train travels to reach its destination when there are other trains in the network. For freight trains, delays can be of two types, namely, direct and knock-on (or indirect) delays. Direct delays to trains are a consequence of minor delays at a station. These are not as a result of other trains traveling along the same lines. Knock-on delays are those which are induced into the system due to a direct and/or knock-on delay to another train in the network. It is transferred from one train to, possibly, all the other trains in the vicinity.

The capacity of a railway network and the delay across it are closely related. The delays encountered by trains under different operating assumptions can be used to evaluate the capacity of a section of a network, which is referred to as a *subnetwork*. Capacity can be defined as the maximum number of trains that can traverse a network, or a section of a network, without resulting in a deadlock. Burdett and Kozan (2006) define absolute capacity as the theoretical capacity value that is realized only when critical sections of a network are saturated. On the other hand, actual capacity is the number of trains that can safely coexist in a network, or a portion of it, when interference delays are taken into consideration. Both measures of capacity are measured over time. Absolute capacity can be used as an upper bound for planning purposes.

The actual capacity of any section of a railway network cannot be a unique value, and it is neither easily defined nor quantified. It depends on the average minimum headway time between consecutive trains, the signalling system, train speeds, trackage configuration etc. For instance, single-tracks with sidings can accommodate more trains and enable crossings and overtakes than those without. Signals can increase capacity by reducing the required headway between trains. Delay estimation and capacity analysis in railway transportation is dependent on various operational aspects. The first aspect is the trackage configuration. The network can consist of single, double, triple or even more track. Single tracks are common in the case of North America while double and triple tracks are common in Europe. Normally, the level of complexity in urban areas is higher than in rural areas because they contain many junctions. A train can block the movement of other trains when it tries to cross over at a junction from one line to another. The second aspect is the variation in speed limits on different track segments and junctions. Furthermore, passenger trains and freight trains can have different maximum speeds even though their paths may use the same tracks, but not necessarily at the same time. If a train passes a junction by changing lines, the speed limit at the junction will be enforced. While single speed limits are common on networks in rural areas, multiple speed limits are common in metropolitan areas. A lower speed-limit over a subnetwork tends to increase travel time delays. The third aspect is the characteristic of each train in the rail network such as priority, train length, speed, acceleration rate and deceleration rate. Generally, passenger trains have higher priority than freight trains. If two trains want to seize the same track simultaneously, the train with the lower priority should wait and stop until the train with higher priority passes. Sometimes trains cannot be dispatched at their maximum speed because of the track speed limit. Acceleration and deceleration rates need to be considered in order to increase or reduce speed without violating the speed limit. This results in a nonlinear function to represent the movement of trains.

This report is organized as follows. In Section 2 we present a brief review of the work done in developing routing and scheduling models for railways, and the analytical and simulation modeling techniques that have been used in the past for estimating travel-time delays. In Sections 3, 4 and 6 we discuss our main research accomplishments - an integrated capacity management model that uses regression-based delay estimation equations to route trains such that delays are kept at a minimum. We present an approximation scheme to solve this capacity management model in Section 7 and, finally, end with a few concluding remarks and prospective directions for future research in Section 9.

2 Literature Review

There is clearly a need among US freight railroads for better analytical tools to manage their capacity and scheduling. In order to have the ability to minimize travel times and delays in delivering freight goods, we should be able to have each train start from its origin at a time and travel on a route that minimizes travel time, meet/pass interferences and expected delays. To accomplish this, it is imperative to have an integrated routing and capacity management model, instead of a simple scheduling model that assumes the routes to be fixed. Although there exists a slew of problems that have been studied in rail transport research, there has been little work in developing an integrated approach for the capacity planning and scheduling problems that considers all the complicating aspects of rail operations. In particular, in the capacity planning area, there are no models to guide planners in determining whether to accept or reject a shipment based on how it would affect the travel-time delay in a network. Although there has been some work in the scheduling area, it does not take into account alternative routes. We next briefly review the prior work on rail operations. Next, in sections 2.1 and 2.2 we present an overview of prior work done in the areas of rail routing, scheduling and delay predictions to reflect our two main contributions, namely, a routing and scheduling model and delay estimation techniques.

2.1 Rail Routing and Scheduling

For a given railroad network, planning can be broadly classified into three levels, namely, train routing, train scheduling (or timetabling) and train dispatching (see Cordeau et al, 1998). *Train routing*, done at a strategic level, determines the track segments each train needs to be routed on, so as to minimize the expected delay for each train, while considering the capacity of these segments. The routing problem also deals with assigning origin-destination pairs to cars, assembling and disassembling cars into blocks, and grouping and ungrouping blocks into trains. The *train scheduling* problem addresses the issue of developing operational timetables, usually for passenger trains, taking into account train speeds, and headways and buffer time between trains. This is performed at the tactical level. The *train dispatching* problem deals with real-time operations involving precise synchronization of freight and passenger train movements on the lines of the physical railway network. Given a train timetable, the train dispatching problem determines a feasible plan of meets and overtakes that satisfies a system of constraints.

There are many problems that fall into the general category of the routing problem. These include blocking models and makeup models. The railroad *blocking model* determines which cars should be assembled into which blocks at which yards, and places emphasis on the movement of cars as opposed to the movement of trains. Newton (1996) models this as a network optimization problem with nodes as yards and arcs as potential blocks. The objective is to minimize the cost of delivering all commodities, and is solved using Dantzig-Wolfe decomposition. Barnhart et al. (2000) formulate the blocking problem as a network design problem, and use a Lagrangian relaxation approach to solve the mixed integer program. These optimization-based formulations of the blocking problem do not account for the complexity of this problem, and do not produce good quality solutions that can be implemented. Ahuja et al. (2007) propose a very large-scale neighborhood (VLSN) search algorithm to solve real-life instances of the blocking problem. This algorithm has two main subroutines: constructing the initial feasible solution, and re-optimizing the blocking arcs emanating from a node. The *makeup model* assigns blocks to trains, and the routing model determines the routing and frequency of trains. Crainic et al. (1984) propose a nonlinear MIP by integrating the blocking, the makeup and the routing models all together. Keaton (1992) uses Lagrangian relaxation for solving similar combined blocking and routing problems.

In this research work, we focus on train routing and train scheduling. By *routing*, we mean given a set of possible routes in going from an origin to a destination or a set of destinations, a train is made to travel on a route with the least expected delay. Delays occur when trains traveling in either the same direction or opposite direction meet, thus requiring one of the trains to be pulled over for the other to cross or overtake it. Routing is typically done at a macro level, that is, it does not deal with which siding to travel on, which crossing should be used, which track segment of a double- or triple-track should the train travel on etc. We note that most of the prior work in the area of railway planning deals with railway scheduling without routing. However, models that address both routing and scheduling aspects make it easier to find a timetable for all the planned trains, and hence improve service levels and reduce delays.

There is no dearth in the amount of research done in train scheduling. We list some of the most recent and important literature in this area. Huntley et al.(1995) propose a routing and scheduling model with an objective to minimize operational costs. It is solved using simulated annealing techniques and the output is the sequence of train links to be followed by each demand block from origin to destination. Higgins et al. (1995) address the problem of scheduling trains on a single line track when the priority of each train in a conflict depends on an estimate of the remaining crossing and overtaking delay. This priority is used in a branch and bound procedure to allow the determination of optimal solutions. Carey and Lockwood (1995) propose a model, algorithms and strategies for the train pathing (or routing) and timetabling problem. As the pathing problem is computationally intensive, they develop heuristics to solve sub-problems to route a single train while holding the order of all the trains fixed. Hallowell and Harker (1998) consider the problem of scheduling trains on a partially double track rail line by accounting for delays due to meet/pass interference. Gorman (1998) also considers an integrated routing and scheduling problem, and uses genetic and tabu search algorithms to solve it. The model has binary variables for potential train services and assignment of demand blocks to a train. The objective is to minimize the sum of fixed costs of trains and marginal cost per car. Caprara et al. (2002) propose a graph theoretic model for the timetabling problem on a one-way, single track network. They develop an integer program model and use a Lagrangian relaxation approach to solve it heuristically. Dessouky et al. (2006) propose a mixed integer programming model for train dispatching in a complex network by considering speed limits and headways. They use branch and bound techniques to solve the problem. In most of the research work done so far on railway routing, the routing almost always refers to railway cars routing and not train routing. Also, most of the papers mentioned above present models that assume train routes to be known, thus concentrating on train scheduling and dispatching alone.

2.2 Capacity Management & Delay Estimation

For efficient and cost-effective train routing, a modeling strategy needs to account for track capacity. Ideally, we would like to route the trains such that the various sections of the track infrastructure are, more or less, uniformly utilized. This would reduce travel time and the probability of experiencing knock-on delays. More importantly, incorporating capacity constraints into routing and scheduling models enables us to determine if the existing trackage can handle any new trains, and identify the sections where additional tracks could be added in order to increase the capacity of the entire network. A network (or sub-network) is saturated if the addition of a new train results in a sudden increase in the delays for any or all trains. Thus, capacity analysis is vital in identifying the traffic thresholds at which travel times would exceed service standards. Some work has been done in developing capacity management models.

2.2.1 Analytical Models

One of the earliest analytical models on capacity and delay assessment was developed by Frank (1966). He studied delay on a single track with unidirectional and bidirectional traffic. By restricting only one train on each link between sidings and using single train speeds and deterministic travel times, he estimated the number of trains that could travel on the network. Petersen (1974) extended this work to accommodate for two different train speeds. He assumed independent and uniformly distributed departure times, equally spaced sidings and a constant delay for each encounter between two trains. Chen et al. (1990) extended Petersen's model to present a technique to calculate delay for different types of trains over a specified single track section as a function of the schedules of the trains and the dispatching policies. They assumed sidings to be equally distributed, that faster trains can overtake slower trains, meets and overtakes occur only between 2 trains at a time, and there exists a fixed probability $P_{i,j}$ of a train i getting delayed by a train j. This modeling technique was extended by Parker et al. (1990) to a partially double-track rail network which consisted of a single-track section with sidings and double-track sections. Similar to the previous work, trains depart according to their scheduled departure times. The train to be delayed during a meet (or overtake) is determined by a trade-off between the lateness of the train with respect to its schedule and the overall priority of the train. Carey et al. (1994) studied the effects of knock-on delays between two trains on a single-track. They used non-linear regression to develop stochastic approximations of the relation between scheduled headways and knock-on delays, and tested these approximations by conducting detailed stochastic simulation of the interactions between trains as they traverse sections of the network. Özekici et al. (1994) used Markov chain techniques to study the effects of various dispatching patterns and arrival patterns of passengers on knock-on delays and passenger waiting times. Given a travel time probability density function for a train on a track link, a departure time transition matrix was constructed for the calculation of the expected departure delay. Higgins et al. (1998) presented an analytical model to quantify the positive delay for individual passenger trains, track links and schedule as a whole in an urban rail network. The network they considered has multiple unidirectional and bidirectional tracks, crossings and sidings. Yuan (2006, 2008) proposed probability models that provide a realistic estimate of knock-on delays and the use of track capacity. The proposed model reflects speed fluctuation due to signals, dependencies of dwell times at stations and stochastic interdependencies due to train movements. D'Ariano (2008) studied delay propagation by decomposing a long time horizon into tractable intervals to be solved in cascade, and using advanced Conflict Detection and Resolution with Fixed Routes (CDRFR) algorithms. These algorithms are used to detect and globally solve train conflicts on each time interval.

Queuing theory is another methodology that has been used for estimating delay in railroads. Greenberg et al. (1988) presented queuing models for predicting dispatching delays on a low speed, single track rail network supplemented with sidings and/or alternate routes. Train departures are modeled as a Poisson process, and the slow transit speed and deterministic travel times enable them to travel with close headways. This work assumes sidings to have infinite capacity. Huisman et al. (2001) investigated delays to a fast train caught behind slower ones by capturing both scheduled and unscheduled movements. This is modeled as an infinite server $G/G/\infty$ re-sequencing queue, where the running time distributions for each train service are obtained by solving a system of linear differential equations. Wendler (2007) presented an approach for predicting waiting times using a $M/SM/1/\infty$ queuing system with a semi-Markovian kernel. The arrival process is determined by the requested train paths. The description of the service process is based on an application of the theory of blocking times and minimum headway times.

A bottleneck approach is one way to determine the absolute capacity of a network, by identifying the maximum number of trains that can travel through the track segments constituting a bottleneck in a given time period. De Kort et al. (2003) considered the problem of determining the capacity of a planned railway infrastructure layout under uncertainties for an unknown demand of service. The capacity assessment problem for this generic model is translated into an optimization problem. Burdett and Kozan (2006) developed capacity analysis techniques and methodologies for estimating the absolute (theoretical) traffic carrying ability of facilities over a wide range of defined operational conditions. Specifically, they address the factors on which the capacity of a network depends on, namely, proportional mix of trains, direction of travel, length of trains, planned dwell times of trains, the presence of crossing loops and intermediate signals in corridors and networks. Gibson et al.

(2002) also developed a regression model to define a correlation between capacity utilization and reactionary delay. Landex et al. (2006) and Kaas (1998) discussed techniques to calculate capacity utilization for railway lines with single and multiple tracks, as per the UIC (International Union of Railways) 406 method.

2.2.2 Simulation Models

Simulation techniques can be used to study direct, knock-on and compound delays and ripple effects from conflicts at complex junctions, terminals, railroad crossings, network topology, train and traffic parameters. The compound interaction effects of these factors cannot be effectively captured in an analytical delay estimation model. Petersen et al. (1982) present a structured model for rail line simulation. They divide the rail line into track segments representing the stretches of track between adjacent switches and develop algebraic relationships to represent the model logic. Dessouky et al. (1995) use a simulation modeling methodology to analyze the capacity of tracks and delay to trains in a complex rail network. Their methodology considers both single and double-track lines and is insensitive to the size of the rail network. Their model has a distinctive advantage of accounting for track speed-limits, headways, and actual train lengths, speed-limits acceleration and deceleration rates in order to determine the track configuration that minimizes congestion delay to trains. This work is extended by Lu et al. (2004). Hallowell et al. (1998) improve upon the work by Parker et al. (1990) by incorporating dynamic meet/pass priorities in order to approximate an optimal meet/pass planning process. Extensive Monte Carlo simulations are conducted to examine the application of an analytical line model for adjusting real-world schedules to improve on-time performance and reduce delay. Krueger (1999) uses simulation to develop a regression model to define the relationship between train delay and traffic volume. The parameters involved are network parameters, traffic parameters and operating parameters.

3 Research Accomplishments

As mentioned previously, there is very little work done in developing train routing models. Although Carey and Lockwood (1995) address this problem, their technique only provides a local solution since each train is routed individually. Moreover, similar to many papers, they test their routing model on a single unidirectional line, thereby not accounting for the complexities and non-linearities of railway operations. During the two years of this research project, our efforts were mainly concentrated in addressing this gap. The tasks that were accomplished during this period can be broadly classified as follows.

1. Develop an integer programming (IP) model which optimally routes the trains between their respective origin and destination points. An "optimal" route is defined as one that has the least expected delay in travel time associated with it.

2. Develop a regression-based delay estimation technique using simulation models. This can be used to estimate delay across a sub-network at various traffic levels.

3. Extend this delay estimation technique to predict delays for generic single- and double-track network segments, thereby, saving the effort of running simulations on individual segments.

4. Design a solution procedure to solve the aforementioned IP, and compare the performance of the procedure to that of other heuristic procedures, such as Lu et al. (2004).

A large portion of the research on railway planning handles train scheduling and dispatching at a micro-level, to determine every track segment, junction and siding each train travels on. With the aid of aggregation techniques, which are explained further on, our modeling procedure is carried out at a macro-level. The integrated routing and capacity management model is capable of incorporating route flexibility as opposed to having a train follow a fixed path from origin to destination, and of handling large size railway networks with many more trains. These features constitute the primary differences between our routing model and the scheduling model proposed by Dessouky et al. (2006).

In addition to routing trains, the IP model also decides the order in which the trains leave their origin points, the times at which they depart each node and the sequence in which they pass each other at crossings. A novel feature is the incorporation of track capacity into the routing model. The use of capacity-delay correlations enables the adjustment of the capacity of a track or sub-network to have admissible delays, and to reject or defer shipment(s) that would overload the network. In the following sections, we present a detailed description of the routing and scheduling IP model, the regression-based delay estimation techniques and compare the results obtained by running this integrated routing and capacity management model on the railway network in the Southern California region to those obtained by using the construction heuristic proposed by Lu et al. (2004) on the same network.

