UNIVERSITÉ DE MONTRÉAL

AFFECTATION DES LOCOMOTIVES ET DES WAGONS AUX TRAINS DE PASSAGERS

JEAN-FRANÇOIS CORDEAU
DÉPARTEMENT DE MATHÉMATIQUES
ET DE GÉNIE INDUSTRIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIAE DOCTOR (Ph.D.)
(MATHÉMATIQUES DE L'INGÉNIEUR)
JUIN 1999

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée:

AFFECTATION DES LOCOMOTIVES ET DES WAGONS AUX TRAINS DE PASSAGERS

présentée par: CORDEAU Jean-François

en vue de l'obtention du diplôme de: Philosophiae Doctor

a été dûment acceptée par le jury d'examen constitué de:

M. SAVARD Gilles, Ph.D., président

M. SOUMIS François, Ph.D., membre et directeur de recherche

M. DESROSIERS Jacques, Ph.D., membre et codirecteur de recherche

M. GAUVIN Jacques, Ph.D., membre

M. QUEYRANNE Maurice, Ph.D., membre

Remerciements

Je remercie très sincèrement M. François Soumis, directeur de recherche, pour son aide tant didactique que matérielle. Sa disponibilité, son jugement et ses connaissances m'ont permis de progresser très rapidement. Je le remercie également de la flexibilité dont il a fait preuve en me laissant explorer des avenues qui lui paraissaient au départ peu prometteuses.

Au cours des trois dernières années, j'ai aussi beaucoup profité de l'aide de M. Jacques Desrosiers, codirecteur de recherche. Sa collaboration à la rédaction des articles fut précieuse pour en améliorer à la fois le contenu et la présentation.

Je désire par ailleurs souligner le travail considérable de MM. Guy Desaulniers et Norbert Lingaya qui ont contribué significativement à plusieurs aspects du projet chez VIA Rail.

Je tiens enfin à remercier MM. Paolo Toth et Daniele Vigo de l'Université de Bologne qui m'ont accueilli pendant trois mois en 1996.

Le travail présenté dans cette thèse a bénéficié de bourses du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) et du Fonds pour la formation de chercheurs et l'aide à la recherche (FCAR).

Résumé

Cette thèse traite du problème d'affectation des locomotives et des wagons aux trains dans le contexte particulier du transport de passagers. Étant donné un horaire de trains et un ensemble d'unités d'équipement disponibles, le problème consiste à affecter à chaque train prévu à l'horaire un nombre suffisant de locomotives et de wagons compatibles, tout en respectant certaines contraintes relatives à l'opération des trains et à l'utilisation du matériel roulant. En dépit de sa ressemblance avec le problème d'affectation des locomotives aux trains de marchandises, ce problème exige une approche différente en raison de la nature des interactions qui existent entre les différents types d'équipement.

Afin de résoudre le problème, nous proposons différentes approches originales basées sur des modèles multi-flots comportant à la fois des contraintes et des variables additionnelles. Dans ces modèles, une unité de flot représente une pièce d'équipement ou encore un groupe de pièces utilisées sur un même train. L'affectation des différents types d'équipement ne peut cependant se faire de manière individuelle et indépendante. Les variables et contraintes additionnelles ont donc pour rôle de traduire les multiples interactions liant les locomotives et les wagons affectés à chaque train. Une partie importante de la thèse est consacrée au développement de méthodes de décomposition permettant de traiter efficacement ces interactions.

Nous présentons d'abord une revue complète de la littérature reliée à l'utilisation de l'optimisation mathématique en transport ferroviaire. Cette revue porte sur les principaux problèmes rencontrés en transport de marchandises et en transport de passagers. Pour chaque catégorie de problèmes, nous proposons une classification des

modèles et décrivons leurs principales caractéristiques en insistant sur leur structure et sur les méthodes de résolution utilisées. Les modèles décrits sont regroupés en deux grandes catégories: les modèles de routage et les modèles de fabrication d'horaires et d'affectation d'équipement. Cette dernière catégorie inclut le problème d'affectation des locomotives et des wagons aux trains de passagers.

Nous proposons ensuite trois approches pour résoudre ce problème. La première approche se fonde sur un modèle très complet incorporant un large éventail de possibilités et de contraintes nécessaires dans une application pratique. Ce modèle a été développé en fonction des besoins spécifiques d'une entreprise canadienne mais peut néanmoins être adapté à diverses situations. En plus des contraintes d'entretien et des possibilités de substitution entre certains types d'équipement, la formulation comporte des pénalités pour réduire le couplage et le découplage de wagons durant les connexions entre deux services consécutifs. Ce modèle en nombres entiers est résolu par une méthode de séparation et d'évaluation progressive dans laquelle les relaxations linéaires sont résolues par une décomposition de Dantzig-Wolfe. Cette approche est au coeur d'un système complet maintenant en opération chez VIA Rail.

La seconde approche est basée sur un modèle plus simple mais possédant une structure très flexible. Ce modèle simplifié traduit les difficultés fondamentales du problème découlant des combinaisons de pièces d'équipement et de leur effet sur la vitesse d'opération, mais n'incorpore pas les éléments plus complexes tels que les contraintes d'entretien ou les possibilités de substitution. La formulation utilisée diffère de la précédente et se prête bien à une approche de résolution basée sur une décomposition au niveau des variables. Nous proposons donc une approche de décomposition de Benders qui, grâce à certaines techniques permettant d'accélérer l'algorithme, s'avère très efficace. L'approche est également comparée avec des méthodes alternatives, basées sur la relaxation lagrangienne ou la décomposition de

Dantzig-Wolfe, dont la performance est de loin inférieure en raison de la formulation du problème.

La dernière partie de la thèse présente des extensions au modèle simplifié qui ont pour but de le rendre mieux adapté à des applications réelles. Nous décrivons donc une formulation étendue incorporant les contraintes d'entretien, les possibilités de substitution ainsi que les pénalités pour le couplage et le découplage de pièces d'équipement. Les contraintes d'entretien sont introduites en remplaçant le problème de flot associé à chaque type d'équipement par un problème multi-flots. Ces ajouts alourdissent considérablement le modèle, mais un algorithme efficace basé sur la décomposition de Benders est obtenu en résolvant d'abord une relaxation du problème dans laquelle les contraintes d'entretien ne sont pas imposées. Ceci permet d'obtenir une très bonne approximation de la solution optimale du problème et de générer un ensemble de contraintes accélérant ensuite considérablement l'algorithme. De plus, la génération de coupes Pareto-optimales permet d'obtenir un gain de vitesse très appréciable sur certaines instances. Pour les plus grandes instances, les problèmes multi-flots sont résolus par une décomposition de Dantzig-Wolfe. Cette dernière approche combine donc la décomposition de Benders et la décomposition de Dantzig-Wolfe à l'intérieur d'une méthode de séparation et d'évaluation progressive.

En somme, les principales contributions de cette thèse sont de proposer des modèles détaillés et flexibles pour l'affectation des locomotives et des wagons aux trains de passagers, d'adapter différentes méthodes de décomposition pour résoudre ces modèles, et de présenter des idées permettant de les résoudre efficacement à l'aide de la décomposition de Benders. L'utilité pratique des approches présentées est par ailleurs confirmée par leur application à des problèmes réels.

Abstract

This dissertation addresses the problem of assigning locomotives and cars to trains in the special context of passenger transportation. Given a train schedule and a set of available equipment units, the problem is to provide each train with a sufficient number of compatible locomotives and cars while satisfying supplementary constraints pertaining to train operations and rolling stock characteristics. Despite the similarities between this problem and that of assigning engines to freight trains, the former requires a different approach because of the the nature of the interactions that exist between the different types of equipment.

To solve the locomotive and car assignment problem, we propose a number of different approaches based on multi-commodity network flow models with additional constraints and variables. In these models, the flow represents units of equipment or groups of units that are used together on the same train. Because the assignment of the different types of equipment cannot be made individually and independently, the role of the additional variables and constraints is to reflect the numerous interactions that link the locomotives and cars assigned to each train. An important portion of the dissertation is devoted to the development of decomposition approaches that facilitate the efficient treatment of these interactions.

We first present a complete review of the literature concerning optimization models in rail transportation. This survey describes the main problems that are treated in freight and passenger transportation. For each category of problems, we propose a classification of the proposed models and describe their important characteristics by focusing on their structure and the solution methods proposed to solve them. The

models are grouped in two main categories: routing models and scheduling models. The latter category includes the locomotive and car assignment problem which is the topic of this dissertation.

We then propose three approaches for solving this problem. The first approach is based on a very complete model including a large array of possibilities and constraints that are necessary in a practical application. This model was developed according to the specific needs of a Canadian railway but can however be customized to deal with various situations. Besides maintenance constraints and substitution possibilities between equipment types, the formulation incorporates penalties for switching cars during a connection between two successive train services. The integer programming problem is solved by a branch-and-bound method in which the linear relaxations are optimized through a Dantzig-Wolfe decomposition. This approach is the core of a complete system that is now implemented at VIA Rail.

We next present an alternative model that is simpler but possesses a very flexible structure. This model addresses the fundamental difficulties of the problem that arise when combining equipment units of different types, but does not incorporate more complex features such as maintenance constraints or substitution possibilities. The formulation differs from the previous one and is well suited for a variable decomposition approach. We thus propose a solution approach based on Benders decomposition which, with the help of some refinements that yield a significant speed improvement in the algorithm, turns out to be quite effective. The approach is also compared with alternative methods, based on Lagrangian relaxation and Dantzig-Wolfe decomposition, whose performance is largely inferior because of problem formulation.