In the second year of our research, we concentrated our efforts on developing a simulation-based delay estimation methodology that is capable of estimating travel-time delays on single-track and double-track rail networks. A majority of the prior work on delay estimation and capacity assessment for railway networks does not explicitly consider the vital and complex interactions between traffic, operating and network parameters. In the case of the analytical models, heavy assumptions are made in order to maintain the complexity of the problem within solvable bounds, thereby rendering the problem to be far off from real-life rail operations. Furthermore, these models may be incapable of recognizing the dynamic nature of capacity and knock-on delays involving more than two trains. More often than not, delay or capacity estimation is unlikely to be the final step in railway operations planning. Instead, a dispatcher might use these estimated values in railway routing and scheduling, that is, to route a set of trains over tracks with the minimum expected delay so as to minimize the overall system delay. For such purposes, it would be beneficial to design simple delay estimation models that could be easily integrated with or incorporated into a routing, scheduling or dispatching model. Analytical models requiring algorithms to solve a system of equations might not be the best option for this purpose. Simulation models, on the other hand, enable us to develop simple, yet accurate, algebraic relationships that better capture the stochastic nature of the interactions between the traffic, operating and network parameters, and their impact on travel time delays.

Due to these reasons, we used simulation techniques to develop accurate and simple delay estimation models that can then be used with railway routing and scheduling models. Delay is modeled as a function of traffic parameters such as train lengths, acceleration and deceleration rates, operating parameters such as prevailing speed-limits, headways etc., and network characteristics such as double or single-track, number of crossings or sidings and the spacing between them. Once a delay estimation equation has been generated by our technique, it can be used to estimate delay and capacity of a network section with physical attributes within the range of those of the generic networks used in the experiments. That is, individual simulations need not be run for every aggregated network section.

4 Routing and Scheduling Model

In this section, we provide a formal description of the integrated train routing and capacity management model. We use the same network definition of a rail system as proposed by Lu et al. (2004). Given a railway network consisting of main tracks, sidings, junctions and platforms, it can be converted to a general network $G = (N, A)$, where N is the node set and A is the arc set. Each node $j \in N$ contains a set of segment resources, whose lengths are equal to the length of the longest train to be routed. The time spent by a train at a node j will be at least equal to the time needed to traverse the segment(s) included in that node. The capacity of these nodes depends on the number and type of track segments included in the node. The stations existing in the network are also represented by nodes. The capacity of these station-nodes is greater than 1, in order to accommodate crossing and overtaking. The nodes are connected to each other by directed arcs $a \in A$ which do not have any length or speed limit associated with them. There are two directed arcs, one for each direction, between any two nodes i and j. To replicate the physical network, only one of these two arcs can be simultaneously occupied. Similarly, each node j can be occupied by an additional train as long as the capacity of the segments constituting the node is not violated. Hence, the time taken by the trains to travel in the transformed network will be exactly the same as in the physical network. In Figure 1, we show an example of how to translate an actual track configuration into network G.

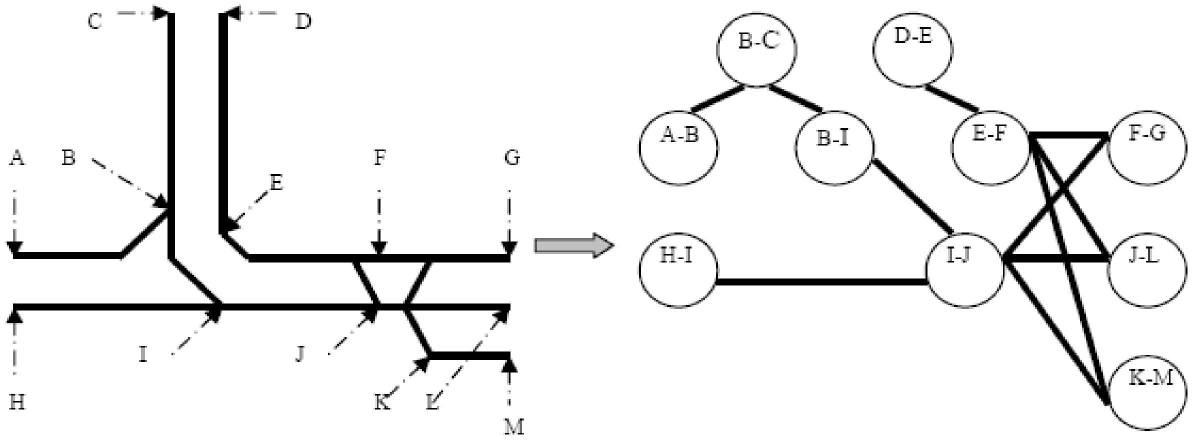


Figure 1: Conversion from a physical network to a general network

For each train $h \in H$, the origin and destination stations are known, but the route can be either known or unknown. If the route is unknown, the IP model would route and schedule all the trains. In this case, each train is routed along the track segments (nodes) with the least overall delay. On the other hand, if the route is known, then the IP model would just be a train scheduling model. As in any train scheduling model, the time of departure for each train from its origin, and the meet/pass order are optimally determined by the model.

To represent a train h traveling from node i to node j at time t, we introduce a binary variable indexed in i, j, h and t, where this variable is equal to 1 if train h travels from node i to node j at time t. The IP model is run over a time horizon T. As in any network, we have trains traveling in both directions. This modeling procedure generates an IP model with a large number of binary variables that, even for small-scale applications, is not solvable in any reasonable time using standard solvers such as CPLEX. For example, suppose we have 14 trains, with 7 in each direction, to be routed over a network with a 100 directed arcs and the model is run till T = 300. This gives us 420,000 binary variables. To make the problem more tractable, we define binary variables Y to denote a train h traveling from node i to node j *by* time t (instead of *at* time t as described above). This reduces the number of variables in each constraint, thereby making the model sparser. We also perform pruning by defining Y only for possible directions of travel for each train. For instance, for a train h that could travel only from node a to node b on its way from its origin to its destination, we define Y indexed only in a, b, h, T, and not in b, a, h, T. This pruning step halves the number of binary variables. In contrast to Dessouky et al. (2006), we ignore train lengths, thereby reducing the number of constraints. We believe this assumption does not significantly skew the results due to our procedure of dividing the network into segments equal to the train lengths, and also due to node capacity constraints included in the IP model. We also make an assumption that the trains have very high acceleration and deceleration rates to enable them to instantly change their speeds to follow the speed limits of the track segments they are traveling on.

4.0.3 Notation

We consider a set of H trains traveling through a railway network. As outlined in Lu et al. (2004), each node in the transformed network has two ports: port 0 and port 1. Port 0 indicates the starting point of travel for a train moving in the node from one direction. Port 1 indicates the starting point of travel in the opposite direction of port 0. If a train h enters a node from port 0 (respectively, port 1), then it must leave from port 1 (respectively, port 0). We use M as a large constant to express non-linear relationships through linear constraints. We define the following *parameters* and *binary decision variable*:

H_1: set of trains heading from 0 to 1

H_2: set of trains heading from 1 to 0

N: set of nodes representing track segments

A: set of directed arcs

A_1: set of arcs directed from ’1’ of the previous node to ’0’ of the next node

A_2: set of arcs directed from ’0’ of the previous node to ’1’ of the next node

v_i: velocity limit of node i

l_i: length of node i

c_i: capacity of node i

r_h: earliest departure (or release) time of train h

o_h: origin node of train h

d_h: destination node of train h

T: time horizon

s_1: set of (A_1, H_1)

s_2: set of (A_2, H_2)

S: $s_1 \cup s_2$

M: a very large number

$Y_{i,j,h,t}$: flow variables defined such that (i, j, h) in S, take a value of 1 if train $h \in H$ traverses on (i,j) $\in A$ at any time between 0 and t and 0 otherwise

4.0.4 IP Model

The integrated train routing and capacity management model is given below. The objective function of this integer programming model is to minimize the sum of all the arc traversal times for each train, thereby indirectly minimizing the total travel time (and delay) for all trains. Weights can be associated with the travel time of each train to represent any priority that may exist. Here, we give an objective function assuming all trains are of equal importance. This can be written as follows:

$$\textbf{IP}: \quad \begin{array}{ll} \text{Minimize} & \displaystyle\sum_{(i,j,h)\in S} \sum_{t=0}^{T} t[Y_{i,j,h,t} - Y_{i,j,h,t-1}] \\ \\ \text{subject to} & \text{constraints } (1) - (10) \end{array}$$

where the constraints are explained in detail below.

The following constraint (1) ensures that a train does not leave its origin station prior to its earliest departure time.

$$\sum_{(i,j)|(i,j,h)\in S, i=o_h} Y_{i,j,h,r_h-1} = 0 \qquad \forall h \in H \tag{1}$$

The following constraint (2) ensures that a train must eventually leave its origin station after the earliest departure time. Each train is allowed the flexibility to leave at an appropriate time so as to minimize the total delay in the network.

$$\sum_{(i,j)|(i,j,h)\in S, i=o_h} [Y_{i,j,h,T} - Y_{i,j,h,r_h-1}] = 1 \qquad \forall h \in H \tag{2}$$

The following constraint (3) enforces the condition that a train has to arrive at its destination station.

$$\sum_{(i,j)|(i,j,h)\in S, j=d_h} Y_{i,j,h,T} = 1 \qquad \forall h \in H \tag{3}$$

Constraint (4) enforces the condition that a train must take at least (length of segment/velocity limit over that segment) units of time to traverse the node representing that segment. Constraint (5) takes care of flow conservation, thereby ensuring that every train entering a node leaves the node after a certain amount of time.

$$\sum_{i|(i,j,h)\in S} Y_{i,j,h,\tau} - \sum_{k|(j,k,h)\in S} Y_{j,k,h,\tau+\lceil l_j/v_j \rceil} \geq 0 \qquad \forall \left(j \in N | j \neq o_h, j \neq d_h\right), \quad h \in H,$$
$$\forall \tau \in 0, \ldots, T - \lceil l_j/v_j \rceil \tag{4}$$

$$\sum_{i|(i,j,h)\in S} Y_{i,j,h,T-\lceil l_j/v_j \rceil} - \sum_{k|(j,k,h)\in S} Y_{j,k,h,T} = 0 \qquad \forall \left(j \in N | j \neq o_h, j \neq d_h\right),$$
$$\forall h \in H \tag{5}$$

Constraints (6)-(7) are capacity constraints for the nodes. In each time period τ the number of trains traveling in node j must be less than c_j.

$$\sum_{i,h|(i,j,h)\in s_1} Y_{i,j,h,\tau} - \sum_{k,h|(j,k,h)\in s_1} Y_{j,k,h,\tau} \leq c_j \qquad \forall j \in N, \quad \tau \in 0, \ldots, T \tag{6}$$

$$\sum_{i,h|(i,j,h)\in s_2} Y_{i,j,h,\tau} - \sum_{k,h|(j,k,h)\in s_2} Y_{j,k,h,\tau} \leq c_j \qquad \forall j \in N, \quad \tau \in 0,\ldots,T \tag{7}$$

Constraints (8)-(9) ensure that trains traveling in opposite directions cannot enter the same node at the same time. If a node is already occupied by a train, then a train traveling in the opposite direction would either need to wait for this node to be freed or may choose an alternative node to reach its destination. Trains traveling in the same direction can occupy the same node at the same time, provided that the capacity constraint for that node is not violated.

$$\sum_{a,h|(a,j,h)\in s_2} Y_{a,j,h,t} - \sum_{k,h|(j,k,h)\in s_2} Y_{j,k,h,t} \leq M\left[1 - \sum_{k|(k,j,h')\in s_1} (Y_{k,j,h',t} - Y_{k,j,h,t-1})\right]$$
$$\forall \left(j \in N | j \neq o_h, j \neq d_h\right), \quad h' \in H, \quad t \in 0,\ldots,T \tag{8}$$

$$\sum_{a,h|(a,j,h)\in s_1} Y_{a,j,h,t} - \sum_{k,h|(j,k,h)\in s_1} Y_{j,k,h,t} \leq M\left[1 - \sum_{k|(k,j,h')\in s_2} (Y_{k,j,h',t} - Y_{k,j,h,t-1})\right]$$
$$\forall \left(j \in N | j \neq o_h, j \neq d_h\right), \quad h' \in H, \quad t \in 0,\ldots,T \tag{9}$$

Constraint (10) ensures that an arc is not simultaneously occupied by trains traveling in the opposite direction, since this could lead to a deadlock.

$$\sum_{h\in H|(i,j,h)\in S} (Y_{i,j,h,t} - Y_{i,j,h,t-1}) + \sum_{h\in H|(j,i,h)\in S} (Y_{j,i,h,t} - Y_{j,i,h,t-1}) \leq 1$$
$$\forall t \in 1,\ldots,T, \quad (i,j) \in A_1 | \left(j,i\right) \in A_2 \tag{10}$$

In order to make significant contributions to real-world railway routing and scheduling, we need to be able to run our IP model for large railroad networks. That is, we need to do train routing and scheduling at a macro-level, as opposed to a micro-level. For this purpose, in addition to the aforementioned pruning, we resort to *Aggregation*. Aggregation refers to combining a suitable and sizeable portion of the network under consideration into a single node in the graph G. Clearly, aggregation helps to reduce the problem size by reducing the number of nodes and arcs in G for a given network. This would help us to carry out train routing and scheduling over a larger network, such as the entire Southern California region, with many more trains. In the process, we increase the capacity of a node to represent the combined capacity of all the segments included in it. We follow some fundamental guidelines for aggregation - each node can have only similar type of track segments. We cannot combine single, double and triple track segments into a single node. For a given network G, we select a sub-network with end-points, say, A and B, that conforms to this condition. We calculate the total number of paths from A to B, denoted by c. Now, in G, we substitute this sub-network with two nodes, one for each direction and each with capacity $\lfloor c/2 \rfloor$. If the capacity constraint for a node allows more than one train to exist in a node at the same time, then we allow for the possibility of these co-existing trains to overtake each other. In the next section, we compare the total travel time from our IP model under aggregation with the travel time attained by using the simulation model by Lu et al. (2004).

4.1 Planning Model Experimental Results

In this section, we compare the results from our IP model with the results obtained from a greedy construction heuristic implemented in the simulation model presented in Lu et al. (2004), and the mixed integer program for train scheduling presented in Dessouky et al. (2006). Recall the model by Dessouky et al. (2006) uses actual train lengths but does not allow for flexible routing. The network used for running the three models is shown below in Figure 2. For running the IP model, we transform this physical network into a general network as described in the previous section.

The transformed network has 42 nodes and 52 directed arcs in each direction. Similar to Dessouky et al. (2006), we have 14 trains traveling across the above rail network. All trains are assumed to be ready at time zero. The origin and destination for the trains are given below in Table 1 below.

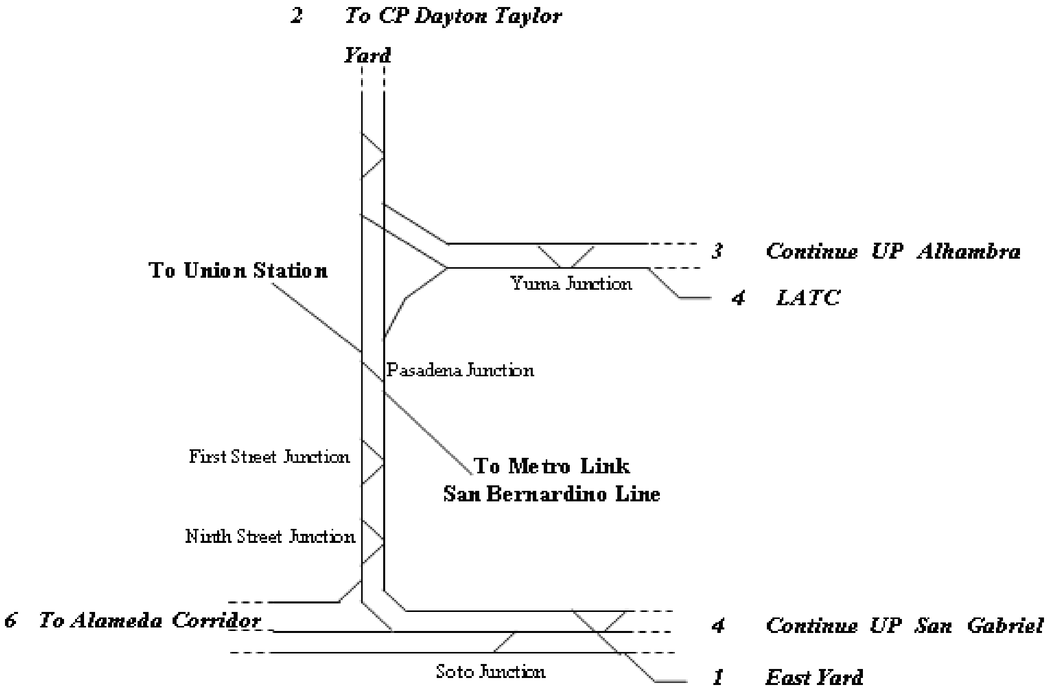


Figure 2: A Portion of the Rail Network near Downtown Los Angeles

RouteID	Origin	Destination
1, 13	CP Dayton Taylor Yard	Alameda Corridor
2,14	Alameda Corridor	CP Dayton Taylor Yard
3	LATC	Alameda Corridor
4	Alameda Corridor	LATC
5	CP Dayton Taylor Yard	LATC
6	LATC	CP Dayton Taylor Yard
7	Union Station	Metrolink San Bernardino Line
8	Metrolink San Bernardino Line	Union Station
9	CP Dayton Taylor Yard	East Yard
10	East Yard	CP Dayton Taylor Yard
11	East Yard	Alameda Corridor
12	Alameda Corridor	East Yard

Table 1: Route Data Specification

The IP model for routing these 14 trains is run using CPLEX 9.0 solver on a Linux server with a 3.06GHz Intel Xeon CPU for 2.0 CPU hours. The data files for the simulation model were set up for the same physical network, and for the 14 trains with their respective origins and destinations. It is run using AweSim! 2.0 simulation software. The results are tabulated in Table 2 below.

Model	Travel time (hr)	Decrease delay	Increase delay
new IP model (flexible)	LB:13.33 UB:15.13	no train length, flexible path	discretized time. t=5 minutes
new IP model (fixed)	15.53	no train length	discretized time, fixed path. t=4 minutes
Greedy heuristic, Lu et al. (2004)	14.103	continuous time, flexible path	heuristic approximation, train lengths
IP model, Dessouky et al. (2006)	LB:13.6 UB:14.5	continuous time	train lengths, fixed path

Table 2: Total travel times for 14 trains (in hours)

In the first row, we present the results from running the train routing IP model, presented in the previous section, over the network shown in Figure 2. To make the problem size tractable so that it can be solved by CPLEX in a reasonable amount of time, we modify t to represent 5 minutes. After running the routing model for 2 CPU hours, we get the lower and upper bounds on the optimal solution, which are presented in the table above. The second row represents the results we derive from our new IP model by forcing the trains to follow the same routes as used by Dessouky et al. (2006). For this, we resort to a 4 minute rounding to run our IP model in CPLEX. The third row represents the results obtained from the simulation model, presented in Lu et al. (2004), by having the trains follow the route determined by the new IP model under flexible routing. Finally, in the fourth row, we show the results presented by Dessouky et al. (2006) for the same network and set of trains.

In Table 2, the third column lists the features and assumptions of each modeling strategy that cause the travel times of the 14 trains to decrease, and the fourth column lists the reasons that cause the travel times to increase. The simulation model presented by Lu et al. (2004) and the IP model by Dessouky et al. (2006) have the advantage of using continuous time in their modeling procedure. However, using actual train lengths together with continuous time modeling limits the size of the railway network to which these two models can be applied. The integrated routing and capacity management IP model has an advantage of being capable of performing train routing on large real-life railway networks with many more trains. There are two reasons for this to be possible, namely, assuming train lengths to be negligible and using aggregation techniques. In our model, we discretize time to control the IP problem size so that it can be solved by CPLEX.

5 Aggregation

In order to be able to perform routing of trains over a large network and estimate travel time precisely, we test our IP model by applying aggregation to a portion of the railway network in Southern California, shown in Figure 3. It stretches from the *CP Dayton Taylor Yard* to the *City of Industry*, through *Downtown Los Angeles*, a total of nearly 50 miles.

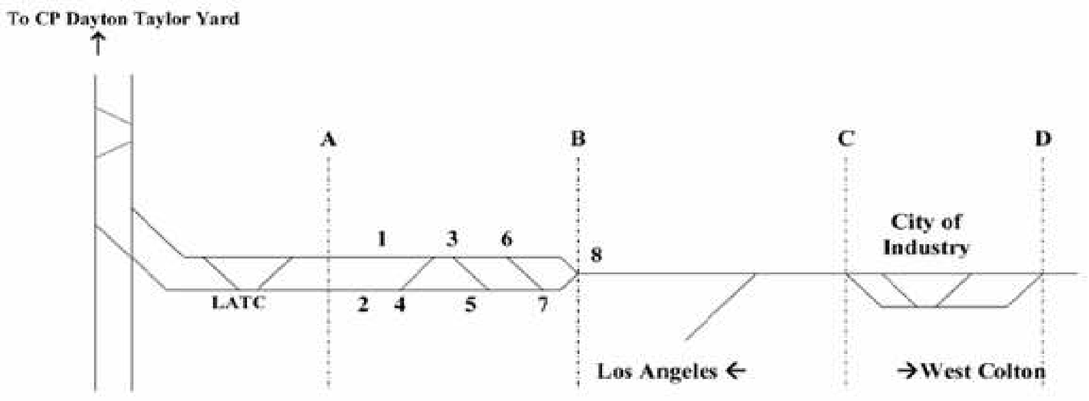


Figure 3: Union Pacific - Alhambra Rail Network near Downtown Los Angeles

The trains used for this network are given in Table 3. All trains are assumed to be ready at time zero.

The transformed network G has 38 nodes and 35 arcs. In order to be able to accurately compare the results from our IP model with the results from the simulation model, presented in Lu et al. (2006), we resort to a one minute rounding in our IP. Hence, even though the

RouteID	Origin	Destination
1, 7	City of Industry	CP Dayton Taylor Yard
2, 8	CP Dayton Taylor Yard	City of Industry
3, 9	City of Industry	LATC
4, 10	LATC	City of Industry
5, 11	LATC	CP Dayton Taylor Yard
6, 12	CP Dayton Taylor Yard	LATC

Table 3: Route Data Specification

network in Figure 3 is slightly smaller than the network in Figure 2, due to a one minute rounding, the IP problem size is larger in the case of the network in Figure 3. In order to solve the routing IP model for this network with a minute rounding, we can only have trains 1 to 6 from Table 3. We use N_a to denote the set of aggregated nodes.

There are two sections in G that can each be aggregated into a single node - section A-B and section C-D. For section A-B, there exist six paths between A and B, namely, 1-3-6-8, 1-3-6-7-8, 1-3-5-7-8, 2-4-5-7-8, 2-4-3-6-8 and 2-4-3-6-7-8. Hence, this sub-network can be substituted with two nodes, one for each direction and each with capacity equal to three. In this case, due to aggregation, we can have trains 1 to 10 from Table 3 traveling through the network, still using a minute rounding. Similarly, section C-D has four paths and can be substituted with two nodes, one for each direction and each with capacity equal to two. This allows us to have trains 1 to 12 traveling through the network using a one minute rounding.

The IP model, presented in this paper, and the simulation model, presented in Lu et al. (2006), are each made to run on the network given in Figure 3 for three scenarios - non-aggregated network, aggregation for section A-B, aggregation for sections A-B and C-D. In the case of the IP model, each scenario is made to run for 4.0 CPU hours, using the CPLEX 9.0 solver. The results are presented in Table 4.

From these results, we gather that our aggregation technique, when combined with the integrated routing and capacity management IP model, performs well in comparison to the simulation model.

Model	# of nodes	# of arcs	# of trains	Capacity of each aggregated node	Travel time (hr)
IP model	38	35	6	n/a	3.68
Simulation	38	35	6	n/a	3.82
IP model, aggregated section A-B	30	27	10	3 (A-B)	8.38
Simulation	38	35	10	n/a	8.56
IP model, aggregated sections A-B and C-D	33	31	12	3 (A-B), 2(C-D)	LB:10.41 UB:10.84
Simulation	38	35	12	n/a	10.89

Table 4: Comparison of results from IP model with aggregation and simulation model

6 Delay Estimation Techniques

In Sections 4 and 5, we presented the routing and scheduling IP model and used aggregation techniques to solve this model for large railway networks. However, the only property of the individual nodes that was considered while building an aggregated node was the length of the track segments. Track configurations, interactions between trains and knock-on delay effects were not explicitly considered. Hence, the actual travel time of the trains, from their origins to their respective destinations, might be quite different from the values obtained from the IP model. Some key aspects of real-life railway operations that were not accounted for in the aforementioned strategy include the following.

1. Nodes N in G that make up an aggregated node j' can have different capacities $c_{j'}$ and speed-limits $v_{j'}$. Our strategy presented above computes $c_{j'}$ for each j' in N_a in a simplified fashion, without considering the variations in the speed limits of the track segments and the trains.

2. At a given point in time, the nodes constituting an aggregated node j' can have multiple trains traveling on the track segments they represent. In addition, these trains could be traveling in opposite directions.

3. The delay across an aggregated node j' is dependent upon the physical characteristics of the track segments constituting j'. The type of trackage (single-track, double-track

etc), the number of meet-pass points, spacing between the meet-pass points, track speed-limit and the number of sidings need to be considered.

4. The delay across an aggregated node j' is dependent upon the current traffic in j'. That is, the number of trains of various types, their direction of travel and speed-limits.

Due to these reasons, we had to develop a delay estimation technique that accounts for the complexities and non-linearities that exist in real-world railway operations. Instead of the simple aggregation step presented previously, we now develop a technique that accurately estimates the delay for each aggregated node j' as a function of the trackage configuration represented by the nodes in N constituting j', as well as the prevailing traffic in j'. Once we obtain these delay functions, they could be inserted into the IP model in order to route and schedule the trains so as to minimize their travel time. The capacity of each j' can be set such that the time taken by an incoming train to traverse j' is within delay thresholds.

As mentioned in Section 2, there are various ways to obtain delay estimates that have been used in prior research on railway operations; these include queueing theory, stochastic approximation methods, simulation methods, analytical models etc. We decided to use simulation models to obtain data on delay values, and then use regression to develop delay estimates as a function of the train traffic.

The simulation model used for this purpose was developed using AweSim! 2.0 software and presented in Lu et al.(2004). This model considers multiple trackage configurations in the same rail network with multiple speed limits while accounting for the acceleration and deceleration limits of the trains. Freight trains are assumed to arrive at origin stations following a stochastic arrival process. Train movement is a continuous process while the scheduling and dispatching of trains are triggered by discrete events. The continuous motion of train movements is approximated by dividing the movement in small discrete steps. A deadlock-free dispatching algorithm is embedded into the simulation model to determine the optimal run times for a train under multiple speed limits, and to decide the movement of each train in the network considering whether to continue moving at the same speed, to accelerate or decelerate, or to stop. The algorithm also determines the next track to be seized from among the multiple alternative tracks. All the complexities of real-world railway operations such as trackage configurations, track speed-limits, number of meet-pass points, number of sidings and train lengths and speed-limits are built into the system. The authors prove this algorithm to be deadlock-free, while attempting to keep the train delays to a minimum. The modeling methodology does not depend on the size of the network and is insensitive to the trackage configuration. Thus, changes to the trackage configuration require changes only to the input data files. As mentioned in Section 4.1, the route for each train needs to be given in order for the simulation model to schedule them. For a given set of trains and routes, the simulation model has been shown to give efficient schedules with minimum travel delay. We ran this simulation model over each sub-network that we intended to aggregate, in order to get delay estimates across the aggregated node representing the sub-network.

6.1 A Delay Updating Procedure

Given that for different routes we can estimate delay variations using a simulation model, we propose an iterative procedure to update delay as we determine optimal routes. A step-by-step explanation of our methodology to route and schedule trains along the routes with minimum delay is provided below.

1. Given a railway network N and a set of trains H that need to be routed along this network, identify the various routes along which each train could be routed.

2. Identify suitable sub-networks along each route that can be aggregated according to the procedure explained on page 22.

3. Let $h_{j',i}$ denote the number of trains that will be traveling through node j' in iteration i. Assume some initial value for the number of trains traveling on each route in each direction. This gives us the initial ($i = 0$)total number of trains that pass through each aggregated node, $h_{j',0}$, for all j' in N_a.

4. Set counter $i \leftarrow 0$, $\delta_{j'} \leftarrow$ M (some large number)

5. Repeat until $\delta_{j'} \leq \epsilon, \ \forall \, j' \in N_a$, ϵ is a very small number.

a. $i \leftarrow i+1$.

b. Run the simulation model on the sub-network constituting an aggregated node j', with the number of trains in each direction being equal to $h_{j',i}$. Record the time taken to travel through j' by each train. Repeat this step for all aggregated nodes.

c. Develop a regression model for each aggregated node to represent delay in traversing it as a function of the trackage configuration and the existing traffic.

d. The capacity $c_{j'}$ for each j' is set to be equal to the maximum number of trains that can co-exist in j', without the delay for each train exceeding the threshold value.

e. In the network N, substitute the aggregated sub-networks with their respective aggregated nodes. The length $l_{j'}$ of an aggregated node is equal to the total track length of the aggregated sub-network it represents. The speed-limit $v_{j'}$ of an aggregated node is equal to the weighted speed-limit of the sub-network, with the weights being proportional to the track lengths.

f. Incorporate the delay functions [step (c)] and capacities [step(d)] for the aggregated nodes into the IP model from Section 4.0.4.

g. Run the modified IP model and get new estimates $h_{j',i}$ for the number of trains traveling through each j'.

h. $\delta_{j'} = (h_{j',i}$ - $h_{j',i-1})$

6. Input the final routes and schedules into the simulation model to get the final travel times and delay values for all trains. The necessity for this step arises due to the assumptions made in the IP model (Section 4) regarding the train lengths and acceleration and deceleration rates.

To begin with this iterative process, we first identified the parameters to be recorded in each simulation run, which were to be later used to develop the regression-based delay functions. The parameters selected were those which impact the delay experienced by the trains traveling through a sub-network. For this purpose, the parameters considered included the trackage configuration, traffic and operating parameters. As explained in Krueger (1999), by focusing on how changes in these parameters affect delay, we account for the dynamic nature of delay and capacity. Hence, the IP model would be able to efficiently route the trains along the routes with minimum delay. Since we ran the simulation model for each aggregated sub-network individually, the trackage configuration remained unchanged and, hence, was not considered as one of the factors for regression. In the simulation model, the number of trains scheduled is equal to $h_{j',i}$. The ends of the sub-network under consideration are treated as the origins and destinations of the trains passing through it. Trains arrive at their respective origins according to a Poisson process. The simulation was started and made to run over a 100 days time horizon, with statistics being cleared after 10 days, that is, after steady state has been reached. The traffic and operating parameters of the system were captured at the time of arrival of a randomly selected train at either ends of the sub-network. For a fixed $h_{j',i}$, the arrival rate of the trains were varied and the simulation run to get a good representation of the system delay at varied levels of traffic. The traffic and operating parameters recorded were the following.

1. X_i: the number of trains of type i in the subnetwork.

2. D_1: the number of trains that enter the subnetwork from the opposite direction after the entry and before the exit of the aforementioned randomly selected train.

3. D_2: the number of trains already present in the subnetwork when the randomly selected train enters the subnetwork and traveling in a direction opposite to it.

Once the simulation-generated data has been gathered, it is fed into a statistical software such as Minitab to generate a regression equation for delay as a function of the traffic and operating parameters. For each aggregated node, we generated regression equations with these parameters, and also with the interaction effects between the traffic parameters. Finally, the regression equation with a reasonable value of adjusted R-squared and number of significant parameters is selected to be incorporated into the routing and scheduling IP model.

6.2 Generating the delay function: An example

In this section, we illustrate the procedure of developing the delay equation for a sub-network. The network being considered here is a Union Pacific - Alhambra railway track between the CP Dayton Taylor Yard and the City of Industry, shown below in Figure 4. It is a double-track segment with crossings, and is 5.79 miles in length.