The last part of the dissertation presents extensions to the simplified model that make it more appropriate for real-life applications. We thus describe an extended

formulation incorporating maintenance constraints, substitution possibilities and penalties for car switching. Maintenance constraints are introduced by replacing the network flow problem for each type of equipment by a multi-commodity network flow problem. These additions make the model more difficult to solve but an efficient algorithm based on Benders decomposition is obtained by first solving a relaxation of the model in which maintenance constraints are removed. This yields a very good approximation of the optimal solution and allows the generation of a set of cuts which then considerably accelerate the algorithm. In addition, the generation of Pareto-optimal cuts produces a considerable speed improvement when solving certain instances. To solve larger instances, the multi-commodity network flow models are solved with a Dantzig-Wolfe decomposition. This last approach thus combines Benders decomposition and Dantzig-Wolfe decomposition within a branch-and-bound method.

In short, the main contributions of this dissertation are to present detailed and flexible models for the assignment of locomotives and cars to passenger trains, to adapt several decomposition methods for solving these models, and to give valuable insight on the efficient implementation of Benders decomposition. In addition, the practical usefulness of these approaches is confirmed by their application to real problems.

Table des matières

Liste des tableaux

Liste des figures

Introduction

Bien que peu populaire en Amérique du Nord, le transport ferroviaire de passagers est néanmoins très répandu à travers le monde. En Suisse, par exemple, les Chemins de Fers Fédéraux (CFF) transportent environ 250 millions de passagers annuellement et la distance totale parcourue chaque jour par leurs trains sur le réseau d'environ 3000 kilomètres correspond à deux fois le tour de la Terre. Pour ce faire, les CFF utilisent plus de 500 locomotives et 4500 wagons répartis en un grand nombre de catégories. Compte tenu de la taille de la population de ce pays et de la superficie de son territoire, le transport ferroviaire y est donc extrêmement populaire. En Italie, il y a près de 16 000 kilomètres de voies ferrées employées pour le transport de passagers. En 1995, 450 millions de passagers ont utilisé le train pour parcourir au total plus de 50 milliards de kilomètres, et l'entreprise d'état italienne FS disposait de plus de 1500 locomotives et 12 000 wagons pour s'acquitter de sa tâche. En France, plus de deux millions de passagers utilisent le transport ferroviaire chaque jour alors qu'environ 10 000 trains parcourent plus d'un million de kilomètres. En Inde, finalement, dix millions d'individus utilisent le train quotidiennement et les Chemins de Fers Indiens possèdent près de 40 000 wagons réservés au transport de passagers.

La popularité du transport par train s'explique de plusieurs manières. D'abord, il s'agit d'un moyen de transport peu polluant et hautement sécuritaire. Ensuite, le transport ferroviaire constitue un mode très pratique puisqu'il permet de se déplacer rapidement en échappant aux bouchons de circulation qui sont fréquents dans la plupart des grandes villes. Lorsque la distance à parcourir est assez courte et que la fréquence des trains est élevée, il est souvent plus rapide de prendre le train que d'utiliser sa voiture. Finalement, le transport par train libère le passager du devoir

de conduire et lui évite ainsi bien des ennuis, tout en lui permettant de travailler ou de se reposer en se déplaçant.

En dépit de l'importance du transport ferroviaire de passagers, on observe dans plusieurs pays un recul marqué de ce mode par rapport au transport routier et au transport aérien. Selon une étude récente (KOPECKY, 1998), la part de marché du transport ferroviaire en Europe est passée de 10% en 1970 à seulement 6% en 1997. La principale raison expliquant ce déclin serait l'insatisfaction des clients quant au service offert, à sa fiabilité, et à son coût trop élevé. Dans l'espoir de rendre le transport par train plus compétitif, des efforts importants ont donc été entrepris par la plupart des transporteurs afin d'améliorer la qualité du service et de réduire les coûts.

Le transport ferrovaire de passagers est une activité très complexe qui côtoie le transport ferroviaire de marchandises et partage avec lui une partie de ses ressources. Plusieurs niveaux de planification et de contrôle des opérations sont donc nécessaires afin d'assurer le bon fonctionnement du système. La planification stratégique consiste principalement en des décisions ayant des implications durant plusieurs années telles que l'acquisition de matériel roulant et les décisions de construction ou d'abandon de segments de voie ferrée. Le niveau tactique concerne la planification à moyen terme qui doit être révisée à tous les trois ou quatres mois selon l'évolution de la demande. La préparation de l'horaire des trains et du plan d'utilisation de l'équipement en sont des exemples. Le niveau opérationnel touche finalement aux décisions de très court terme prises en considérant une information ponctuelle détaillée.

Une part très importante des ressources consacrées à la planification par les entreprises ferroviaires vise en fait les problèmes rencontrés au niveau tactique. La seule préparation d'un horaire coordonnant l'ensemble du service offert est une tâche ardue faisant intervenir de multiples facteurs. Cet horaire doit d'abord être adapté à la répartition géographique et temporelle de la demande. Il doit ensuite faciliter le

voyage des passagers en minimisant les temps de connexion pour les itinéraires les plus courants. Il doit aussi tenir compte des horaires des trains étrangers puisque certains clients empruntent des trains opérés par différents transporteurs au cours d'un même voyage. Cet horaire doit finalement respecter une série de contraintes provenant de la configuration du réseau, de ses politiques d'utilisation, et du matériel disponible.

Une fois l'horaire préparé, il faut ensuite décider de l'affectation de l'équipement aux trains. Cette affectation doit non seulement satisfaire les besoins de chaque train prévu à l'horaire mais également respecter un grand nombre de contraintes imposées par le mode d'utilisation et les caractéristiques du matériel roulant disponible. Évidemment, la séparation de la planification tactique en un problème de fabrication d'horaire et un problème d'affectation d'équipement peut conduire à une solution sous-optimale. Cette approche est néanmoins inévitable en raison de la très grande taille des problèmes.

L'objet de cette thèse est le développement de modèles mathématiques et de méthodes d'optimisation pour l'affectation des locomotives et des wagons aux trains de passagers. Nous nous intéressons plus particulièrement au problème de planification tactique visant à déterminer une affectation de l'équipement disponible aux trains prévus à l'horaire tout en respectant certaines contraintes opérationnelles. Les modèles et les méthodes que nous proposons peuvent aussi être utilisés au niveau stratégique afin d'évaluer la composition optimale de l'ensemble d'équipement nécessaire pour assurer le service décrit par un horaire représentatif.

Dans le cas du transport ferroviaire de marchandises, la formation des trains et l'affectation des locomotives se font habituellement de façon séquentielle. En effet, il serait impensable d'utiliser un modèle où chaque wagon serait représenté explicitement. Ceci donnerait lieu à un modèle beaucoup trop gros pour les méthodes dont on dispose actuellement. On sépare donc le problème de façon à établir d'abord

un plan de transport suivi d'un plan d'affectation des locomotives. En d'autres termes, on décide d'abord des trains que l'on va former et, une fois les caractéristiques de ces trains connues, on décide de l'affectation des locomotives disponibles.

Le transport des passagers se distingue du transport de marchandises pour deux principales raisons: les wagons sont beaucoup moins nombreux et le caractère périodique de la demande fait en sorte qu'il est possible de parvenir à une meilleure planification en traitant à la fois les locomotives et les wagons. En effet, dans la plupart des pays, les trains de passagers fonctionnent selon un horaire révisé à chaque trois ou quatre mois selon l'évolution de la demande. Ainsi, le nombre et les caractéristiques des wagons utilisés sur chaque train varient très peu d'une semaine à l'autre à l'intérieur d'un même trimestre. Il est donc possible d'obtenir un plan global d'utilisation des locomotives et des wagons qui sera répété de façon cyclique pendant quelques mois. De cette manière, on peut réduire à la fois les coûts d'opérations et le nombre d'unités d'équipement nécessaires pour assurer le service.

La principale difficulté du problème d'affectation des locomotives et des wagons aux trains de passagers provient des incompatibilités et interdépendances qui existent entre les différents types d'équipement utilisés par une entreprise donnée. En effet, il est souvent impossible d'utiliser un certain type de wagon avec un certain type de locomotive pour des raisons techniques ou d'homogénéité. Ainsi, même si tous les types d'équipement disponibles peuvent être utilisés sur tous les trains, certaines combinaisons peuvent être interdites. Il existe par ailleurs une interdépendance très forte qui lie les types d'équipement entre eux. Cette interdépendance provient de la vitesse d'opération du matériel qui est déterminée par la plus lente des composantes d'un train. Or, la vitesse d'opération est une donnée très importante. Contrairement au transport de marchandises où les trains fonctionnent souvent sans horaire précis ou peuvent dévier sans trop de conséquences de l'horaire prévu, le transport de passagers

est organisé selon un horaire qui doit être respecté de manière précise afin d'assurer la satisfaction des usagers.