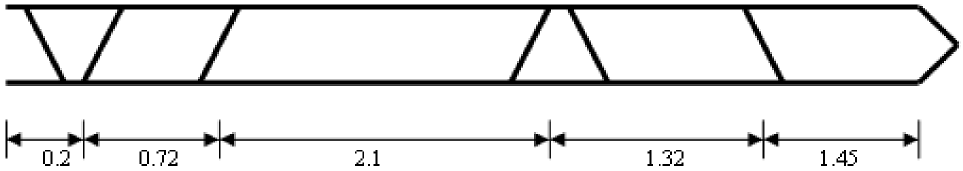
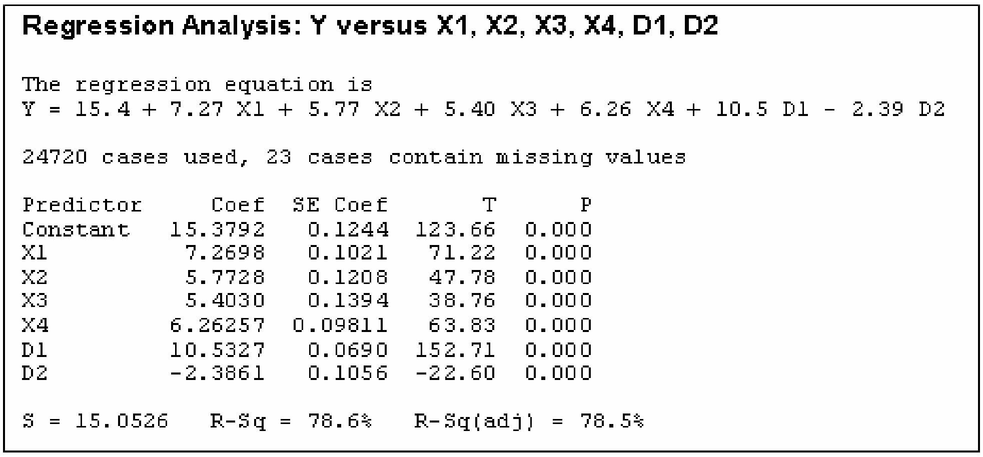


Figure 4: Illustration of a sub-network to be aggregated

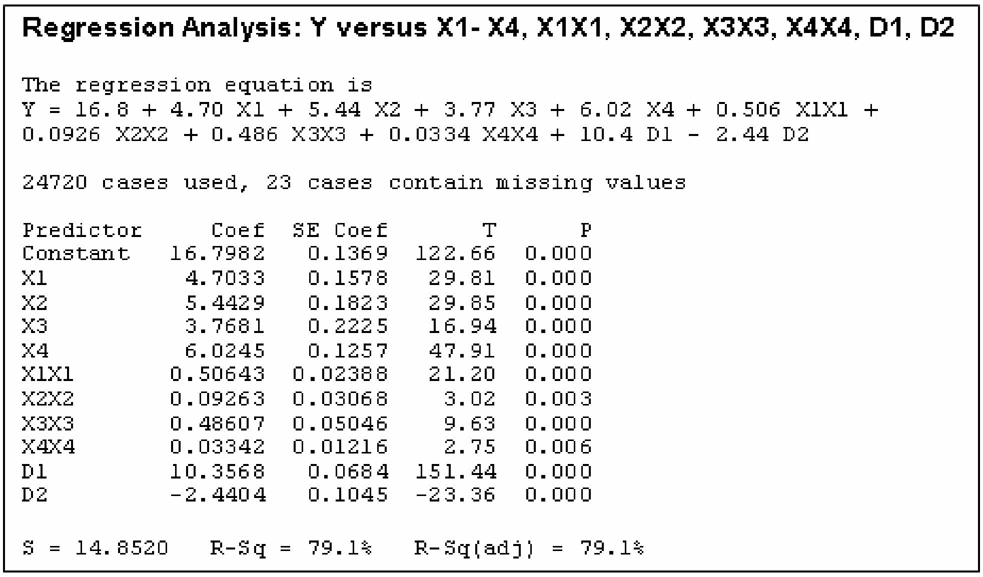
As explained above, the simulation model was run for this sub-network over 100 days. We considered four types of trains for this sub-network - long double stack (8000 feet), other intermodal (6000 feet), carload (6500 feet) and oil (5000 feet). The traffic and operating parameters were recorded when randomly selected trains enter the sub-network. The arrival rates of the trains were altered in order to get a good representation of the impact of the traffic and operating parameters on delay. Over 25,000 data points were collected in this manner. The following regression models were developed.

1. X_i's, D_1 and D_2.
2. X_i's, quadratic interaction effects of X_i's, D_1 and D_2.
3. X_i's, quadratic and cross-product interaction effects of X_i's, D_1 and D_2.
4. X_i's, quadratic, cross-product and cubic interaction effects of X_i's, D_1 and D_2.



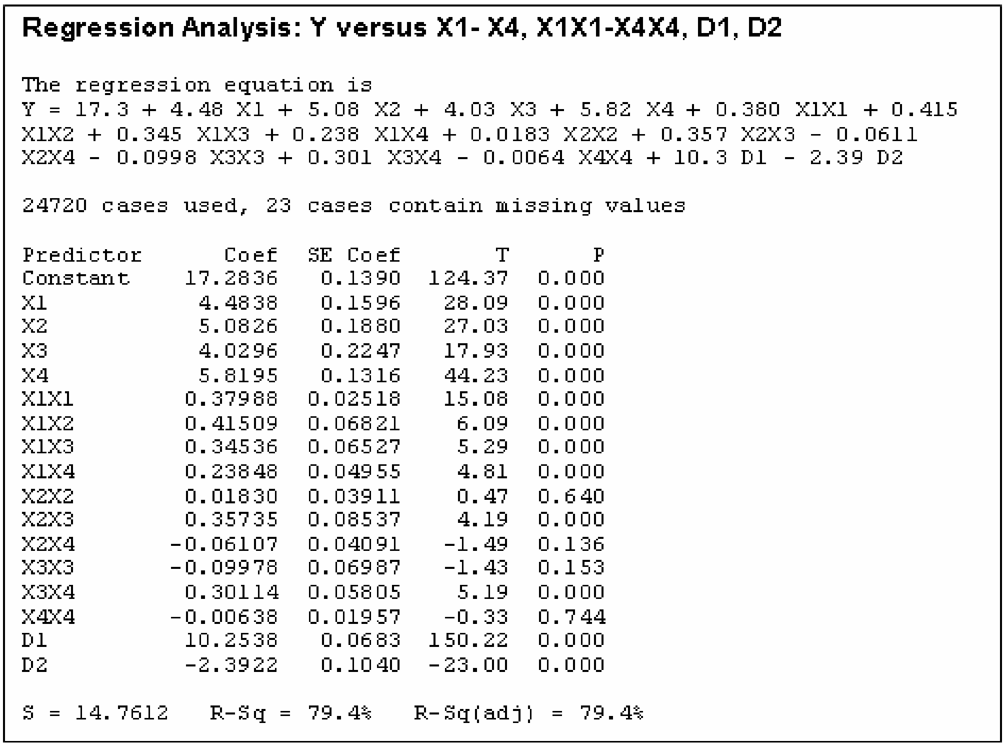
```
Regression Analysis: Y versus X1, X2, X3, X4, D1, D2

The regression equation is
Y = 15.4 + 7.27 X1 + 5.77 X2 + 5.40 X3 + 6.26 X4 + 10.5 D1 - 2.39 D2

24720 cases used, 23 cases contain missing values

Predictor      Coef  SE Coef       T      P
Constant    15.3792   0.1244  123.66  0.000
X1           7.2698   0.1021   71.22  0.000
X2           5.7728   0.1208   47.78  0.000
X3           5.4030   0.1394   38.76  0.000
X4          6.26257  0.09811   63.83  0.000
D1          10.5327   0.0690  152.71  0.000
D2          -2.3861   0.1056  -22.60  0.000

S = 15.0526   R-Sq = 78.6%   R-Sq(adj) = 78.5%
```

Figure 5: Regression equation for delay



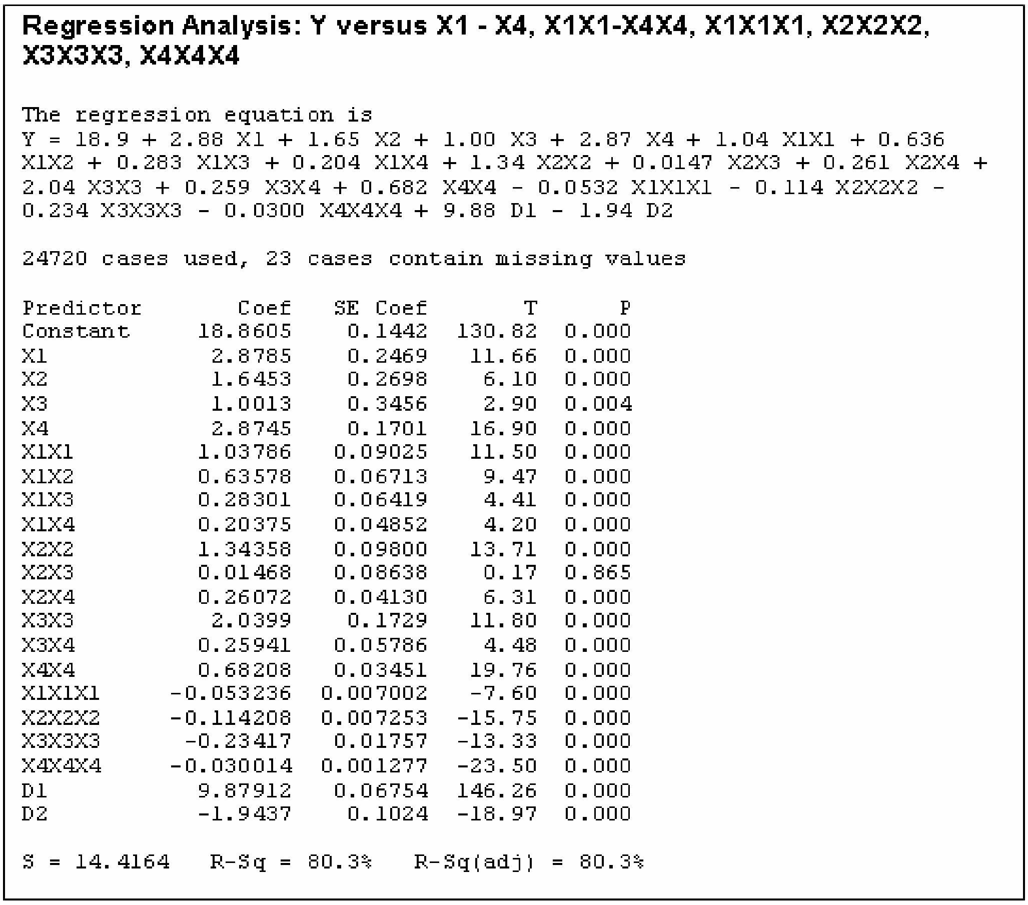
```
Regression Analysis: Y versus X1- X4, X1X1, X2X2, X3X3, X4X4, D1, D2

The regression equation is
Y = 16.8 + 4.70 X1 + 5.44 X2 + 3.77 X3 + 6.02 X4 + 0.506 X1X1 +
0.0926 X2X2 + 0.486 X3X3 + 0.0334 X4X4 + 10.4 D1 - 2.44 D2

24720 cases used, 23 cases contain missing values

Predictor      Coef  SE Coef       T      P
Constant    16.7982   0.1369  122.66  0.000
X1           4.7033   0.1578   29.81  0.000
X2           5.4429   0.1823   29.85  0.000
X3           3.7681   0.2225   16.94  0.000
X4           6.0245   0.1257   47.91  0.000
X1X1        0.50643  0.02388   21.20  0.000
X2X2        0.09263  0.03068    3.02  0.003
X3X3        0.48607  0.05046    9.63  0.000
X4X4        0.03342  0.01216    2.75  0.006
D1          10.3568   0.0684  151.44  0.000
D2          -2.4404   0.1045  -23.36  0.000

S = 14.8520   R-Sq = 79.1%   R-Sq(adj) = 79.1%
```

Figure 6: Regression equation for delay with quadratic terms



```
Regression Analysis: Y versus X1- X4, X1X1-X4X4, D1, D2

The regression equation is
Y = 17.3 + 4.48 X1 + 5.08 X2 + 4.03 X3 + 5.82 X4 + 0.380 X1X1 + 0.415
X1X2 + 0.345 X1X3 + 0.238 X1X4 + 0.0183 X2X2 + 0.357 X2X3 - 0.0611
X2X4 - 0.0998 X3X3 + 0.301 X3X4 - 0.0064 X4X4 + 10.3 D1 - 2.39 D2

24720 cases used, 23 cases contain missing values

Predictor        Coef   SE Coef        T      P
Constant      17.2836    0.1390   124.37  0.000
X1             4.4838    0.1596    28.09  0.000
X2             5.0826    0.1880    27.03  0.000
X3             4.0296    0.2247    17.93  0.000
X4             5.8195    0.1316    44.23  0.000
X1X1          0.37988   0.02518    15.08  0.000
X1X2          0.41509   0.06821     6.09  0.000
X1X3          0.34536   0.06527     5.29  0.000
X1X4          0.23848   0.04955     4.81  0.000
X2X2          0.01830   0.03911     0.47  0.640
X2X3          0.35735   0.08537     4.19  0.000
X2X4         -0.06107   0.04091    -1.49  0.136
X3X3         -0.09978   0.06987    -1.43  0.153
X3X4          0.30114   0.05805     5.19  0.000
X4X4         -0.00638   0.01957    -0.33  0.744
D1            10.2538    0.0683   150.22  0.000
D2            -2.3922    0.1040   -23.00  0.000

S = 14.7612   R-Sq = 79.4%   R-Sq(adj) = 79.4%
```

Figure 7: Regression equation for delay with quadratic and cross-product terms



Regression Analysis: Y versus X1 - X4, X1X1-X4X4, X1X1X1, X2X2X2, X3X3X3, X4X4X4

```
The regression equation is
Y = 18.9 + 2.88 X1 + 1.65 X2 + 1.00 X3 + 2.87 X4 + 1.04 X1X1 + 0.636
X1X2 + 0.283 X1X3 + 0.204 X1X4 + 1.34 X2X2 + 0.0147 X2X3 + 0.261 X2X4 +
2.04 X3X3 + 0.259 X3X4 + 0.682 X4X4 - 0.0532 X1X1X1 - 0.114 X2X2X2 -
0.234 X3X3X3 - 0.0300 X4X4X4 + 9.88 D1 - 1.94 D2

24720 cases used, 23 cases contain missing values

Predictor        Coef    SE Coef        T      P
Constant      18.8605     0.1442   130.82  0.000
X1             2.8785     0.2469    11.66  0.000
X2             1.6453     0.2698     6.10  0.000
X3             1.0013     0.3456     2.90  0.004
X4             2.8745     0.1701    16.90  0.000
X1X1          1.03786    0.09025    11.50  0.000
X1X2          0.63578    0.06713     9.47  0.000
X1X3          0.28301    0.06419     4.41  0.000
X1X4          0.20375    0.04852     4.20  0.000
X2X2          1.34358    0.09800    13.71  0.000
X2X3          0.01468    0.08638     0.17  0.865
X2X4          0.26072    0.04130     6.31  0.000
X3X3           2.0399     0.1729    11.80  0.000
X3X4          0.25941    0.05786     4.48  0.000
X4X4          0.68208    0.03451    19.76  0.000
X1X1X1      -0.053236   0.007002    -7.60  0.000
X2X2X2      -0.114208   0.007253   -15.75  0.000
X3X3X3       -0.23417    0.01757   -13.33  0.000
X4X4X4      -0.030014   0.001277   -23.50  0.000
D1            9.87912    0.06754   146.26  0.000
D2            -1.9437     0.1024   -18.97  0.000

S = 14.4164   R-Sq = 80.3%   R-Sq(adj) = 80.3%
```

Figure 8: Regression equation for delay with quadratic, cross-product and cubic terms

We assume a 5% significance level. In Figure 5, all the traffic and operating parameters are significant, and the regression equation has an adjusted R-squared value of 78.5%. As can be seen from Figures 6, 7 and 8, there is very little improvement in the adjusted R-squared value as we add the interaction terms. Since these traffic and operating parameters need to be mathematically expressed in order to be included into the IP model, and since we preferred to keep the IP model free of any non-linear expressions, we decided to go with the regression equation for delay with just the first-order traffic and operating parameters. By doing so, we did not compromise on the quality of the delay function because of an insignificant variation in the adjusted R-squared value.

6.3 Delay function in IP Model

Once the delay equations have been derived, they need to be inserted into the routing and scheduling IP model to get the new number of trains routed over each route (and aggregated node). In the IP formulation presented in Section 4.0.4, constraint (4) enforces the time a train should take to traverse a node by considering the speed limit over the node's segment. However, in reality, the maximum speed a train can travel is the minimum of its speed limit and the segment's speed limit, denoted by v_min_j. So, we modify constraint (4) accordingly.

$$\sum_{i|(i,j,h)\in S} Y_{i,j,h,\tau} - \sum_{k|(j,k,h)\in S} Y_{j,k,h,\tau+\lceil l_j/v\_min_j\rceil} \geq 0 \qquad \forall\left(j \in N | j \neq o_h, j \neq d_h\right), \quad h \in H$$

$$\forall \tau \in 0, \ldots, T - \lceil l_j/v\_min_j\rceil \quad \text{(4a)}$$

In the case of an aggregated node, the speed limit of the sub-network is the weighted speed-limit of all the track segments making up the sub-network, as explained previously.

The regression-based delay equation in Figure 5 above is represented by the following equations (11) and (12) . We separate the delay expression for the two directions, denoted by $\Delta 1_{j,\tau}$ and $\Delta 2_{j,\tau}$. Note that $\Delta 1_{j,\tau}$ and $\Delta 2_{j,\tau}$ are indexed in τ because delay is being expressed as a function of the traffic and the operating parameters, both of which vary with time.

$$\begin{aligned}\Delta 1_{j,\tau} = {} & \kappa_j + x1_j * \left[\sum_{i|i,j,h \in S_{LDT}} Y_{i,j,h,\tau} - \sum_{k|k,j,h \in S_{LDT}} Y_{j,k,h,\tau} \right] \\ & + x2_j * \left[\sum_{i|i,j,h \in S_{OIM}} Y_{i,j,h,\tau} - \sum_{k|k,j,h \in S_{OIM}} Y_{j,k,h,\tau} \right] + x3_j * \left[\sum_{i|i,j,h \in S_{OIL}} Y_{i,j,h,\tau} - \sum_{k|k,j,h \in S_{OIL}} Y_{j,k,h,\tau} \right] \\ & + x4_j * \left[\sum_{i|i,j,h \in S_{CARL}} Y_{i,j,h,\tau} - \sum_{k|k,j,h \in S_{CARL}} Y_{j,k,h,\tau} \right] + d1 * \left[\sum_{l|l,k,h \in S_2} Y_{l,k,h,\tau} - \sum_{j|k,j,h \in S_2} Y_{k,j,h,\tau} \right] \\ & + d2 * \left[\sum_{k|k,j,h \in S_2} Y_{k,j,h,\tau} - \sum_{i|j,i,h \in S_2} Y_{k,j,h,\tau-1} \right] \quad \forall \quad j \in N_a, \tau \in 1, \ldots, T \end{aligned} \tag{11}$$

$$\begin{aligned}\Delta 2_{j,\tau} = {} & \kappa_j + x1_j * \left[\sum_{i|i,j,h \in S_{LDT}} Y_{i,j,h,\tau} - \sum_{k|k,j,h \in S_{LDT}} Y_{j,k,h,\tau} \right] \\ & + x2_j * \left[\sum_{i|i,j,h \in S_{OIM}} Y_{i,j,h,\tau} - \sum_{k|k,j,h \in S_{OIM}} Y_{j,k,h,\tau} \right] + x3_j * \left[\sum_{i|i,j,h \in S_{OIL}} Y_{i,j,h,\tau} - \sum_{k|k,j,h \in S_{OIL}} Y_{j,k,h,\tau} \right] \\ & + x4_j * \left[\sum_{i|i,j,h \in S_{CARL}} Y_{i,j,h,\tau} - \sum_{k|k,j,h \in S_{CARL}} Y_{j,k,h,\tau} \right] + d1 * \left[\sum_{l|l,k,h \in S_1} Y_{l,k,h,\tau} - \sum_{j|k,j,h \in S_1} Y_{k,j,h,\tau} \right] \\ & + d2 * \left[\sum_{k|k,j,h \in S_1} Y_{k,j,h,\tau} - \sum_{i|j,i,h \in S_1} Y_{k,j,h,\tau-1} \right] \quad \forall \quad j \in N_a, \tau \in 1, \ldots, T \end{aligned} \tag{12}$$

κ_j, $x1_j$, $x2_j$, $x3_j$, $x4_j$, $d1$ and $d2$ are the constant and the coefficients of X_1, X_2, X_3, X_4, D_1 and D_2 respectively from the regression-based delay function. S_{LDT}, S_{OIM}, S_{OIL} and S_{CARL} are subsets of S for the long double stack, other intermodal, oil and carload types of train respectively. For the IP model, it is not easy to mathematically express D_1, the number of trains that can potentially enter the sub-network in the opposite direction between the entry and exit times of the train under consideration, since this is something that occurs in the future. So, we make an approximation by expressing it as the number of trains traveling in the opposite direction that are in the adjacent nodes k, $\forall (j,k) \in A$, at the time of entry of a train h into an aggregated node j from, say, i. For a train entering an aggregated node J at, say, τ, D_2 is expressed mathematically as the number of trains traveling in the opposite direction that have entered j by τ, but have not yet departed from j. Now, in addition to constraint (4a), we need to introduce two more delay constraints for the aggregated nodes, accounting for $\Delta 1_{j,\tau}$ and $\Delta 2_{j,\tau}$. For this purpose, we introduce the following equations.

According to constraint (13), $Z_{i,j,h}$ equals the time train h enters j from i. Evidentally, it is equal to zero for all other values of i, j and τ for a given train.

$$Z_{i,j,h} = \sum_{\tau=1}^{T} \tau * [Y_{i,j,h,\tau} - Y_{i,j,h,\tau-1}] \quad \forall \quad j \in N_a, (i,j,h) \in S|(i,j) \in A, h \in H \tag{13}$$

Constraints (14) and (15) enforce the minimum time a train entering an aggregated node j takes to travel through it, as per the regression-based delay equations.

$$\sum_{k|(j,k)\in A_1} Z_{j,k,h} - \sum_{i|(i,j)\in A_1} Z_{i,j,h} + \left[1 - \sum_{i|(i,j)\in A_1} (Y_{i,j,h,\tau} - Y_{i,j,h,\tau-1})\right] * M \geq \Delta 1_{j,\tau}$$
$$\forall \quad h \in H, j \in N_a, \tau \in 1,\ldots,T \tag{14}$$

$$\sum_{k|(j,k)\in A_2} Z_{j,k,h} - \sum_{i|(i,j)\in A_2} Z_{i,j,h} + \left[1 - \sum_{i|(i,j)\in A_2} (Y_{i,j,h,\tau} - Y_{i,j,h,\tau-1})\right] * M \geq \Delta 2_{j,\tau}$$
$$\forall \quad h \in H, j \in N_a, \tau \in 1,\ldots,T \tag{15}$$

The most limiting of constraints (4a), (14) and (15) determines the actual delay experienced by a train in traveling through an aggregated node.

6.4 A Generic Delay Estimation Procedure

In this section, we extend the regression-based delay estimation procedure, presented above, to generic rail networks. A methodology, based on *Design of Experiments* techniques, is used to predict delays in railway networks, while capturing interactions between the network, traffic and operating parameters. As explained in detail below, generic networks are first constructed to represent the range of the physical attributes of the actual networks for which delays are to be estimated. Next, we run simulations representing train movements through these generic networks, and record the relevant system state data. Finally, a regression analysis is made to run on the collected data. This regression equation is shown to accurately estimate the travel time delay on an actual network that has its physical attributes within the extreme limits of the networks used in the experiments.

Once again, we used the simulation model discussed in Section 5 to run simulations on the generic networks. Considering the Downtown Los Angeles - Inland Empire Trade Corridor as an example, the authors show that the delays experienced by the trains as per the simulation model are very close to the real-world travel time delays. This simulation model is used to study the impact of the network topology and traffic parameters on the delay experienced by trains in traversing a subnetwork. We assume trains can accelerate and decelerate instantaneously to obey track speed limits, and the simulation model is modified accordingly. Hence, the maximum speed of a train at each instant of time is simply set to the constraining speed-limit of the track segment. Furthermore, a Poisson arrival process is assumed for each train. The control parameters for each simulation are as follows:

1. λ_i: the arrival rate of each type of train.

2. L: the length of the subnetwork, that is, the distance in miles between the start and the end of the subnetwork.

3. V: the speed-limit of the subnetwork. The free running time of the train over the subnetwork is inversely proportional to the minimum of V and the train speed-limit.

4. C: the number of crossings or sidings for a double- or single-track respectively. They are assumed to be uniformly distributed. These enable a smooth flow of traffic within the subnetwork. Typically, delay reduces with an increase in C.

5. S: the spacing over a subnetwork. This is defined as the portion of the subnetwork over which crossings (or sidings) are uniformly distributed. If crossings are uniformly distributed over the entire track length (i.e., $S = 1$), then trains can more easily overtake and/or cross each other, than if all the crossings are concentrated at one end of the network segments. Therefore, delay increases with a decrease in S due to possible interactions between trains on two consecutive crossings (or sidings).

Among the above control parameters, L, V, C and S are utilized to represent various subnetwork configurations in order to study the impact of these four on the travel time delay. In our work, to be able to build a generic delay estimation model for a single-track or a double-track, we assume that each of these four parameters can take three different values which are labelled as LO, MID and HI. These three levels can be thought of as representing the lowest, middle and highest subnetwork length, speed-limits, crossings (or sidings) and spacing that can be found in the complex railway network under consideration. There are 3^4, or 81 subnetwork configurations that need to be simulated in order to build the delay model. However, due to the need for efficiency, we invoke a response surface methodology tool known as *fractional factorial design*. We develop a one-third fractional factorial design, wherein we assume third-order and higher interactions between the four control parameters to be negligible, and instead concentrate our efforts in studying the main effects and the two-factor interactions. According to standard rules (Montgomery, 1984), we choose 27 of these 81 designs so as to get a good representation of the interaction effects. In response surface terminology, this is called a 3^{4-1} design. In this way, the generic delay model developed would be able to estimate delay, with high precision, on a host of subnetworks within the extreme values of the four topological parameters. An important thing to note here is that we do not mix double-track subnetworks with single-track subnetworks. The generic delay model is developed separately for each of them.

The simulation is run for each of the 27 subnetwork configurations using AweSim! 2.0 (Pritsker and O'Reilly, 1999), by altering the data files. For each subnetwork configuration, simulations are run for a fixed ratio of the types of trains traveling through the subnetwork, and various values of λ_i for each train type. Two stations are assumed to be present at either end of a subnetwork, and there are an equal number of trains travelling in either direction. Trains are made to travel between their respective origin and destination stations. Furthermore, we also assume that there are no stations in between the origin and destination stations. During each run, the state of the system is recorded at the arrival times of randomly selected trains at their respective origin station. Similar to the the recorded data in Section 5, at each randomly sampled time instant we once again record X_i, D_1 and D_2.

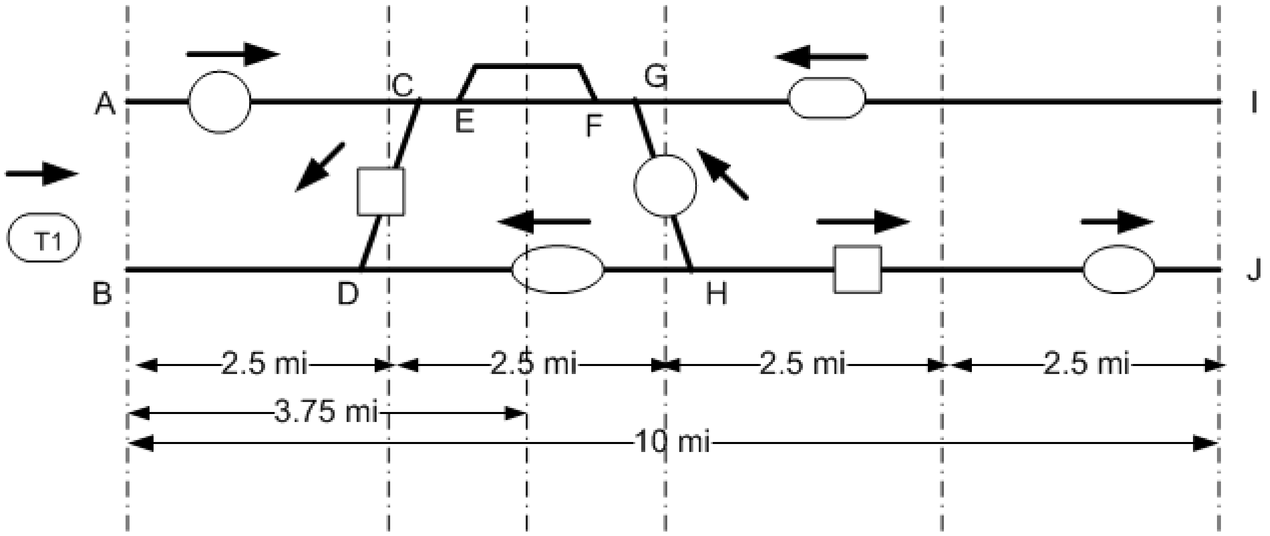


Figure 9: The double-track railway segment with 7 moving trains at the time train T_1 is deciding to enter on A or B

Figure 9 above, shows a double-track subnetwork of length $L = 10$ mi with crossings CD and GH and siding EF, i.e., $C = 3$. The speed limit V over this subnetwork is 35 mi/hr. The crossings and the sidings are uniformly distributed over $\frac{3}{4}$th or 75% of the length of the subnetwork, i.e., $S = 0.75$. $T1$ is a train that is about to enter the network. There are four train types, represented by rectangles, squares, circles and ellipses. Of the 7 trains already existing in the subnetwork, 3 are traveling in the same direction as $T1$ would upon entering, and 4 are traveling in the opposite direction. Hence, $D2 = 4$. At time of entry of $T1$, X_1 (squares) = 2, X_2 (rectangles) = 1, X_3 (circles) = 2 and X_4 (ellipses) = 2. D_1 represents the trains that would enter through IJ from the adjacent subnetwork(s) after $T1$ enters through AB and before it exits through IJ. X_i, D_1 and D_2 are called *covariate parameters* because they can be altered only by changing the control parameter(s), in this case, λ_i. D_1 and D_2 represent traffic moving in the opposite direction, relative to the randomly selected train, and therefore impact delay by providing "resistance" to its smooth flow.

Then, we run a single regression analysis over the data collected from the 27 subnetwork configurations, using *Minitab*. The parameters used are the X_i's, D_1, D_2, L, V, C and S, and the response variable is the travel time delay experienced by the train, Y. A normal probability plot and a plot of the residuals $(y_i - \hat{y}_i)$ versus the predicted response $\hat{Y}$ (fitted response value from the regression analysis) are plotted. This is done to examine the fitted model to ensure that it provides an adequate approximation to the true system, and to verify that none of the least squares regression assumptions are violated. We also run regressions with quadratic and cross-product interaction effects of the X_i's and the network topological parameters, in order to study their effects on the delay.

In the next section, we present an example wherein we build these generic delay models for single and double-track subnetworks for the railway network in the Los Angeles area. On a side note, we use the following nomenclature in the remaining sections of this paper: *actual delay* refers to the delay experienced by the trains in real-world rail operations, *simulation delay* refers to the travel time delay experienced by the trains as per the simulation model by Lu et al. (2004), and *predicted delay* refers to the delay estimated from the delay estimation equation (obtained from the regression analysis), that is expected to be experienced by the trains in traveling through a network.

6.5 Case Study: Los Angeles area Railway Network

The Ports of Los Angeles and Long Beach are the busiest ports on the West Coast. Three railroad lines, Union Pacific - Alhambra, Union Pacific - San Gabriel and Burlington Northern Santa Fe operate service from Los Angeles downtown to the ports. Travel time delays from the simulation model on this network have been shown by Lu et al. (2004) to be close to real-world delay values. The trackage in this region is primarily a combination of single and double-tracks. Crossings and sidings are provided for the purpose of train meets and overtakes, thereby ensuring a smooth traffic flow. Four types of trains primarily travel on these tracks - long double stack (8000 feet), intermodal (6000 feet), carload (6500 feet) and oil (5000 feet). The speed-limits of these trains are 70, 55, 50 and 40 mi/hr respectively. In our experiments, we assume a fixed ratio of these four train types. For each subnetwork, multiple simulation runs are performed, each with a different combination of the λ_i's. The primary purpose of this is to get a good representation of the system space, that is, how the delay varies with different values of X_i, for a fixed setting of the network topology parameters.

6.5.1 Delay Estimation for a Double-track Subnetwork

For the purpose of designing networks to build a generic delay model for a double-track subnetwork, the three grades of values listed in Table 5 below were selected for the network topology parameters.