En plus de ces difficultés fondamentales du problème, de nombreuses contraintes régissent l'affectation des locomotives et des wagons. D'abord, les ressources sont généralement limitées et les planificateurs doivent tenir compte des limites sur le nombre d'unités disponibles de chaque type. Ensuite, afin de respecter la réglementation et d'effectuer des travaux mineurs, chaque unité doit être inspectée à intervalle régulier à l'un des centres d'entretien disponibles. Dans certains cas, cette contrainte a peu d'impact puisque l'entretien peut être effectué à l'une quelconque des stations où les trains s'arrêtent à la fin d'un service. Dans d'autres situations, seules quelques stations possèdent l'équipement et le personnel nécessaires pour effectuer les opérations d'entretien. Il faut alors s'assurer que chaque pièce d'équipement soit régulièrement acheminée vers l'une des stations appropriées. Plusieurs contraintes proviennent également des caractéristiques spécifiques du réseau physique. Par exemple, le découplage d'un wagon lors de l'arrêt à une station requiert la présence d'une voie d'évitement afin d'en permettre le garage jusqu'à ce qu'il soit couplé à un autre train. Finalement, d'autres éléments de planification tels que les possibilités de substitution doivent être pris en compte et compliquent encore davantage le problème.

Bien que le problème d'affectation des locomotives aux trains de marchandises partage certains traits communs avec le problème d'affectation des locomotives et des wagons aux trains de passagers, ce dernier possède donc des caractéristiques qui en font un problème plus difficile à résoudre et qui exigent une approche différente.

Puisque le problème d'affectation des locomotives et des wagons n'a été l'objet que de très peu de recherches, le premier objectif de cette thèse est de proposer un cadre de modélisation du problème qui en capte les difficultés fondamentales tout en possédant la flexibilité nécessaire pour l'adaptation à divers contextes pratiques. Ce cadre a

donc pour but de saisir les difficultés propres à la combinaison de différents types d'équipements présentant des incompatibilités et des interdépendances. Sa structure vise par ailleurs à permettre l'introduction de diverses contraintes et possibilités additionnelles relatives au fonctionnement d'un système de transport ferroviaire. Un second objectif, intimement lié au premier, est d'adapter différentes méthodes de décomposition pour résoudre les modèles proposés et de comparer leur performance. Plus précisément, cette comparaison vise les possibilités offertes par la relaxation lagrangienne, la décomposition de Dantzig-Wolfe et la décomposition de Benders. Une partie de ce travail consiste à évaluer dans quelle mesure ces méthodes peuvent s'adapter aux variations apportées aux modèles. Le dernier objectif de la thèse est de démontrer l'utilité pratique des modèles et des méthodes proposés. À cet effet, des tests sont réalisés à partir de données réelles fournies par une entreprise canadienne.

Au premier chapitre, nous présentons une revue détaillée de la littérature récente concernant l'emploi de modèles d'optimisation en transport ferroviaire. Cette revue déborde largement du cadre de l'affectation des locomotives et des wagons, et traite de la plupart des problèmes de planification et de contrôle rencontrés en transport par train de marchandises ou de passagers. Nous proposons une classification des différents modèles proposés dans la littérature et insistons plus particulièrement sur la structure de ces modèles ainsi que sur les méthodes utilisées pour les résoudre. Nous présentons d'abord les modèles de routage utilisés en transport de marchandises, suivis des modèles de fabrication d'horaires et d'affectation qui sont utilisés à la fois en transport de marchandises et en transport de passagers. Cette dernière catégorie inclut les problèmes d'affectation des locomotives et des wagons aux trains de marchandises et aux trains de passagers.

Au second chapitre, nous décrivons un modèle et une méthode de résolution développés en fonction des besoins spécifiques de l'entreprise canadienne VIA Rail. Ce modèle tient compte des très nombreuses caractéristiques du réseau ainsi que

des politiques de fonctionnement de l'entreprise. En particulier, il incorpore des contraintes d'entretien et des pénalités pour le couplage et le découplage de wagons qui compliquent considérablement le modèle. Le problème est résolu à l'aide d'une approche de génération de colonnes dans laquelle les colonnes correspondent à des chemins débutant à l'unique centre d'entretien, couvrant un certain nombre de trains, et se terminant au centre d'entretien dans les délais requis. La méthode de résolution consiste en une heuristique en deux phases qui permet d'alléger le modèle, au prix d'une certaine détérioration de la qualité de la solution.

Le troisième chapitre décrit un modèle simplifié pour lequel différentes méthodes de résolution exactes sont comparées. Puisque le modèle possède une structure très appropriée pour une décomposition primale des variables, nous présentons d'abord une approche de décomposition de Benders. Lorsque sont fixées la combinaison d'équipement utilisée sur chaque train ainsi que les séquences de trains qui seront couverts par le même équipement, le problème se décompose en des sous-problèmes de flot dans un réseau. Plusieurs concepts sont utilisés pour accélérer l'algorithme de résolution. En particulier, l'ajout à l'initialisation de contraintes valides au problème maître permet de réduire considérablement les temps de calcul et d'obtenir des solutions optimales en quelques minutes. Nous comparons également cette approche avec une relaxation lagrangienne et une décomposition de Dantzig-Wolfe.

Au dernier chapitre, nous décrivons finalement trois extensions importantes du modèle simplifié décrit au chapitre précédent. Nous considérons d'abord l'ajout des contraintes d'entretien. Ceci se fait en remplaçant les sous-problèmes de flot par des problèmes multi-flots. Ces derniers sont résolus par l'algorithme du simplexe ou par une décomposition de Dantzig-Wolfe. Nous considérons aussi l'ajout de pénalités pour limiter les modifications apportées aux trains durant les connexions entre deux services consécutifs. Finalement, le modèle étendu incorpore également la possibilité de substituer une pièce d'équipement à une autre. Ces deux dernières

extensions s'ajoutent très simplement au modèle par l'introduction de variables entières supplémentaires dans le problème maître. Un algorithme de résolution très efficace est obtenu en résolvant d'abord la relaxation du problème sans les contraintes d'entretien. De plus, des coupes de Benders non dominée sont générées en résolvant à chaque itération un problème auxiliaire. Cette approche permet de résoudre à l'optimalité des problèmes de grande taille avec tout l'éventail des contraintes présentes dans une application réelle.

Remarquons finalement que le problème d'affectation des locomotives et des wagons aux trains appartient à la classe des problèmes NP-difficiles dans le cas où plusieurs types de locomotives et plusieurs types de wagons sont utilisés. Il s'agit en effet d'une généralisation du problème d'affectation de véhicules à des itinéraires. Or, BERTOSSI *et al.* (1987) ont démontré que ce problème est NP-difficile dans le cas où le nombre de types de véhicules est supérieur à 1.

Chapitre 1

A Survey of Optimization Models for Train Routing and Scheduling

Jean-François Cordeau, Paolo Toth et Daniele Vigo, *Transportation Science* 32, pages 380–404, 1998.

Contrairement aux domaines du transport aérien et du transport routier qui ont été l'objet d'innombrables publications en recherche opérationnelle au cours des dernières décennies, le transport ferroviaire n'est parvenu à attirer l'attention des chercheurs que plus récemment. Plusieurs raisons peuvent expliquer ce constat. Tout d'abord, les problèmes pratiques rencontrés en transport ferroviaire sont généralement de très grande taille. Ensuite, les politiques de fonctionnement des transporteurs sont souvent difficiles à traduire en langage mathématique ou donnent lieu à des modèles dont la résolution est difficile. Enfin, la seule tâche de recueillir l'information nécessaire pour alimenter les modèles proposés requiert des systèmes de traitement de l'information dont peu d'entreprises disposaient par le passé. Heureusement, cette situation change rapidement et on observe depuis une dizaine d'années un intérêt croissant pour l'utilisation de la recherche opérationnelle dans l'espoir d'améliorer à la fois la rentabilité des entreprises et la qualité du service qu'elles offrent.

Cet article présente une revue de la littérature concernant l'utilisation de méthodes d'optimisation en transport ferroviaire. Nous décrivons d'abord brièvement les

processus de planification et de contrôle des opérations ferroviaires. Cette description fait ressortir les liens qui existent entre les différentes facettes de l'organisation et permet d'introduire une taxonomie des problèmes étudiés. Les modèles décrits dans l'article sont regroupés en deux grandes catégories: les modèles de routage et les modèles de fabrication d'horaires et d'affectation d'équipement.

Les problèmes de routage concernent exclusivement le transport de marchandises. Ils comprennent toutes les politiques de fonctionnement déterminant les étapes successives suivies par les wagons de marchandises, du chargement initial chez le client jusqu'à la livraison chez le destinataire. Plus précisément, ces politiques visent le routage de la marchandise dans le réseau, le regroupement des wagons pour la formation des trains, et le routage des trains eux-mêmes. Plusieurs modèles ont été proposés pour chaque catégorie de problèmes mais peu d'entre eux intègrent l'ensemble des politiques. De plus, seuls de rares modèles incorporent la dimension temporelle du problème. Finalement, ces problèmes sont intimement reliés au problème de la distribution des wagons vides. Pourtant, ce problème est généralement traité indépendamment en considérant comme données les différentes politiques de fonctionnement.

Les problèmes de fabrication d'horaires et d'affectation d'équipement concernent quant à eux la dimension temporelle de la planification et du contrôle des opérations. Parmi ceux-ci, on retrouve le problème de l'utilisation des voies ferroviaires. Puisqu'un grand nombre de trains doivent habituellement se partager les voies d'un réseau donné, une coordination précise du mouvement des trains est nécessaire afin de maximiser l'utilisation des voies et d'assurer la sécurité sur le réseau. Les problèmes d'affectation concernent par ailleurs l'utilisation du matériel roulant et, en particulier, l'affectation des locomotives et des wagons aux trains. Ces problèmes possèdent une dimension temporelle importante et leur solution fournit en fait un horaire d'utilisation de l'équipement à l'intérieur d'une période donnée.