	LO	**MED**	**HI**
Length (mi)	5.0	12.5	20.0
Speed-limit (mi/hr)	15	35	55
Crossings	1	3	5
Spacing (%)	0.50	0.75	1.00

Table 5: Settings for network topology parameters for double-track simulations

These values reflect the range of the four parameters within which a majority of the double-track subnetworks in the Los Angeles area network lie. As described in the previous section, we now develop a one-third fractional factorial design on which the simulations are to be run. The 27 treatment combinations that are used to run the simulations are shown in Table 6.

	L	**V**	**C**	**S**		**L**	**V**	**C**	**S**
1	12.5	55	1	0.50	15	12.5	55	5	0.75
2	20.0	35	3	1.00	16	12.5	15	1	1.00
3	5.0	35	1	1.00	17	12.5	15	3	0.75
4	5.0	15	1	0.50	18	20.0	55	3	0.75
5	5.0	35	3	0.75	19	5.0	15	3	1.00
6	20.0	35	1	0.5	20	5.0	35	5	0.5
7	20.0	15	5	1.00	21	20.0	35	5	0.75
8	5.0	15	5	0.75	22	20.0	15	1	0.75
9	12.5	35	3	0.5	23	20.0	55	5	0.50
10	20.0	55	1	1.00	24	12.5	35	1	0.75
11	20.0	15	3	0.5	25	5.0	55	3	0.50
12	12.5	15	5	0.5	26	12.5	35	5	1.00
13	5.0	55	5	1.00	27	12.5	55	3	1.00
14	5.0	55	1	0.75					

Table 6: 27 parameter combinations considered in the one-third fractional factorial design for double-track simulations

For each of the 27 treatment combinations, 25 simulation runs are made, each with a different combination of the λ_i values for the four train types. This is done to obtain travel time estimates under various network operating conditions, described by the X and D variables. The simulations are run at the real-world daily peak traffic conditions, thus representing a stationary process. Therefore, the variance of the observed values is constant. As explained previously, in each simulation run, the state of the system is recorded at random intervals of time, each triggered by the arrival of a randomly selected train at its respective origin station. In each simulation run, approximately 1000 data points are recorded in this manner. Finally, all the data collected from these 27x25 simulations are combined to fit a regression model. The results are plotted in the graphs below.

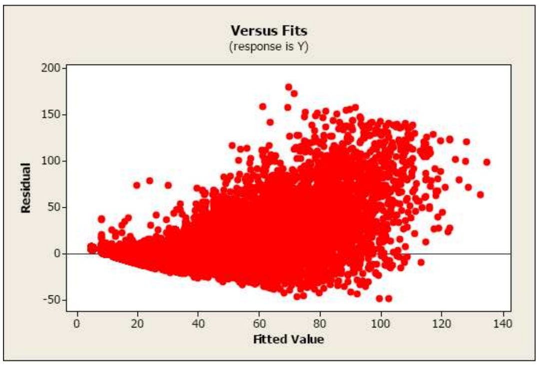


Figure 10: Residuals vs. predicted response for the double-track subnetwork simulation

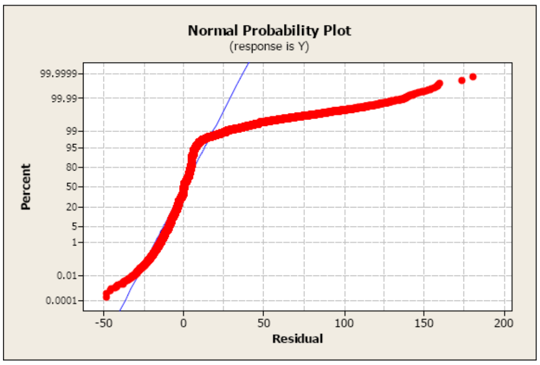


Figure 11: Normal probability plot for the double-track subnetwork simulation

In the plot of the residuals versus the predicted response $\hat{Y}$, the general impression should be that the residual scatter randomly on the display, suggesting that the variance of the predicted response is constant for all values of the mean of $\hat{Y}$. However, in Figure 10, our plot exhibits a funnel-shaped pattern, which indicates that the variance of the predicted response depends on its mean value. In Figure 11, it is apparent that the normality assumption is being violated. A remedial procedure for these abnormalities is to transform the response variable Y. A Box-Cox transformation procedure is carried out, and the transformation parameter that minimizes the sum of squares of error is selected. For the experiment presented above, a natural log transformation has the best effect in improving the fit of the model to the data.

In addition to the single-order effects, we also fit regression models to a data set that includes interactions in the X_i's and the network parameters. In Table 7, we retrace the backward elimination procedure. As per this procedure, we start with the single and higher-order effects of the X's, and the single-order and higher-order interaction effects of the network topology parameters. The higher-order effects of the four network topology parameters L, V, C and

S comprise of the quadratic effects represented by LL, VV, CC and SS, and the interaction effects represented by LV, LC, LS, VC, VS and CS. After running regression analysis, we delete those effects that have no or least impact on the adjusted R-sq value. Next, we run regression analysis with just the effects that were not deleted in the previous step. This is done iteratively. We stop when we are left with just the statistically significant terms in the regression equation.

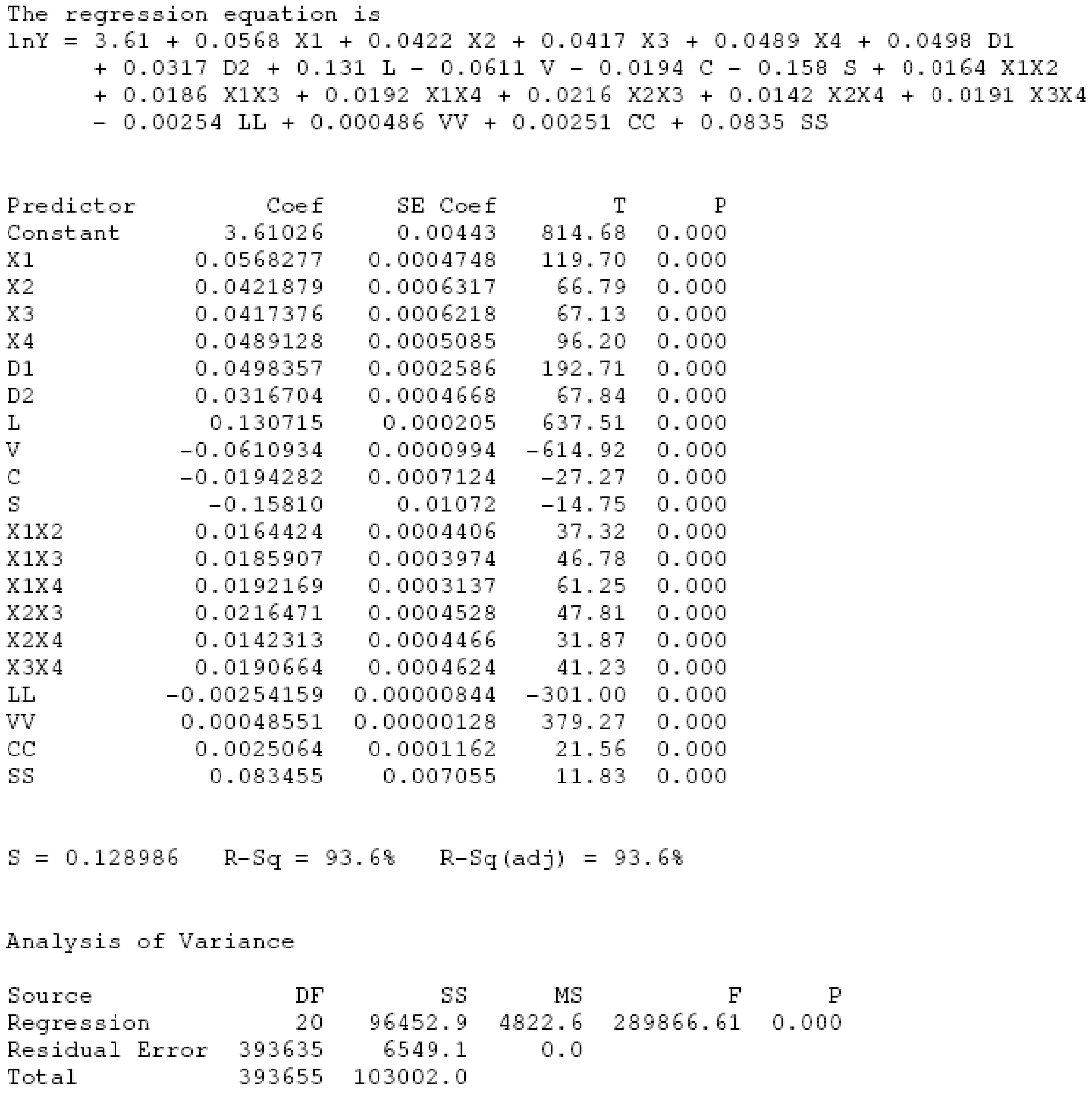
The size of the data sets obtained from the simulation runs creates a problem when assessing statistical significance. Specifically, the large degree of freedom for error makes all of the candidate regression terms significant at typical alpha levels. So, instead of using P-values as criteria for model selection, we use the relative magnitudes of the coefficients and the adjusted R-squared value. In effect, we are eliminating terms that provide negligible contributions to the predicted values. Since the P-values of all the terms are always significant, we eliminate the term with the smallest coefficient value. An important observation to be made from the table above is that the network topology variables in the fractional factorial design have been chosen so as to keep them linearly independent of each other. By the virtue of this design, if any of the network topology interaction terms has a low coefficient value and is chosen to be eliminated, then all the topology interaction terms that have been so chosen can be simultaneously eliminated from the regression equation.

	Regression Parameters	Adjusted R-Sq (%)	Eliminated Term(s)
1	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X1,X1X2,X1X3,X1X4,X2X2, X2X3,X2X4,X3X3,X3X4,X4X4, LL,LV,LC,LS,VV,VC,VS,CC, CS,SS	93.7	–
2	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X1,X1X2,X1X3,X1X4,X2X2, X2X3,X2X4,X3X3,X3X4,X4X4, LL,VV,CC,SS	93.6	LV,LC,LS,VC,VS,CS
3	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X1,X1X2,X1X3,X1X4,X2X2, X2X3,X2X4,X3X4,X4X4, LL,VV,CC,SS	93.6	X3X3
4	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X2,X1X3,X1X4,X2X2,X2X3, X2X4,X3X4,X4X4,LL,VV,CC,SS	93.6	X1X1
5	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X2,X1X3,X1X4,X2X3,X2X4, X3X4,X4X4,LL,VV,CC,SS	93.6	X2X2
6	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X2,X1X3,X1X4,X2X3,X2X4, X3X4,LL,VV,CC,SS	93.6	X4X4
7	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X2,X1X3,X1X4,X2X3,X2X4,X3X4	89.9	LL,VV,CC,SS
8	X1,X2,X3,X4,D1,D2,L,V,C,S	87.9	X1X2,X1X3,X1X4, X2X3,X2X4,X3X4

Table 7: Backward elimination for double-track subnetwork simulation.

By comparing the regression models shown in Table 7, we notice that the one in row 6 has the least number of significant terms without a drastic reduction in the adjusted R-squared value. This regression model is shown in Figure 12. In a physical sense, this regression model suggests that the delay experienced by a train is impacted by the heterogenic mix of traffic flowing in the opposite direction. Furthermore, in addition to their single-order effects, quadratic interations of the network parameters influence delay. In this manner, we derive a regression model that defines an exponential relation between delay and traffic, operating and network parameters.

The regression equation below can be used to estimate delay (Y) for an actual double-track subnetwork.



```
The regression equation is
lnY = 3.61 + 0.0568 X1 + 0.0422 X2 + 0.0417 X3 + 0.0489 X4 + 0.0498 D1
      + 0.0317 D2 + 0.131 L - 0.0611 V - 0.0194 C - 0.158 S + 0.0164 X1X2
      + 0.0186 X1X3 + 0.0192 X1X4 + 0.0216 X2X3 + 0.0142 X2X4 + 0.0191 X3X4
      - 0.00254 LL + 0.000486 VV + 0.00251 CC + 0.0835 SS


Predictor         Coef     SE Coef       T      P
Constant       3.61026     0.00443  814.68  0.000
X1           0.0568277   0.0004748  119.70  0.000
X2           0.0421879   0.0006317   66.79  0.000
X3           0.0417376   0.0006218   67.13  0.000
X4           0.0489128   0.0005085   96.20  0.000
D1           0.0498357   0.0002586  192.71  0.000
D2           0.0316704   0.0004668   67.84  0.000
L             0.130715    0.000205  637.51  0.000
V           -0.0610934   0.0000994 -614.92  0.000
C           -0.0194282   0.0007124  -27.27  0.000
S             -0.15810     0.01072  -14.75  0.000
X1X2         0.0164424   0.0004406   37.32  0.000
X1X3         0.0185907   0.0003974   46.78  0.000
X1X4         0.0192169   0.0003137   61.25  0.000
X2X3         0.0216471   0.0004528   47.81  0.000
X2X4         0.0142313   0.0004466   31.87  0.000
X3X4         0.0190664   0.0004624   41.23  0.000
LL         -0.00254159  0.00000844 -301.00  0.000
VV          0.00048551  0.00000128  379.27  0.000
CC           0.0025064   0.0001162   21.56  0.000
SS            0.083455    0.007055   11.83  0.000


S = 0.128986   R-Sq = 93.6%   R-Sq(adj) = 93.6%


Analysis of Variance

Source              DF        SS      MS          F      P
Regression          20   96452.9  4822.6  289866.61  0.000
Residual Error  393635    6549.1     0.0
Total           393655  103002.0
```

Figure 12: Detailed regression results for model 6 in the double-track simulation

The next logical step is to test our delay modeling methodology. As part of this step, we adopt two validation strategies. First, we randomly chose five treatment combinations of the 54 that were not used in the one-third fractional factorial design. The performance of the generic delay estimation model for these five network configurations is shown in rows 1-5 in Table 8 below. In the second validation strategy, we choose a subnetwork existing in the Los Angeles area, and test the performance of our delay estimation model on this subnetwork.

This result is shown in row 6 in Table 8 below.

	L	V	C	S	Relative Error, Mean (%)	Relative Error, Median (%)	Percent within 20% rel. error
1	5.0	15	1	1.00	14.55	9.09	87.82
2	12.5	35	3	0.75	9.06	5.75	88.65
3	20.0	35	5	1.00	11.22	5.75	81.77
4	5.0	55	3	0.75	4.92	0.78	93.40
5	12.5	55	3	0.50	9.19	6.78	88.03
6	6	36.67	3	1.00	20.35	20.34	78.28

Table 8: Validation of the double-track delay model. 1-5 are from the 54 unused topological subnetwork configurations. 6 is a real rail network

For a given network topology, the simulation delay values for trains are derived from running the simulation model. The relative error is defined as the absolute value of the difference between the simulation delay and predicted (from the delay estimation equation) delay divided by the simulation delay. The mean and the median of the relative error are given in columns 6 and 7. Our observation from these tests is that the delay estimation model estimates data with a high accuracy under normal, expected levels of traffic. But, it also has a tendency to overestimate delay under conditions of high traffic in a subnetwork that could potentially lead to a deadlock. These values are small in number and, therefore, are not removed while collecting descriptive statistics. Instead, they are considered as extreme values. Hence, in this case, the median of the relative error proves to be a more robust measure than the mean, and looking at the median of the relative error gives an estimate of the effect of these extreme values on the mean of the relative error. The final column lists the portion of the data set with a corresponding relative error within 20%, which gives an estimate of the number of these extreme values.

We next investigate the conditions when our delay model provides small relative error terms since the previous analysis showed that the relative error can be large under extreme heavy traffic. We consider two cases in this analysis: light and medium traffic. In other words, we present the performance of our delay estimation model for double-tracks without including the extreme values of traffic indicative of a high degree of congestion. In Table 9 below, we compare the predicted delay value with the delay value obtained from the simulation model for the same six network configurations listed in Table 8. We compute the portion of the data set, with relative error in the delay values within 10%, for two different cases. In the row labeled 'Case 1', we select the data where just the right number of trains co-exist in the network so as to maintain the minimum safety distance, that is, the number of trains $\leq 2L/1.5$, assuming trains are 1.5 miles long and the safety distance is of a similar length. In other words, we compare the predicted and simulation delay values for low traffic densities without any queues at either station. The values listed in this row show that our delay estimation model performs fairly well under low traffic conditions. In the row labeled 'Case 2', we select data values where the quantity *number of trains*$/2L$ is $\leq 80\%$. This can be thought of as the utilization of the network being $\leq 80\%$. The maximum number of trains in this case will be higher than the number of trains in Case 1. Since all cannot be accommodated simultaneously, there might be some queuing occurring at either or both stations. Under this case of medium traffic densities, our delay estimation model continues to perform well, as more than 90% of the data is within 10% relative error for all the six network configurations.