La contribution de cet article est de tracer un portrait récent et complet de l'utilisation de la recherche opérationnelle en transport ferroviaire. En plus de proposer une classification des différents problèmes ayant été étudiés dans la littérature, il fournit une description détaillée des principaux modèles en insistant plus particulièrement sur leur structure et sur la méthode de résolution retenue. En somme, l'article et les très nombreuses références qu'il contient constituent un bon point de départ pour quiconque s'intéresse à l'application de la recherche opérationnelle au transport ferroviaire de marchandises ou de passagers.

A Survey of Optimization Models for Train Routing and Scheduling

JEAN-FRANÇOIS CORDEAU

École Polytechnique de Montréal

PAOLO TOTH and DANIELE VIGO

Università di Bologna

July 1998

Abstract

The aim of this paper is to present a survey of recent optimization models for the most commonly studied rail transportation problems. For each group of problems, we propose a classification of models and describe their important characteristics by focusing on model structure and algorithmic aspects. The review mainly concentrates on routing and scheduling problems since they represent the most important portion of the planning activities performed by railways. Routing models surveyed concern the operating policies for freight transportation and railcar fleet management, whereas scheduling models address the dispatching of trains and the assignment of locomotives and cars. A brief discussion of analytical yard and line models is also presented. The emphasis is on recent contributions, but several older yet important works are also cited.

1.1 Introduction

The rail transportation industry is very rich in terms of problems that can be modeled and solved using mathematical optimization techniques. However, the related literature has experienced a slow growth and, until recently, most contributions were dealing with simplified models or small instances failing to incorporate the characteristics of real-life applications. Previous surveys by ASSAD (1980b, 1981) and HAGHANI (1987) suggest that optimization models for rail transportation were not widely used in practice and that carriers often resorted to simulation. This situation is somewhat surprising given the considerable potential savings and performance improvements that may be realized through better resource utilization. It is also contrasting with the rapid penetration of optimization methods in other fields such as air transportation (YU, 1998).

In fact, the development of optimization models for train routing and scheduling was for a long time hindered by the large size and the high difficulty of the problems studied. Important computing capabilities were needed to solve the proposed models, and even the task of collecting and organizing the relevant data required installations that very few railroads could afford. As a result, practical implementations of optimization models often had a limited success, which deterred both researchers and practitioners from pursuing the effort.

In the last decade however, a growing body of advances concerning several aspects of rail freight and passenger transportation has appeared in the operations research literature. The strong competition facing rail carriers, the privatization of many national railroads, deregulation, and the ever increasing speed of computers all motivate the use of optimization models at various levels in the organization. In addition, recently proposed models tend to exhibit an increased level of realism

and to incorporate a larger variety of constraints and possibilities. In turn, this convergence of theoretical and practical standpoints results in a growing interest for optimization techniques. Hence, although simulation-based approaches are still widely used to evaluate and compare different scenarios, one witnesses a sustained development of optimization methods capable of producing high-quality solutions to complex problems within short computing times.

Problems facing rail transportation planners can be grouped into a number of classes according to the facet of the organization that is concerned. The most common approach is to represent the rail transportation system as a network whose nodes represent yards or stations and whose arcs represent lines of track on which trains carry passengers or freight. One then distinguishes between local problems involving only a node or an arc of the network, and global problems involving multiple entities. Rail transportation problems can also be classified into categories according to the planning horizon considered. At the strategic level, one is mainly concerned with the acquisition or construction of durable resources that will remain active over a long period of time. The tactical level is related to medium and short term issues, and generally involves the specification of operating policies that are updated every few months. Finally, the daily tasks that are performed by taking account of the fine detail of the system belong to the operational level. This popular hierarchical approach is explained in greater detail by ASSAD (1980a), who also gives numerous examples of problems that pertain to each category.

In this paper, we intend to review most of the recent contributions dealing with train routing and scheduling with regard to both freight and passenger transportation. We will thus cover all three levels of planning but focus our attention on global problems of train management. Because of the large size and the high degree of heterogeneity that characterize most models, we have opted for a textual description.

A more involved comparison of mathematical formulations would require focusing on a much smaller subset of models.

Most reviewed models have been proposed during the last decade although we also cite several older but important works. Apart from a few exceptions, the survey concentrates on published and easily accessible material. We have also elected to limit ourselves to contributions dealing specifically with rail transportation, even though a lot of work done in the related areas of road and air cargo transportation is certainly relevant to the rail context. Finally, the field of railway crew management will not be treated here but we instead refer the interested reader to recent work by CAPRARA *et al.* (1997).

The paper is organized as follows. Section 1.2 introduces the necessary background and definitions concerning the reviewed material. Models for train routing and train scheduling are reviewed in Sections 1.3 and 1.4, respectively. Conclusions and an account of current research trends are presented in the last section.

1.2 Background and Definitions

We now give a brief description of railroads and introduce some terminology that will be used throughout the text. A more detailed account of rail operations and freight transportation is presented in the book by BECKMANN *et al.* (1956). The authors also provide an interesting introduction to rail modeling and optimization.

The first part of the review is devoted to routing problems in the context of rail freight transportation. Demand for freight transportation is usually expressed in terms of tonnage of certain commodities to be moved from an origin to a destination.

Given these demands, the railroad must establish a set of operating policies that will govern the routing of trains and freight.

For every origin-destination pair of traffic demand, the corresponding freight may be shipped either directly or indirectly. When demand is important enough, delivery delays are obviously minimized by using direct trains as opposed to sending the traffic through a sequence of links. However, when demand does not warrant the dispatching of direct trains, delays are inevitable. Either the traffic is consolidated and routed through intermediate nodes, or freight cars have to wait at the origin node until sufficient tonnage has been accumulated.

To benefit from economies of scale, trains are thus often formed by grouping cars with various commodities and having different origins and destinations. These trains operate between particular nodes of the network, called *classification yards*. At these yards, cars are separated, sorted according to their final destination, and combined to form new outbound trains. However, because the classification process requires considerable resources, cars are not reclassified at every yard on their trip from origin to destination. Instead, cars with different final destinations but sharing some initial portion of their trips are assembled into *blocks*. Cars in the same block may then pass through a series of intermediate classification yards, being separated and reclassified only after they have reached the destination of the block. The *blocking policy* specifies what blocks should be built at each yard of the network and which cars should go into each block.

In each yard, blocks are built on *classification tracks* where they await the departure of an outbound train. The list of potential blocks that may go into each outbound train is specified by the *makeup policy*. Also, when a train passes through an intermediate classification yard, it may leave or pick up blocks of cars. A block left by an inbound train is either transfered to a different train or it is broken up and

its cars are reclassified. Hence, although the origin and destination of a block may correspond to those of a train, a block may also switch trains several times before reaching its final destination.

Every loaded movement on a rail network leads to a supply of empty cars at destination. Therefore, if transportation demand is unbalanced, steps must be taken to reposition empty cars and avoid their accumulation in some parts of the network where more traffic is directed. Even if traffic is balanced in the long run, this need not be the case in the short term. Repositioning empty freight cars can thus help the railroad offer better service to its customers by reducing the average time they have to wait for cars, and decrease the capital investment associated with equipment ownership. The *freight car management problem* consists of dynamically distributing empty cars in the network to improve the railroad's ability to promptly answer requests for empty cars while minimizing the costs associated with their movement.

The second part of the survey discusses models that deal with the temporal dimension of train management. Scheduling problems appear in both freight and passenger transport, albeit in slightly different forms. In the case of freight transportation, trains sometimes operate without schedules and simply depart when they have accumulated sufficient tonnage. Although this practice is still very common in North America, it is seldom seen in Europe where freight trains usually operate according to published schedules just as they do in the case of passenger transportation. When freight trains do not operate according to a schedule, potential time slots must still be assigned to them.

Although train timetabling is usually performed at the tactical level of planning, real-time operations necessitate precise synchronization of freight and passenger train movements on the lines of the physical railway network. The lines can be made of a single track, as is often the case in North America and in most developing countries,

or may contain two or more tracks, as is common in Europe. To allow trains traveling in different directions on a single-track line to meet, *sidings* are located at regular intervals along the line. These short track sections allow one train to pull-over and free the way for the other one. Sidings are also used to permit a fast train to pass a slower one. Given a train timetable, the *train dispatching problem* determines a feasible plan of meets and overtakes that satisfies a system of constraints on the operation of trains.

Finally, a related scheduling problem concerns the use of the rolling equipment stock. Because of the high capital expenditures associated with locomotives, a major concern to every railway is to maximize the use of these resources. The basic *locomotive assignment problem* consists of assigning a set of locomotives to cover all scheduled trains at minimum cost while satisfying some side constraints such as compatibility restrictions and maintenance requirements. Although freight trains generally contain a large number of cars and several engines, passenger trains use a small number of cars coupled with a few locomotives. In the case of passenger trains, it is thus possible to perform the simultaneous assignment of both types of equipment to the trains.

1.3 Routing Problems

Operating plans for rail freight transportation indicate the train connections to be provided, the blocks to be built in each yard, and the assignment of blocks to trains. In addition, train timetables must be developed to specify the departure and arrival times of trains. These closely intertwined policies should ideally be determined concurrently to identify the most efficient way of delivering all traffic while satisfying a set of technological constraints on train and yard capacity. However, because this

leads to a very difficult problem, a sequential approach is often adopted. For example, a blocking plan may be developed first, followed by a train routing and makeup plan. Very frequently, train timetables are specified last and are designed around the routing plans. Operating plans are usually updated every few months but weekly or daily adjustments must be made to account for demand variability.

Most optimization models for train and freight routing are defined over a network whose nodes represent origins, destinations or intermediate transfer points for the traffic to be routed. The arcs then represent existing or possible train connections between these points that are often aggregated to represent the activities of a wider geographic area.

Because yard activities constitute an important part of freight transportation operations, we first present a brief review of analytical models developed to analyze yard performance under different configurations or traffic conditions. Although these are not optimization models per se, they may appear within the objective or constraint structure of large-scale routing models. We then present network models that address the blocking, makeup and routing problems. Models aimed specifically at the freight car management process are described in the last section. In each section, models are presented in ascending chronological order.

1.3.1 Analytical Yard Models

Yard policies concern the specification of the activities to be performed in the yards of a rail network. More precisely, they indicate how trains entering each yard should be inspected and disassembled, and how cars should be sorted and reassembled into blocks that will form new outbound trains. Although reclassification work is also performed to some extent in less-than-truckload and air cargo transportation, the

delays associated with these activities are usually negligible for these modes while they constitute a large portion of the overall transit time for rail freight. As explained by KEATON (1989), car time in intermediate terminals occurs in classification and assembly operations and while waiting for the departure of an outbound train, but also as a result of yard congestion. Car time is also spent in origin and destination terminals where cars wait either for the departure of an outbound train or for delivery to the receiver by a local train.

Two types of classification yards are in common use. *Flat yards* use engines to move cars from an inbound train to classification tracks. In *hump yards*, this work is performed by gravity: cars detached from an inbound train are pushed over the top of a hump and roll down to the appropriate track. Early work on yard modeling was realized by CRANE *et al.* (1955) who presented an analysis of a particular hump yard and discussed the queuing processes identified in inspection and classification operations. A simple model for the location of a classification yard was then proposed by MANSFIELD and WEIN (1958).

A more detailed analysis of railyard operations was performed by PETERSEN (1977a,b) who developed queuing models to represent the classification of incoming traffic and the assembly of outbound trains. In these queuing models, the basic units of arrival are complete trains to be processed. The author also modeled the delay to a railcar from the end of classification to the start of the train assembly operation with a bulk queue, and observed that this delay is a minor source of yard congestion in comparison with classification and assembly operations. The models are used to compute the probability distribution of connection times for various levels of traffic given known service times. In the second paper, expressions are derived to relate the classification and assembly times to the physical characteristics of the yard and traffic attributes. The accuracy of the models was validated using historic data from two

railroads. An insightful description of railyards is also presented in the first of these papers.

TURNQUIST and DASKIN (1982) modeled yard operations from the perspective of freight cars, rather than from the perspective of trains. They thus developed queuing models for classification and connection delays that consider individual cars as the basic units of arrival. Their approach also differs from that of PETERSEN in the sense that connection to an outbound train and assembly are treated as a single operation. Expressions for the mean and variance of classification and connection delays are derived under the assumption of Poisson arrivals using a batch-arrival and a batch-service queuing model, respectively. The authors also demonstrated how their model may be used to evaluate the effects of train dispatching strategies on the mean and variance of delay. In particular, they analyzed two strategies that consist, respectively, of scheduling trains at regular intervals, and dispatching trains when a given number of cars become available.

A different approach to the problem of predicting yard time distributions was studied by MARTLAND (1982) who described a methodology for estimating the total connection time of cars passing through a classification yard. The model is based on a function, calibrated using actual data from the railroad, that relates the probability of making a particular train connection to the time available to make that connection and other variables such as traffic priority and volume. The function can be adjusted through different techniques such as regression analysis or simulation experiments. The approach, which has been tested and implemented by several railroads, is proposed as an aid to planning but also as a way to control operations by setting standards for train connection performance.

Other analytical models concern the performance and resource requirements of sorting strategies that specify what blocks should be assigned to each available

classification track and how individual cars should be handled. Early work on this topic was performed by SIDDIQEE (1972) who compared four sorting and train formation schemes in a railroad hump yard. A screening technique and a dynamic programming approach were suggested by YAGAR *et al.* (1983) to optimize humping and assembly operations.

DAGANZO *et al.* (1983) investigated the relative performance of different multi-stage sorting strategies. In multi-stage sorting, several blocks are assigned to each classification track, and cars must be resorted during train formation. Equations are derived for the service time per car of triangular sorting in both flat yards and hump yards. In a series of three papers, different classification strategies were also analyzed and compared by DAGANZO (1986, 1987a,b), who gave expressions for the switching work and space requirements. In the last two papers, the author considered dynamic blocking in which the assignment of blocks to classification tracks is allowed to vary through time.

Finally, AVRAMOVIĆ (1995) modeled the physical process of cars moving down the hump of a yard. This process is represented by a system of differential equations that incorporate several factors, such as hump profile and rolling resistance, affecting the movement of a car. The model can be used in the design of a hump yard to evaluate the strength of track retarders that regulate the speed of cars.

1.3.2 Network Routing Models

We now discuss network optimization models that address different problems related to freight train routing. We first review models dealing with the blocking policy, followed by models addressing the train routing and makeup problem. Compound

models that integrate blocking, makeup, and scheduling decisions are discussed last. The characteristics of the most important contributions are summarized in Table 1.1.

Table 1.1: Characteristics of network routing models

Authors	Problem type	Planning horizon	Objective function	Model structure	Solution approach
Bodin et al. (1980)	Blocking	Tactical	Min operating and delay costs	Nonlinear MIP	Heuristic
Assad (1983)	Blocking	Operational	Min total classification	Shortest path	Dynamic programming
Van Dyke (1986)	Blocking	Tactical	Min operating costs	Shortest path	Heuristic
Newton (1996)	Blocking	Tactical	Min operating costs	NDP with node budget	Dantzig-Wolfe decomposition
Crainic et al. (1984)	Routing/ makeup	Tactical	Min operating and delay costs	Nonlinear MIP	Heuristic decomposition
Haghani (1989)	Routing/ makeup	Operational	Min operating and delay costs	Nonlinear MIP	Heuristic decomposition
Keaton (1989)	Routing/ makeup	Tactical	Min operating and time costs	Linear MIP	Lagrangian relaxation
Keaton (1992)	Routing/ makeup	Tactical	Min operating and time costs	Linear 0-1 IP	Lagrangian relaxation
Martinelli and Teng (1996)	Routing/ makeup	Tactical	Min transit time	Nonlinear 0-1 IP	Neural networks
Marín and Salmerón (1996)	Routing/ makeup	Tactical	Min operating costs	Nonlinear IP	Local search heuristics
Morlok and Peterson (1970)	Compound	Tactical	Min operating and time costs	Linear MIP	Branch-and-bound
Huntley et al. (1995)	Compound	Tactical	Min operating costs	Nonlinear MIP	Simulated annealing
Gorman (1998)	Compound	Tactical	Min operating costs	Linear 0-1 IP	Genetic search

Blocking models

A blocking policy is usually specified as follows: cars at yard i which are destined for yard j must be added to a block that will next be shipped to yard k (possibly transiting by other intermediate yards). As explained in the introduction, cars in a block will not be reclassified until the block reaches its final destination. A blocking model thus places the emphasis on the movement of cars as opposed to the movement of trains. Its solution indicates the routing of freight through the network and the distribution of classification work among yards, but does not specify the trains to be

run or the assignment of blocks to trains. Instead, an additional problem must then be solved to determine the routing of trains and their makeup.

One of the first models for car blocking belongs to BODIN *et al.* (1980), who suggested a nonlinear, mixed integer programming formulation of the problem. The model, which is a multi-commodity flow problem with additional side constraints, simultaneously determines the optimal blocking strategies for all the classification yards in a railroad system. Besides flow equations that constitute the backbone of the model, yard capacity and block formation constraints are also considered. In particular, the model imposes upper bounds on the number of cars that may be classified and the number of blocks that may be formed in any given yard. This last constraint originates from the fact that each yard has a limited number of tracks on which blocks may be built. Block length constraints are also taken into consideration and guarantee that the number of cars in each block lies between a lower and an upper bound. Finally, *pure strategy* constraints are present. These constraints ensure that all cars in yard i destined for yard j are shipped to the same next classification yard. The objective function considered seeks to minimize the sum of shipping, processing, and delay costs. Delay costs are represented by piecewise linear functions of the flow on arcs of the network. With some manual intervention, the authors solved an instance with 33 classification yards and found a solution within 3% of a tight lower bound.

ASSAD (1983) proposed a solution approach for a problem defined on a line network composed of n yards, with traffic flowing from yard 1 to yard n. Cars are received at yard 1 in arbitrary order and must be separated as they proceed along the line to allow each successive yard to extract the traffic destined for it. Various classification strategies can be used to distribute the classification work among the yards. For the special case in which all yards have equal traffic, the author showed that the search for a solution minimizing the total work can be restricted to strategies

in which traffic for yard i is separated only after previous traffic types $1, \ldots, i-1$ are already classified. When this assumption does not hold, a dynamic programming formulation of the problem leads to an efficient solution method. The author also discussed extensions to the case in which each yard is a potential source of traffic. It is shown that a dynamic programming formulation can still be used for this problem.

VAN DYKE (1986, 1988) described a heuristic blocking approach that has been tested or implemented by several large railroads. The system is based on an iterative procedure that attempts to improve an existing blocking plan by solving a series of shortest-path problems on a network whose arcs represent available blocks. Traffic is assigned to a particular block if the block is on the least cost path from the origin of the traffic to its destination. The cost of assigning traffic to a block depends on a number of factors such as block priority, traffic priority, physical rail lines traversed, and the characteristics of the origin and destination yards of the block. The solution to these problems determines the least cost distribution of traffic across a set of existing blocks. An interactive procedure allows the user to delete existing blocks or introduce additional blocks in the solution. Block capacity constraints are also taken into account by the heuristic.