Test Config.	**1**	**2**	**3**	**4**	**5**	**6**
Case 1	95.02	94.87	94.10	94.67	93.44	93.32
Case 2	92.54	93.41	90.19	92.75	91.65	90.17

Table 9: Validation of the double-track delay model. Case 1: Low traffic conditions. Case 2: Medium traffic conditions.

6.5.2 Delay Estimation for a Single-track Subnetwork

Delay estimation for a single-track subnetwork is analogous to the delay estimation for a double-track subnetwork. In Table 10 below, we list the three grades of values for the network topology parameters that were used for estimating delay for single-track subnetworks.

	LO	**MED**	**HI**
Length (mi)	10.0	15.0	20.0
Speed-limit (mi/hr)	15	35	55
Crossings	2	3	4
Spacing (%)	0.70	0.85	1.00

Table 10: Settings for network topology parameters for single-track simulations

As in the case of a double-track, we develop a one-third fractional factorial design to run the simulations and develop the delay model. Table 11 below lists the 27 treatment combinations selected from the 81 possible by assuming third-order and higher interactions to be negligible.

	L	**V**	**C**	**S**		**L**	**V**	**C**	**S**
1	10	35	4	0.70	15	20	55	2	1.00
2	10	55	4	1.00	16	20	55	3	0.85
3	15	35	3	0.70	17	20	15	4	1.00
4	20	15	3	0.70	18	10	35	2	1.00
5	10	55	2	0.85	19	20	55	4	0.70
6	15	55	2	0.70	20	15	55	4	0.85
7	15	15	3	0.85	21	10	15	3	1.00
8	10	15	4	0.85	22	15	15	2	1.00
9	20	35	2	0.70	23	15	35	4	1.00
10	15	55	3	1.00	24	20	35	4	0.85
11	10	15	2	0.70	25	20	15	2	0.85
12	20	35	3	1.00	26	10	55	3	0.70
13	15	15	4	0.70	27	15	35	2	0.85
14	10	35	3	0.85					

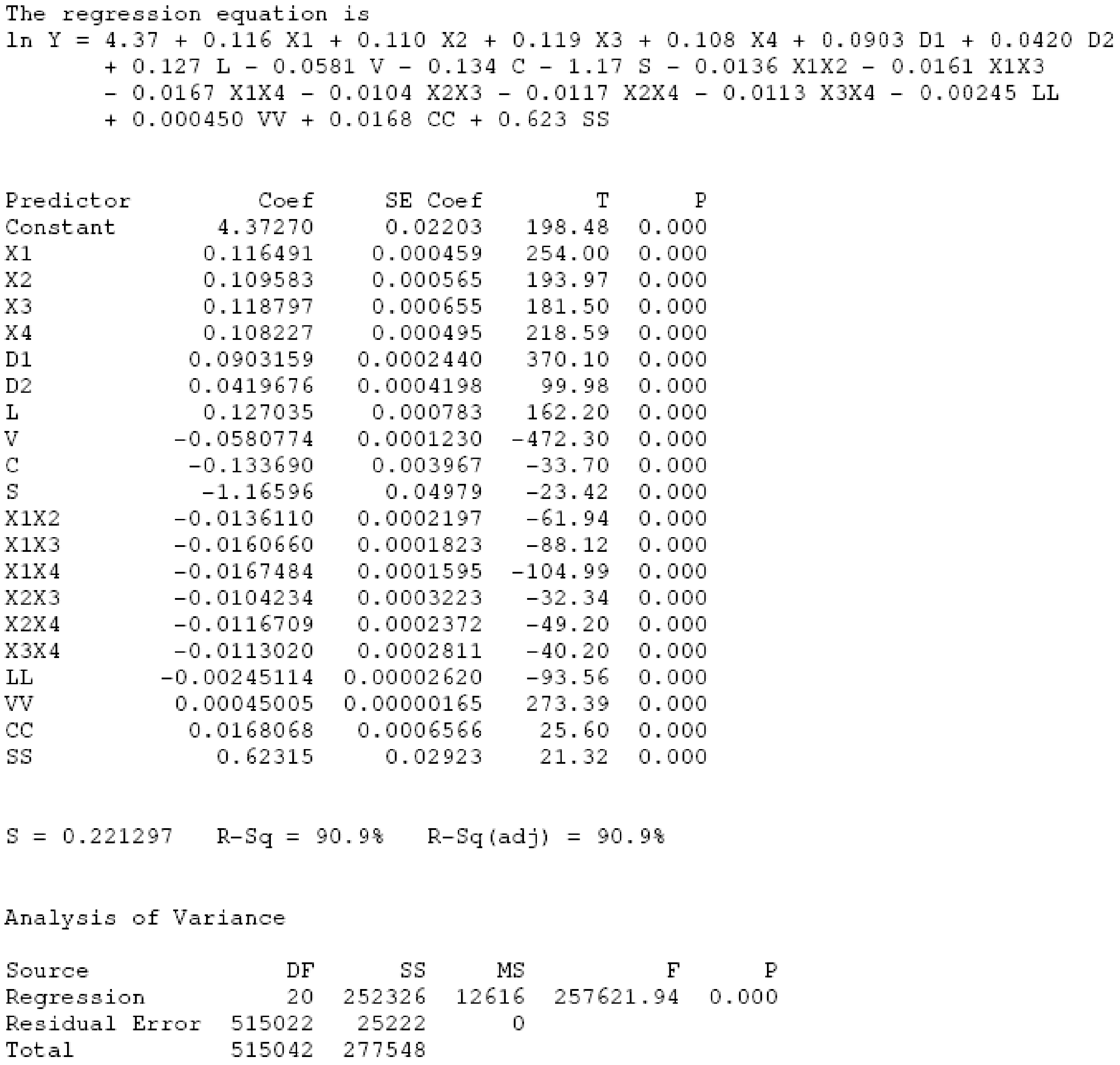
Table 11: 27 parameter combinations considered in the one-third fractional factorial design for single-track simulations

All the data collected from these 27x25 simulations are combined to fit a regression model. This model has an adjusted R-squared value of 79.4%. The normality plot and the plot of the residuals versus the predicted response depict a violation of the normality and homoscedasticity assumptions. A remedial Box-Cox transformation is carried out. Similar to the case of a double-track, the natural logarithm of the response variable, Y, is used as the transformed response. We begin with a regression model containing all the second-order interaction terms of the X_i's and the topology parameters, and by using backward elimination we derive a regression model to estimate delay on a single-track subnetwork.

	Regression Parameters	**Adjusted R-Sq (%)**	**Eliminated Term(s)**
1	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X1,X1X2,X1X3,X1X4,X2X2, X2X3,X2X4,X3X3,X3X4,X4X4, LL,LV,LC,LS,VV,VC,VS,CC, CS,SS	91.1	–
2	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X1,X1X2,X1X3,X1X4,X2X2, X2X3,X2X4,X3X3,X3X4,X4X4, LL,VV,CC,SS	91.0	LV,LC,LS,VC,VS,CS
3	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X1,X1X2,X1X3,X1X4,X2X2, X2X3,X2X4,X3X4,X4X4, LL,VV,CC,SS	91.0	X2X2
4	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X2,X1X3,X1X4,X2X2,X2X3, X2X4,X3X4,X4X4,LL,VV,CC,SS	91.0	X3X3
5	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X2,X1X3,X1X4,X2X3,X2X4, X3X4,X4X4,LL,VV,CC,SS	91.0	X4X4
6	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X2,X1X3,X1X4,X2X3,X2X4, X3X4,LL,VV,CC,SS	90.9	X1X1
7	X1,X2,X3,X4,D1,D2,L,V,C,S, X1X2,X1X3,X1X4,X2X3,X2X4,X3X4	89.4	LL,VV,CC,SS
8	X1,X2,X3,X4,D1,D2,L,V,C,S	88.2	X1X2,X1X3,X1X4, X2X3,X2X4,X3X4

Table 12: Backward elimination for single-track subnetwork simulations.

From Table 12, we note that the regression model in row 6 has the highest adjusted R-squared value with only the significant single-order and interaction terms included. The regression-based delay estimation equation for a single-track subnetwork is given below. The heterogeneity in the traffic flowing in the opposite direction and the quadratic interaction terms of the network parameters affect the delay experienced by a train traveling on a single-track.



```
The regression equation is
ln Y = 4.37 + 0.116 X1 + 0.110 X2 + 0.119 X3 + 0.108 X4 + 0.0903 D1 + 0.0420 D2
       + 0.127 L - 0.0581 V - 0.134 C - 1.17 S - 0.0136 X1X2 - 0.0161 X1X3
       - 0.0167 X1X4 - 0.0104 X2X3 - 0.0117 X2X4 - 0.0113 X3X4 - 0.00245 LL
       + 0.000450 VV + 0.0168 CC + 0.623 SS


Predictor         Coef      SE Coef        T      P
Constant       4.37270      0.02203   198.48  0.000
X1            0.116491     0.000459   254.00  0.000
X2            0.109583     0.000565   193.97  0.000
X3            0.118797     0.000655   181.50  0.000
X4            0.108227     0.000495   218.59  0.000
D1           0.0903159    0.0002440   370.10  0.000
D2           0.0419676    0.0004198    99.98  0.000
L             0.127035     0.000783   162.20  0.000
V           -0.0580774    0.0001230  -472.30  0.000
C            -0.133690     0.003967   -33.70  0.000
S             -1.16596      0.04979   -23.42  0.000
X1X2        -0.0136110    0.0002197   -61.94  0.000
X1X3        -0.0160660    0.0001823   -88.12  0.000
X1X4        -0.0167484    0.0001595  -104.99  0.000
X2X3        -0.0104234    0.0003223   -32.34  0.000
X2X4        -0.0116709    0.0002372   -49.20  0.000
X3X4        -0.0113020    0.0002811   -40.20  0.000
LL         -0.00245114   0.00002620   -93.56  0.000
VV          0.00045005   0.00000165   273.39  0.000
CC           0.0168068    0.0006566    25.60  0.000
SS             0.62315      0.02923    21.32  0.000


S = 0.221297   R-Sq = 90.9%   R-Sq(adj) = 90.9%


Analysis of Variance

Source              DF      SS     MS          F      P
Regression          20  252326  12616  257621.94  0.000
Residual Error  515022   25222      0
Total           515042  277548
```

Figure 13: Detailed regression results for model 6 in the single-track simulation.

The single-track delay model is validated in a similar fashion as the double-track delay model. In Table 13 below, rows 1-5 show the performance of the delay model with respect to the delay obtained by running simulations on 5 randomly chosen subnetwork configurations that were not used in the one-third fractional factorial design. Row 6 shows the performance of the delay model on an actual single-track subnetwork existing in the Downtown Los Angeles to the Ports railway network.

	L	**V**	**C**	**S**	**Relative Error, Mean (%)**	**Relative Error, Median (%)**	**Percent within 20% rel. error**
1	10	35	3	1.00	12.67	10.57	74.36
2	10	55	2	1.00	16.92	14.02	74.10
3	15	15	4	0.85	17.21	12.68	68.76
4	15	55	2	1.00	17.85	15.34	70.31
5	20	35	3	0.70	18.91	14.00	75.19
6	11.43	55	2	1.00	14.11	10.54	71.96

Table 13: Validation of the single-track delay model. 1-5 are from the 54 unused network configurations. 6 is an actual rail network.

The relative error is defined as the absolute value of the difference between the simulation delay and predicted (from the delay estimation equation) delay divided by the simulation delay. The mean and the median of the relative error are given in columns 6 and 7. Our observation from these tests is that the delay estimation model estimates data with a high accuracy under normal levels of traffic. But, it also has a tendency to overestimate delay under conditions of high traffic in a subnetwork that could potentially lead to a deadlock. These values are small in number and, therefore, are not removed while collecting descriptive statistics. Instead, they are considered as extreme values. In the presence of these extreme values, the median of the relative error proves to be a more robust measure than the mean, and looking at the median of the relative error gives an estimate of the effect of these extreme values on the mean of the relative error. The final column lists the portion of the data set with a corresponding relative error within 20%, which gives an estimate of the number of these extreme values.

We next investigate the conditions when our delay model provides small relative error terms since the previous analysis showed that the relative error can be large under extreme heavy traffic. We consider two cases in this analysis: light and medium traffic. In Table 14 below, we compare the predicted delay value with the delay value obtained from the simulation model for the same five network configurations listed in Table 13. We compute the portion of the data set, with relative error in the delay values within 10%, for two different cases. In the row labeled 'Case 1', we select the data where just the right number of trains co-exist in the network so as to maintain the minimum safety distance, that is, the number of trains $\leq (L + C * 1.5)/1.5$, assuming trains and sidings are 1.5 miles long and the safety distance is of a similar length. In other words, we compare the predicted and simulation delay values for low traffic densities without any queues at either station. The values listed in this row show that our delay estimation model performs fairly well under low traffic conditions. In the row labeled 'Case 2', we select data values where the quantity $number\ of\ trains\ /(L + C * 1.5)$ is $\leq$ 80%. This can be thought of as the utilization of the network being $\leq$ 80%. The maximum number of trains in this case will be higher than the number of trains in Case 1. Since all trains cannot be accommodated simultaneously, there might be some queuing occurring at either or both stations. Under this case of light to medium traffic densities, our delay estimation model continues to perform well as is shown in the table below.

Test Config.	**1**	**2**	**3**	**4**	**5**	**6**
Case 1	90.40	91.77	91.23	90.14	89.36	88.76
Case 2	88.65	89.11	87.43	88.89	87.54	86.97

Table 14: Validation of the single-track delay model. Case 1: Low traffic levels. Case 2: Medium traffic values.

7 An Approximation-based Solution Procedure

Using the IP model presented in Section 4 together with the delay estimation techniques discussed in Section 6, we can now tackle the problem of routing and scheduling trains on large and complex rail networks with combinations of track segments of varying lengths, speed-limits, traffic conditions, and number of crossings, junctions and sidings. In this section, we present an approximation-based solution procedure to solve the NP-hard train routing and scheduling problem (IP). It is computationally impractical to identify an optimal solution to problem (IP) for realistic size rail networks. At the planning level, the most important decisions are the identification of the route for each train and the departure times at the origin stations. Hence, our solution procedure focuses on identifying suitable integer solutions from the LP relaxation for these two decisions. The decisions related to the detailed sequence of trains at the subsequent stations in the network can be made later at the operational level. Hence, our proposed procedure determines only the route that each train takes and the order of departure at the origin destination.

Our proposed solution procedure is as follows:

a. Observing the guidelines for aggregation listed in Section 4.0.4, select suitable network sections with similar topological characteristics that can be aggregated.

b. For each subnetwork, generate delay estimation equations as shown in Section 6.2.

c. Include constraints in the IP model reflecting the delay experienced in traveling across each subnetwork. This is similar to the discussion in Section 6.3.

d. For large rail networks, the IP model formulated in step (c) cannot be solved optimally using solvers such as CPLEX. Due to this reason, a LP-relaxation of the problem is solved.

e. The output from solving the LP-relaxation model gives the routes taken by each train in traveling from the origin to the destination station. Since the $Y_{i,j,h,t}$ variables could be fractional, we compare the values of all possible $\sum_{t\in T} Y_{i,d_h,h,t}$ variables for each train h such that $(i, d_h, h) \in S$, and assign the train h to the route with the largest $\sum_{t\in T} Y_{i,d_h,h,t}$ value.

f. Another information derived from the LP-relaxation output is the order of departure from the origin nodes. To circumvent the problem posed by the fractional decision variable values, we solve the following maximal matching problem to decide the departure times for each train from their respective origins.

$$\mathbf{IP}_1 : \quad \begin{array}{ll} \text{Maximize} & \displaystyle\sum_{h\in H, t\in T} \beta_{h,t} X_{h,t} \\ \text{subject to} & \end{array}$$

$$\sum_{h\in H} X_{h,t} = 1 \quad \forall \quad t \in T \tag{16}$$

$$\sum_{t\in T} X_{h,t} = 1 \quad \forall \quad h \in H \tag{17}$$

In the above formulation, $X_{h,t} = 1$ if train h departs its origin at time t, else it is 0. $\beta_{h,t} = Y_{o_h,j,h,t} - Y_{o_h,j,h,t-1}$ such that $(o_h, j, h) \in S$. Constraint (16) enforces the condition that only one train can depart at any time instant t from the origin station, and constraint (17) ensures that train h departs only once over the planning time horizon.