Recently, a column generation algorithm was introduced by NEWTON (1996), who studied the more general network design problem (NDP) with budget constraints. This problem consists of minimizing the cost of flowing a set of commodities through a network while satisfying budget constraints on the fixed cost of the arcs used. The railroad blocking problem is transformed into this general framework by letting the nodes represent the classification yards and the arcs represent potential blocks that can be built. The fixed cost of offering direct service between two yards involves dedicating a sorting track at the origin yard. Hence, there is a separate node-budget constraint for each yard based on the number of sorting tracks available. Flow constraints are also used to restrict the total number of cars that may be

sorted in each yard. The objective function minimizes the cost of delivering all commodities. Express and non-express traffics are treated simultaneously using priority constraints that limit the number of blocks used in delivering each commodity. The problem is solved using a branch-and-bound procedure with bounds computed at each node using Dantzig-Wolfe decomposition (DANTZIG and WOLFE, 1960). Using a labeling algorithm on an acyclic network, blocking paths with a negative reduced cost are generated for each commodity by solving a shortest path problem with a priority constraint. A rounding heuristic is also used to obtain good upper bounds. Disaggregating the bundle constraints that impose common upper bounds on the arcs of the network gives valid inequalities that strengthen the LP relaxation of the master problem. Branching is performed on the binary variables indicating whether an arc is chosen or not. Computational results were presented for instances with 150 nodes, 6000 potential arcs and 1300 commodities. Feasible solutions within a few percent of a known lower bound were found within a few hours on a workstation computer.

Routing and makeup models

Whereas blocking models indicate the routing of freight and the distribution of classification work among the yards of the network, routing and makeup models determine the routing and frequency of trains and the assignment of blocks to trains. In routing and makeup models, the blocking policy may be either determined endogenously or given as an input. These models thus produce a complete train and freight routing plan. However, because they do not provide actual departure times for the trains to be run, an additional scheduling problem must be solved at a later stage. Similar models for the service network design problem in the motor carrier industry were developed, for example, by POWELL and SHEFFI (1989).

Train formation plans are sometimes developed without regard to the concept of car blocking. For example, THOMET (1971) developed a cancelation procedure that gradually replaces direct shipments by a series of intermediate train connections to minimize operation and delay costs. A model for deciding which pairs of yards should be offered direct service to minimize total transit time of cars was also proposed by SUZUKI (1973), whereas LEBLANC (1976) suggested a network design model for strategic planning. One of the first efforts to integrate multiple components of the freight routing problem is credited to ASSAD (1980a) who proposed a multi-commodity network flow model for train routing and makeup that incorporates some level of interaction between routing and yard activities.

A more complex problem was studied by CRAINIC *et al.* (1984) who proposed a model and a heuristic for tactical planning. The model is a nonlinear, mixed integer, multi-commodity flow problem that deals with the interactions between blocking, makeup, and train and traffic routing decisions. Traffic demand is divided into classes in which each class corresponds to an origin-destination pair, together with a commodity type. The model is based on a service network that specifies the feasible routes on which train services may be run. A set of feasible itineraries is defined for each traffic class. An itinerary specifies the train service path followed and the operations that must be performed at each intermediate stop. By selecting the best traffic distribution for each traffic class, one solves the freight routing problem as well as the blocking and makeup problems. The frequency variables associated with the possible train services provide a solution to the train routing problem. The objective function seeks to minimize the sum of operation and delay costs associated with itineraries and train services. By introducing the train service capacity constraints in the objective function, the authors obtain a modified problem for which they use a decomposition scheme that iterates between two problems until the improvement in the objective function after a complete iteration is less than a preset value. The

subproblem determines the best traffic distribution for each traffic class for a given service level, whereas the master problem modifies the service frequencies to improve the solution value considering the given traffic distribution. The subproblem for each traffic class is solved using column generation and a descent algorithm. This solution methodology was explained in greater detail by CRAINIC and ROUSSEAU (1986), who presented a general framework for the design of the service network and the routing of traffic in the context of multi-commodity, multi-mode freight transportation. The model and algorithm were tested on data from the Canadian National Railroads. The instance contained 2613 aggregated traffic classes and a service network with 415 links. Computational results indicated a significant cost reduction over the solution used by the railroad. A comparison with the simulation method used by the company was done by CRAINIC (1984). Readers interested in strategic planning are also referred to the work of CRAINIC *et al.* (1990a).

As was properly highlighted by HAGHANI (1987), there exist intense interactions among the routing of trains, their makeup, their frequency, and the empty car distribution process. However, models that take all these aspects of rail transportation into consideration often get extremely complex if not simply intractable. The traditional approach has thus been to deal separately with the train routing and makeup problem and the empty car distribution problem. This obviously leads to suboptimal decisions, at both the tactical and operational levels. In an effort to counter the tendency of treating the empty car distribution problem at the operational level by assuming that routing and makeup decisions are given, HAGHANI (1989) proposed a formulation and a solution method for a combined train routing and makeup, and empty car distribution problem. The model is also dynamic and deals with temporal demand variability, providing empty car distribution decisions as well as the optimal time interval between consecutive train services between pairs of yards. To account for demand variations from period to period, each yard is replicated a

certain number of times in a time-space network, depending on the period length and the horizon considered. This network has nodes representing inbound and outbound traffic for every yard in the physical network, and links representing routing, classification, delays, and deliveries. The decision variables used concern the flows of loaded cars, empty cars, and engines provided on the different links mentioned. The objective considered is to minimize the total cost defined by routing costs, classification costs, delay costs for classification and connection, and penalty costs. Penalties are imposed for carrying over the demand for empty cars and as a way to deal with boundary conditions on the shipments. Besides traditional flow conservation constraints on the loaded cars, empty cars, and engines, linking constraints ensure that the number of engines provided on each link is compatible with car routing decisions. This mixed integer model has a nonlinear objective function and linear constraints. It is solved with a heuristic decomposition approach that exploits the structure of the problem by solving an integer programming subproblem for the engine flow variables and a linear programming subproblem for the car flow variables. The algorithm was tested on a network with four nodes and five two-way links. On average, the solutions found by the heuristic were within 10% of the lower bound provided by the LP relaxation of the problem.

KEATON (1989) proposed a model and a heuristic method based on Lagrangian relaxation for the combined problem of car blocking and train routing and makeup. The model is based on a set of service networks that specify the possible train connections and blocking alternatives for each origin-destination pair. Upper limits are imposed on the number of blocks that can be formed at any terminal and on the number of cars assigned to any train. The objective function considers train costs, car time costs, and classification costs. The mixed integer programming model uses integer variables for train connections and continuous variables for car flows. By dualizing the constraints that link train variables and car flow variables into the

objective function, one obtains a series of shortest path problems in the continuous variables and knapsack problems in the train variables. When ignoring train size constraints, the model can be solved efficiently with sub-gradient optimization and special update rules for the multipliers. Feasible solutions are improved by using a dual adjustment procedure and a greedy heuristic. A hypothetical rail network was used to generate an instance with 26 terminals and 333 origin-destination pairs. On average, solutions with duality gaps below 10% were obtained. However, when limits on train size are imposed, it becomes very hard to obtain tight lower bounds on the solution values. This model was used by KEATON (1991) to evaluate service-cost tradeoffs for carload freight traffic in the U.S. rail industry. He applied his formulation and solution method to hypothetical rail networks with variable train costs and concluded that the potential for reducing transit times by increasing train connections and frequency was rather limited.

In a subsequent paper by KEATON (1992), pure strategy constraints for blocking and maximum transit times for each origin-destination pair are also considered. The resulting formulation has only binary variables and results in a multi-commodity network flow problem once the train variables are set. By dualizing the linking constraints between train and car flow variables, and constraints that place limits on train size and maximum transit time, the formulation decomposes into two easily solvable subproblems. In fact, a further relaxation is obtained by discarding all constraints on train size, yard volumes, and service levels, and dualizing the linking constraints between train and car flow variables. This relaxation can be solved efficiently using a dual adjustment procedure, and tight lower bounds can be generated. By iteratively solving this relaxation and adjusting the car or train costs in each iteration, a feasible solution to the original problem is finally obtained. However, this approach, called *iterative strategy*, does not yield explicit lower bounds on the cost of the original problem, and thus the quality of the solution obtained cannot be

evaluated precisely. Computational experiments were performed on a set of three rail systems containing about 80 terminals and 1300 to 1500 origin-destination pairs.

Neural networks were used by MARTINELLI and TENG (1996) to solve a train formation problem. For a given distribution of demand, expressed as the number of cars to be moved between each origin-destination pair, the problem is to assign each class of demand to a unique itinerary chosen from a predefined subset. An itinerary specifies a succession of intermediate yards together with the train sequence used. The problem is formulated as a 0-1 integer program with a nonlinear objective function that minimizes the total time spent by cars in the system. A back-propagation neural network model trained with two groups of patterns was used to solve small instances of the problem. Good performance was obtained, as measured by the quality of the solutions, but the computation times were rather long. The data used contained 30 demand classes, 44 trains, and 108 combinations of demand-train assignments.