7.1 Experimental Results

We test the performance of the proposed solution procedure to route and schedule train on large and complex networks. For this, we considered a portion of the rail network in the Los Angeles area, and four other networks generated using the settings for the network topology parameters for double-track and single-track simulations, presented in Tables 5 and 10 respectively. To generate these experimental networks, we first built a skeleton framework shown in Figure 14 below. It contains 3 main routes, 2 stations - $ST1$ and $ST2$, and 8 aggregate nodes - AB, BC, CD, EF, FG, GH, HI and CH. There are two instances where trains traveling between the two stations need to make a routing decision.

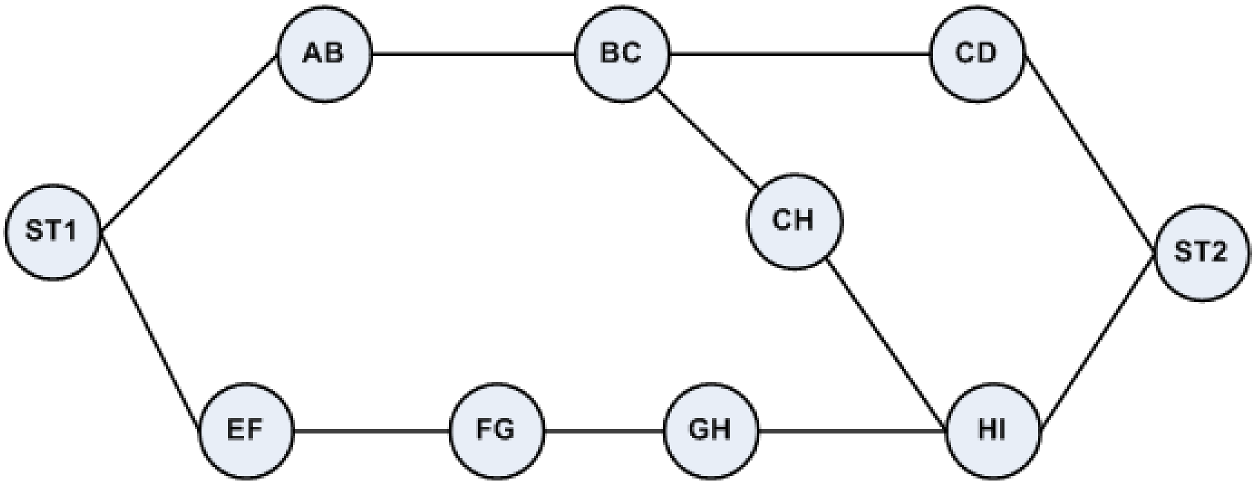


Figure 14: Skeleton framework of the test network.

The next step is to decide the topology of each of the aggregate nodes. For this, we first randomly decided whether an aggregate node would be a single-track or a double-track segment. Then, from the 81 possible combinations of the parameters presented in Tables 5 and 10, each node was assigned a random value of L, V, C and S. The corresponding networks were drawn and the data files required to run simulations and derive the delay expressions for each aggregate node were prepared.

We evaluate the effectiveness of the generated planned routes and departure times from the proposed solution procedure on a simulation model developed by Lu et al. (2004) that captures many of the realistic characteristics of actual rail operations. The simulation model takes into consideration train lengths, acceleration and deceleration rates, and many other attributes of actual rail operations. The performance of our solution methodology, denoted by travel time - LP and delay - LP, was compared with the quality of solution obtained when the routes are determined by a greedy procedure and the order of departure is determined by the construction heuristic developed by Lu et al. (2004). The greedy procedure assigns trains to the route with the least traffic currently on it. The values from this procedure are denoted by travel time - construction and delay - construction. In both cases, we let the construction heuristic built into the simulation model take care of the sequencing of the trains at the intermediate stations. Comparison of these results is presented in Table 15 below.

	Network	**No. of trains**	**LP-relax. obj. value**	**Partial solution obj. value**	**Travel time:construction**	**Delay - construc-**	**Travel time - LP**	**Delay - LP tion**	**% dec. in travel time**	**% dec. in delay**
1	LA rail network	80	2480	2680	292.85	239.79	246.32	191.29	15.89	20.23
2	Network 2	80	3280	3286	90.06	30.14	83.11	23.50	7.73	22.07
3	Network 3	56	1624	1627	144.90	37.95	141.17	33.76	2.58	11.00
4	Network 4	80	2956	3017	256.82	62.24	215.19	47.05	16.21	24.41
5	Network 5	56	1624	1706	142.72	26.31	134.20	23.01	5.97	12.55

Table 15: Validation of the single-track delay model. 1-5 are from the 54 unused network configurations. 6 is an actual rail network.

In the table above, the LP-relaxation (LP-R) objective was obtained by replacing the constraint that the $Y_{i,j,h,t}$ variables should be either 0 or 1 by a weaker constraint that these variables should belong to the interval $[0, 1]$. This procedure expands the feasible solution space, and the NP-hard integer program proposed in Section 4.0.4 is transformed to a related linear program that can be solved efficiently using a commercial optimization solver such as CPLEX. Since the original IP is a minimization problem, an optimal solution to the LP-R would serve as a lower bound to the optimal solution of **IP**. A fractional solution to the LP-R is rounded to obtain routes and order of departures for a given set of trains, while ensuring that none of the constraints are violated. These would constitute a partial feasible solution to **IP**. To measure the quality of this approximated solution, we record their objective function values in Table 15 above in the column *partial solution objective*. The LP-R objective values serve as a lower bound to the partial solution objective values. Comparing the LP-relaxation and partial solution objective values in the table, we can conclude that since these values are close in most cases, the rounding of the fractional solution to the LP-R does not aversely affect the quality of the solution.

In terms of the effectiveness of our procedure to determine routes and departure times at the origin station, Table 15 shows that the percent decrease in travel times and delays by using our methodology are as much as 16% and 24% respectively, which is a significant amount of saving. This means that the approximate routes and the order of departures developed from the fractional solution to the LP-R are of a higher quality than the greedy routes and the departure order used by the construction heuristic.

8 Implementation

This project addresses the focus area of Commercial Goods Movement and International Trade. Train transportation is a practical way to move commodities from ports to inland destinations because of the size of the country. According to a study conducted by the Association of American Railroads, about 40% of all freight in the US is moved by trains and the demand of freight is expected to double by year 2020. This rapid growth has already introduced congestion in rail network systems. Methodologies that can improve efficiency of train movement will dramatically improve the overall logistics network and reduce traffic congestion. In particular, this research also addresses the reduction of congestion and environmental impacts of goods movement in metropolitan areas since efficient rail operations encourages usage of freight transport by rail instead of truck, hence alleviating the traffic congestion along the 710 corridor.

The integrated routing and capacity management model, the delay estimation techniques and the approximation solution scheme developed as part of this funded research project were tested on real-world data on train movement on the Los Angeles area rail network. The performance of our modeling technique was compared with the performance of the construction heuristic proposed by Lu et al.(2004) on a wide variety of rail networks. The algorithm developed as part of our research work outperforms the current state-of-the-art and its implementation would require the existence of suitable optimization software tools, and access to rail data such as traffic volumes, rail network design plans, train lengths, speed-limits, acceleration and deceleration rates, headway regulations etc.

9 Conclusion

In the United States, rail transportation offers a smooth and viable mode to transport freight transcontinentally. Due to the rapid increase in international trade, there has been a significant increase in railway traffic across the nation. In order for rail transportation to continue to be efficient, an in-depth understanding and study of how congestion occurs in railway network, how to prevent it and how to reduce the delay experienced by trains is required. While there exists prior work in the areas of railway routing and delay estimation, there has been little work in trying to route trains along the routes with the least possible delay, while accounting for the complexities and non-linearities that exist in a real-world railway operation. To address this gap, we present a novel approach involving an integrated routing, scheduling and capacity management model that can used to estimate and minimize the delay for a train in going from its origin to its destination. In order to evaluate the IP model's performance more accurately and to solve problems of larger sizes, aggregation techniques are implemented. From the results presented in Section 4.1, we infer that our combined routing and scheduling model performs better than the simulation model. Later on, a regression-based technique is used to estimate the delay across sub-networks represented by an aggregated node. We first develop one regression equation for every aggregated node. Extending on this achievement and using sophisticated statistical techniques we developed a generic regression-based delay function for an aggregated node. As shown is Section 6.5, this procedure can be used to accurately estimate delay on network segments by running simulations on representative networks, thereby saving the effort of running individual simulations. Finally, we propose an approximation-based solution methodology to solve the routing and scheduling problem over large rail networks, by integrating the concepts presented above. Unlike many scheduling models developed so far, the modeling procedure presented here has a distinctive advantage of being adaptable to any railway network configuration and schedule. As part of future work, we propose to develop heuristic solution procedures, such as genetic algorithms, to close the gap between the LP-relaxation and the partial solution objective values presented in Table 15. We also intend to develop tighter lower bounds to the integrated routing and capacity management model.

References

[1] Ahuja, R. K., Jha, K. C., and Liu, J. Solving real-life railroad blocking problems. *Interfaces 37, No. 5* (2007), 404–419.

[2] Barnhart, C., Jin, H., and Vance, P. H. Railroad blocking: A network design application. *Operations Research 48* (2000), 603–614.

[3] Burdett, R. L., and Kozan, E. Techniques for absolute capacity determination in railways. *Transportation Research Part B 40* (2006), 616–632.

[4] Caprara, A., Fischetti, M., and Toth, P. Modeling and solving the train timetabling problem. *Operations Research 50(5)* (2002), 851–861.

[5] Carey, M. Ex ante heuristic measure of schedule reliability. *Transportation Research Part B 33* (1999), 473–494.

[6] Carey, M., and Kwiecinski, A. Stochastic approximation to the effects of headways on knock-on delays of trains. *Transportation Research Part B 28B(4)* (1994), 251–267.

[7] Carey, M., and Lockwood, D. A model, algorithms and strategy for train pathing. *Journal of the Operational Research Society 46* (1995), 988–1005.

[8] Chen, B., and Harker, P. T. Two moments estimation of the delay on a single-track rail line with scheduled traffic. *Transportation Science 24* (1990), 261–275.

[9] Cordeau, J. F., Toth, P., and Vigo, D. A survey of optimization models for train routing and scheduling. *Transportation Science 32(4)* (1998), 380–404.

[10] Crainic, T. G., Ferland, J. A., and Rousseau, J. M. A tactical planning model for rail freight transportation. *Transportation Science 18* (1984), 165–184.

[11] Crainic, T. G., and Gendreau, M. Approximate formulas for the computation of connection delays under capacity restrictions in rail freight transportation. Tech. Rep. 438, University of Montreal, Montreal, Canada, 1985.

[12] D'Ariano, A. *Improving Real-time Train Dispatching: Models, Algorithms and Applications*, t2008/6 ed. TRAIL Thesis Series, The Netherlands, 2008.

[13] de Kort, A. F., Heidergott, B., and Ayhan, H. A probabilistic (max,+) approach for determining railway infrastructure capacity. *European Journal of Operational Research 148* (2003), 644–661.

[14] Dessouky, M. M., and Leachman, R. C. A simulation modeling methodology for analyzing large complex rail networks. *Simulation 65:2* (1995), 131–142.

[15] Dessouky, M. M., Lu, Q., and Leachman, R. C. An exact solution procedure for determining the optimal dispatching times for complex rail networks. Proceedings of the Winter Simulation Conference, 2002, pp. 141–152.

[16] Dessouky, M. M., Lu, Q., Zhao, J., and Leachman, R. An exact solution procedure for determining the optimal dispatching times for complex rail networks. *IIE Transactions 38* (2006), 141–152.

[17] Frank, O. Two-way traffic in a single line of railway. *Operations Research 14* (1966), 801–811.

[18] Gibson, S., Cooper, G., and Ball, B. Developments in transport policy: The evolution of capacity charges on the uk rail network. *Journal of Transport Economics and Policy 36* (2002), 341–354.

[19] Gorman, M. F. An application of genetic and tabu searches to the freight railroad operating plan problem. *Annals of Operations Research 78* (1998), 51–69.

[20] Greenberg, B. S., Leachman, R. C., and Wolff, R. W. Predicting dispatching delays on a low speed, single tack railroad. *Transportation Science 22(1)* (1988), 31–38.

[21] Hallowell, S. F., and Harker, P. T. Predicting on-time line-haul performance in scheduled railroad operations. *Transportation Science 30* (1996), 364–378.

[22] Harker, P. T., and Hong, S. Two moments estimation of the delay on a partially double-track rail line with scheduled traffic. *Transportation Research Forum 30* (1990), 38–49.

[23] Higgins, A., and Ferreira, L. Modeling single line train operations. *Transportation Science 1489* (1995), 9–16.

[24] Higgins, A., and Kozan, E. Modeling train delays in urban networks. *Transportation Science 32(4)* (1998), 251–356.

[25] Higgins, A., Kozan, E., and Ferreira, L. Optimal scheduling of trains on a single line track. *Transportation Research Part B 30B(2)* (1996), 147–161.

[26] Huisman, T., and Boucherie, R. J. Running times on railway sections with heterogeneous train traffic. *Transportation Research Part B 35* (2001), 271–292.

[27] Huntley, C. L., Brown, D. E., Sappington, D. E., and Markowicz, B. P. Freight routing and scheduling at csx transportation. *Interfaces 25(3)* (1995), 58–71.

[28] Kaas, A. H. *Methods to Calculate Capacity of Railways.* PhD thesis, Dept. of Planning, Technical University of Denmark, 1998.

[29] Keaton, M. H. Designing railroad operating plans: A dual adjustment method for implementing lagrangean relaxation. *Transportation Science 26* (1992), 263–279.

[30] Kraay, D. R., and Harker, P. T. Real-time scheduling of freight railroads. *Transportation Research Part B 29B(3)* (1995), 213–229.

[31] Krueger, H. Parametric modeling in rail capacity planning. Proceedings of the Winter Simulation Conference, pp. 1194–1200.

[32] Landex, A., Kaas, A. H., and Hansen, S. Railway operation. Report 4, Centre for Traffic and Transport, Technical University of Denmark, 2006.

[33] Leachman, R. C. Inland empire railroad main line advanced planning study. Tech. rep., Prepared for the Southern California Association of Governments, Contract number 01-077, Work element number 014302, October 1, 2002.

[34] Lu, Q., Dessouky, M. M., and Leachman, R. C. Modeling of train movements through complex networks. *ACM Transactions on Modeling and Computer Simulation 14* (2004), 48–75.

[35] Montgomery, D. C. *Design and Analysis of Experiments*, 2nd ed. John Wiley and Sons, 1984.

[36] Myers, R. H., and Montgomery, D. C. *Response Surface Methodology*, 2nd ed. Wiley Series in Probability and Statistics, 2002.

[37] Newton, H. N. *Network Design Under Budget Constraints with Application to the Railroad Blocking Problem.* PhD thesis, Auburn University, Auburn, AL, 1996.

[38] Özekici, S., and Şengör, S. On a rail transportation model with scheduled services. *Transportation Science 28(3)* (1994), 246–255.

[39] Petersen, E. R. Over the road transit time for a single track railway. *Transportation Science 8* (1974), 65–74.

[40] Petersen, E. R., and Taylor, A. J. A structured model for rail line simulation and optimization. *Transportation Science 16* (1982), 192–206.

[41] Pritsker, A. A. B., and O'Reilly, J. J. *Simulation with Visual SLAM and AweSim*, 2nd ed. John Wiley and Sons, New York and Systems Publishing Corporation, West Lafayette, Indiana, 1999.

[42] Wendler, E. The scheduled waiting time on railway lines. *Transportation Research Part B 41* (2007), 148–158.

[43] Winston, C. The success of the stagger rail act of 1980. Tech. Rep. Page 19, Brookings Institution, September, 2005.

[44] Yuan, J. *Stochastic Modeling of Train Delays and Delay Propagation in Stations.* Ph.D Thesis, Delft University of Technology, The Netherlands, 2006.

[45] Yuan, J., and Hansen, I. A. Optimizing capacity utilization of stations by estimating knock-on train delays. *Transportation Research Part B 41* (2007), 202–217.