In a series of two papers, MARÍN and SALMERÓN (1996a,b) proposed and analyzed the expected performance of local search heuristics for the tactical planning of rail freight networks. Again, the model is based on a service network and considers demands given in terms of origin and destination yards and freight type. Each train service is defined by an origin yard, a set of intermediate yards, a destination yard, and technical characteristics such as speed and capacity. The objective is to minimize car costs, train costs, and investment costs incurred when not enough trains are available. This last term, which uses a crude approximation of the required fleet size, makes the objective function piecewise linear. Because each train service specifies the set of intermediate stations, restrictions on the number of cars transiting in any yard can be imposed. Constraints are also imposed on the number of cars assigned to each service given the chosen service frequency. The three heuristic methods proposed (descent method, simulated annealing, and tabu search) share a common decomposition that separates the routing of the freight cars and the choice of train service frequencies. The

first subproblem, which is solved through a sequential loading algorithm, determines the best routes for a given choice of train frequencies. The second subproblem, which may be solved by inspection, readjusts the train frequencies for the given car routing. In each iteration of the various heuristics, train frequencies are updated according to a move that is chosen from the neighborhood of the current solution, and the car routing subproblem is solved. A reformulation of the problem as a linear program leads to an exact branch-and-bound algorithm that can be used for comparison purposes with the heuristics. Computational tests on four generated networks showed that simulated annealing obtained the best solutions but required more time than the other heuristics. This conclusion was also confirmed by the statistical analysis conducted in the second paper. The largest instance solved contained 82 train services and 150 demand classes.

Compound routing and scheduling models

Routing and makeup models produce a transportation plan that completely describes the routing of freight, the set of trains to be operated and their respective frequency. But because these models do not take scheduling into consideration, it may be difficult to later find a timetable accommodating all planned trains and satisfying line and yard capacity. Hence, compound models, which address both the routing and the scheduling aspects of freight transportation, can significantly help to improve service reliability and reduce costs. The recent work of FARVOLDEN and POWELL (1994) described a similar approach for the motor carrier industry. Also, railroad revenue management models based on profit maximizing and load selection formulations were introduced by CAMPBELL (1996) and KRAFT (1998).

One of the first efforts to integrate both routing and scheduling decisions into a single optimization model is probably the work of MORLOK and PETERSON (1970).

Given a network representing the possible train connections, a binary variable is associated with each train service that may be operated. Each such service is defined by a route in the network, a set of stops, a departure time at the initial node, and additional attributes such as speed and capacity. A second set of binary variables is used to represent the assignment of demand to trains. Additional variables are also introduced to keep track of car time in the network. The costs considered include train and engine crew costs, intermediate yard costs, and car time costs. Besides traditional demand constraints, the model incorporates constraints on the maximum number of cars per train as well as scheduling constraints requiring that certain cars be delivered to given yards before a cut-off time. The model was applied to a very small instance and solved with a branch-and-bound procedure.

A computerized routing and scheduling system was developed by HUNTLEY *et al.* (1995) to help planners at CSX Transportation account for the effects of routing and scheduling decisions in strategic planning. Demand is represented as *batches* that have associated origin and destination yards. Each pair of switching yards in the network defines a link that may accommodate a certain number of trains. The output of the model is the sequence of train links that each batch should follow from origin to destination, as well as the departure times for all train links. The nonlinear objective function minimizes operational costs defined by fuel cost, crew cost, locomotive capital cost, and freight car rental cost. The problem is solved using simulated annealing and a perturbation operator that inserts or deletes a stop from the route of a batch, and adjusts the departure times of the trains. The system was tested on a real problem involving 166 batches and 41 yards. Related field testing showed that the system was useful in analyzing a variety of scenarios, and produced schedules having similar properties to those of the solutions in use by the company, but a smaller cost.

A combination of genetic and tabu search algorithms were used by GORMAN (1998) to address the weekly routing and scheduling problem. To solve the problem for actual train departure times, the time horizon is discretized in hours. Each train may also operate at different speeds and perform a variable sequence of stops on its way from origin to destination. The mathematical formulation has binary variables associated with each potential train service that may be operated during the week. Each possible assignment of demand to a train is also represented by a binary variable. Constraints are imposed on train size to ensure that trains operate on schedule. There are also linking constraints to enforce yard and line capacity. The objective function minimizes the sum of fixed costs of trains and marginal cost per car. The model decomposes into train-scheduling and traffic-assignment components. To solve the problem, the author suggested a classical genetic search procedure in which the population is formed by all possible train schedules. Every time an individual is generated, its cost is evaluated by solving the traffic-assignment problem. Mutations are obtained by either adding or deleting a train, or by shifting a train to an earlier or a later time in the schedule. To improve the performance of the genetic algorithm, each solution is cloned and modified with a tabu search algorithm, thus simulating the use of knowledge-based mutation operators. Computational experiments on data from a major U.S. freight railroad produced solutions that satisfied more constraints and had a smaller cost than the solution actually used by the railroad.

1.3.3 Freight Car Management Models

The utilization cycle of a freight car starts when a client issues an order for empty cars. At a nearby yard, compatible cars are selected and moved to a loading point. Once loaded, they are taken to a classification yard where they are sorted, assembled into blocks, and put onto outbound trains. When a car has reached its final destination,

it is unloaded and, unless it is needed by the receiver, it is returned to the railroad. At this point, the car is available for a new shipment and the cycle may repeat. Very often, however, it will travel empty to a different location where a request must be fulfilled. Because demand for transportation is rarely known long in advance, the railroad must anticipate future requests and manage its fleet accordingly. A good repositioning strategy helps to reduce the size of the fleet and to decrease the delays in delivering empty cars to customers.

Models for fleet management and distribution of empty vehicles were reviewed by DEJAX and CRAINIC (1987). The management of empty railcars shares several characteristics with the distribution of empty containers used in land, maritime, or multimode transportation. Dynamic and stochastic models for the land distribution of empty containers were developed by CRAINIC *et al.* (1990b, 1993). Also, recent work on operations planning in intermodal transportation was performed by NOZICK and MORLOK (1997). Finally, the related problem of dynamic vehicle allocation was initially studied by POWELL (1986, 1987) and later developments have been summarized by POWELL *et al.* (1995).

We now review optimization models for the distribution of empty rail cars. We first discuss models used in the case of a single railroad, followed by models for the case of multiple railroads sharing a fleet of cars under a pooling agreement. The characteristics of the most recent models in each category are summarized in Table 1.2.

Single railroad models

In the first attempts to optimize the distribution of empty freight cars, the process was often represented as a simple network flow problem for which efficient algorithms were available. WHITE and BOMBERAULT (1969) generated a network from a time-

Table 1.2: Characteristics of freight car management models

Authors	Problem type	Planning horizon	Objective function	Model structure	Solution approach
BEAUJON and TURNQUIST (1991)	Single railroad	Tactical	Max expected profits	Nonlinear network	Frank-Wolfe
MORIN (1993)	Single railroad	Operational	Min operating costs	Multi-commodity	Decomposition
SPIECKERMANN and VOSS (1995)	Single railroad	Operational	Min transport costs	Job-shop scheduling	Greedy heuristic
HOLMBERG et al. (1996)	Single railroad	Operational	Min transport and shortage costs	Multi-commodity	Branch-and-bound
ADAMIDOU et al. (1993)	Multiple railroads	Tactical	Max profits	Nash equilibrium	Gauss-Seidel
SHERALI and TUNCBILEK (1997)	Multiple railroads	Strategic	Min fleet size	Network	Heuristic decomposition

space diagram and solved the resulting transshipment problem with a modified out-of-kilter algorithm (FORD and FULKERSON, 1962). Also, static formulations solvable as transportation problems were proposed by ALLMAN (1972) and MISRA (1972). HERREN (1973, 1977) formulated a more complex problem, with a heterogeneous fleet of cars and substitution possibilities, as a minimum cost network flow model that could be solved with a specialized algorithm.

A different approach to freight car management consists of representing the system with an inventory model. One of the first efforts in this direction is the work of AVI-ITZHAK *et al.* (1967) who suggested mathematical models for describing the behavior of car pool systems. PHILIP and SUSSMAN (1977) proposed a discrete event simulation model to determine the optimum inventory level for a single terminal. The inventory management approach was later extended to an entire network by MENDIRATTA and TURNQUIST (1982) who developed a linear programming formulation solvable by a decomposition algorithm.

One of the first contributions dealing with the stochastic nature of the problem is from JORDAN and TURNQUIST (1983), who presented a dynamic network optimization model, based on earlier work by COOPER and LEBLANC (1977), that

takes into account variability in empty car demand and supply, as well as uncertainty in travel times. A methodology based on a combination of linear programming and simulation techniques was proposed by RATCLIFFE *et al.* (1984) to optimize freight car dispatching given known and anticipated demands. Also, a real-life application of linear programming techniques to the daily distribution problem was presented by MARKOWICZ and TURNQUIST (1990).

A combined model for fleet sizing and vehicle distribution and use was described by BEAUJON and TURNQUIST (1991). Their approach takes into account the dynamic nature of these decisions as well as the uncertainty in demand and transit times. They first proposed an exact formulation which can be viewed as a stochastic programming problem or as a stochastic control problem. Because this formulation appears computationally unattractive, a solution method was developed for an approximate reformulation of the problem. The reformulation replaces random variables associated with transportation demand and travel times by their expected value to obtain a network optimization model. The objective function maximizes the expected profit which is defined by the difference between revenues generated by serving demands and costs incurred for vehicle ownership, vehicle movement, and unmet demand. To appropriately model the cost structure of the problem, the concept of *net vehicle pool* is introduced. At each terminal, this quantity represents both the expected vehicle pool and the expected vehicle shortage. Nonlinear costs on the arcs are then used to account for vehicle holding and unmet demand. Because the random travel times are replaced by their expectation, the network approximation introduces an error in representing vehicle arrivals. The solution procedure presented tries to circumvent this weakness by solving a pure network formulation to determine empty vehicle dispatching decisions, and adjusting the size of the net vehicle pools to account for this approximation error by solving a series of unconstrained optimization problems. The nonlinear objective includes functions of the basic decision variables that are neither

convex nor concave because of variance terms. Hence, the network flow problems are solved using a procedure that iteratively fixes the variance terms, solves the resulting concave problems using the Frank-Wolfe algorithm (FRANK and WOLFE, 1956) and updates the variance terms. Numerical experiments performed on instances with up to 70 nodes and 1330 arcs showed that significant improvements are obtained by considering the stochastic nature of the problem.

Decomposition approaches were compared by MORIN (1993) who studied the empty car distribution process at SNCF and formulated the problem as a multi-commodity network flow problem. Each commodity corresponds to a geographical area, and linking constraints ensure flow conservation between adjacent areas. Two formulations that can be solved with sub-gradient algorithms were introduced: a dual decomposition approach that relaxes the linking constraints and a primal decomposition scheme that relies on the introduction of coupling variables. The application of a mixed decomposition approach (MAHEY, 1986) that combines price-directive and resource-directive allocations was also presented with a specialized algorithm that exploits the separability of the problem. Results on a set of data from SNCF indicated that the third method was superior.

SPIECKERMANN and VOSS (1995) formulated the empty railcar distribution problem as a scheduling problem with machines representing railcars and jobs representing requests for cars. The study is realized in the context of a German car rental company that provides empty cars to its customers throughout Europe. All movements are performed by national railways to which the company must pay fees for movements of either loaded or empty cars. The objective of minimizing costs for empty moves translates into minimizing the time-dependent setup costs. The model is solved using a three-stage procedure that is embedded into a greedy heuristic. The first stage finds a feasible solution using the earliest-due-date (EDD) rule. The second stage then tries to improve this solution with respect to an objective of minimizing

the total tardiness in filling the orders. An improvement procedure that tries to reduce the transport costs without increasing tardiness is used last. The algorithm was tested on real data from the company and on randomly generated instances. The largest instance contained 805 requests, 225 railcars, and 205 stations. The system yields a significant cost reduction but computing times exceed several hours in some experiments.

HOLMBERG *et al.* (1998) proposed a multi-commodity network flow model for operational distribution of empty cars. Each commodity corresponds to a type of car, and linking constraints impose limits on the total number of empty cars that may be part of each scheduled train. Train movements are represented on a time-space network. The objective of the model is to minimize transportation and car shortage costs. The value of having a car in inventory at a given terminal after the planning period is also taken into account. A multi-period planning horizon is considered and the operational model is solved using a sliding horizon framework in which decisions associated with the initial segment of the period are implemented whereas the others are reviewed by solving the model over the next segment. The model may also be used at the strategic level to evaluate the consequences of variations in the fleet size. A Lagrangian heuristic method was compared with a simple branch-and-bound procedure. Results obtained on real-life and randomly generated instances led to the conclusion that the model is very tractable. The largest instance solved contained 100 terminals and 20 car types. Substitution possibilities are also treated by extending the basic formulation but no specific results are given for this extension. More details on this approach are given by JOBORN (1995) who also presented an analysis of empty freight car distribution at Swedish State Railways. An approach to determine train frequencies to minimize total costs for running trains and distributing empty cars was also introduced in a related paper by FLISBERG *et al.* (1996).

Multiple railroad models

A traditional repositioning strategy for freight cars consists of returning each unloaded car to its original loading point. This is a very simple and convenient approach given that a significant portion of freight shipments are made from the territory of one railroad to that of another. In the hope of reducing costs associated with empty movements, the concept of car pooling has gradually been introduced. Under a pooling agreement, railroads and shippers agree that cars unloaded at destination can be sent to any of a set of loading points.

A transshipment model to determine daily repositioning decisions that minimize network-wide costs was proposed by KIKUCHI (1985). GLICKMAN and SHERALI (1985) described two optimization approaches for the distribution of pooled cars that focus, respectively, on the benefits to the system as a whole and on the benefits to the individual railroads.

More recently, ADAMIDOU *et al.* (1993) argued that the problem of finding a global profit-maximizing distribution strategy for railroads sharing a fleet of cars is best represented as a generalized Nash equilibrium model. Their model includes coupling variables that link the individual multi-commodity flow subproblems of the railroads and is solved through a Gauss-Seidel algorithm that iterates between these subproblems. When solving the subproblem for a particular railroad, the coupling variables are fixed using the optimal flows obtained when last solving the subproblems for all other railroads. The approach was tested on a large-scale, three-railroad instance generated from actual data, and appeared to be fast and robust. Different solution strategies were compared as well as various demand conditions.

The pooling of railcars used for the transportation of automobiles was studied by SHERALI and TUNCBILEK (1997) who proposed static and dynamic models for

the fleet sizing problem. The static model tends to underestimate the real fleet size required because it is based on time-independent data. The dynamic model is based on a time-space network that represents the movement of empty cars between origins and destinations over the given planning horizon, with an objective of minimizing the fleet size required to satisfy all demands at different points in time. The problem is solved by decomposing the model into a series of smaller subproblems with a shorter, overlapping, temporal horizon. Once a subproblem is solved, the decisions for the initial part of the considered horizon are fixed, and the next subproblem is solved with the augmented flows. Test data instances generated randomly with realistic assumptions were used to evaluate the performance of the algorithm. The models have also been used successfully by the Association of American Railroads.

1.4 Scheduling Problems

While the models of Section 1.3 are mainly concerned with the efficient routing of trains and freight, scheduling models address the temporal dimension of railroad operations. Because the physical rail network is shared by a large number of trains, it is indeed necessary to synchronize their use of the available resources. Also, the scheduling of freight and passenger train movements has an important impact on the quality and level of service provided. Finally, the scheduling of transportation activities is highly dependent upon the availability of rail equipment, such as the locomotives and passenger cars, that are needed to operate trains.

Compound models reviewed in Section 1.3.2 are an attempt at integrating the routing and scheduling aspects of rail freight transportation. However, these two closely intertwined problems are most often treated separately: operating plans are developed first, followed by train schedules that specify tentative departure and arrival

times for the planned trains. The actual dispatching of trains is then performed by taking line capacity and other operational factors into account. This dispatching must often be performed simultaneously with the dispatching of passenger trains that operate in strict accordance with a timetable.

Most early models for train scheduling considered a set of stations connected by a single line. For example, the problem of developing timetables for passenger trains on a line of stations was studied by NEMHAUSER (1969) and SALZBORN (1969). The minimization of the number of railcars needed in a system of radial lines converging to a central station was also studied by SALZBORN (1970). Finally, an efficient approach for allocating demand to regular and express trains when delivering freight on a line network was suggested by ASSAD (1982).

More recently, the problem of finding a periodic train timetable that minimizes total passenger waiting time in stations of a network has received a lot of attention in the literature. Optimization models for that purpose were proposed by CEDER (1991), NACHTIGALL (1996), NACHTIGALL and VOGET (1996) and ODIJK (1996). The strategic problem of choosing a set of operating lines and their frequencies to serve demand and maximize the number of travelers on direct connections was studied by BUSSIECK *et al.* (1996). Also, ZWANEVELD *et al.* (1996) and KROON *et al.* (1997) have proposed models and algorithms for the related problem of routing trains through railway stations. These contributions were reviewed in detail by BUSSIECK *et al.* (1997), who discussed models for several discrete optimization problems in public rail transport. On a similar topic, NACHTIGALL (1995) discussed a problem that appears in passenger information systems and consists of computing shortest paths in a network with arc lengths that vary through time. Finally, NACHTIGALL and VOGET (1997) discussed a model for choosing the track segments to be upgraded to reduce train running times and thus minimize total passenger waiting time.

The following section contains a brief review of analytical models developed to measure the performance of a line relative to the traffic it accommodates, its configuration, operating policies, or other factors. Optimization models for train dispatching are then discussed, followed by models for locomotive assignment.

1.4.1 Analytical Line Models

Several models were proposed to estimate the delay to each train caused by interference on a rail line as a function of dispatching policies, traffic distribution and physical track topology. Early results were given by FRANK (1966) for the case of a single-track line with two-way traffic but a single train speed and equally-spaced sidings. A more elaborate model was then developed by PETERSEN (1974) for trains of different speeds in each direction and sidings that allow for both meets and overtakes. His model assumes uniform and independent distributions of trains in each speed class over the considered horizon. The mean running times for trains in each class are obtained by solving a set of linear equations. Expressions for the expected meet and overtake interference delays on a partially double-tracked line were also developed by PETERSEN (1975). Necessary and sufficient conditions to guarantee that line blocking does not occur were given by PETERSEN and TAYLOR (1983). Queuing models to determine the expected dispatching delays on a single-track line with low-speed traffic and widely-spaced sidings were also described by GREENBERG *et al.* (1988). Finally, KRAFT (1988) extended PETERSEN's approach to take multiple train interactions into account and compared the results with myopic and optimized train dispatching.

CHEN and HARKER (1990) studied a more realistic problem in which trains have scheduled departure and arrival times instead of being randomly distributed over the planning horizon. The mean and variance of travel time are estimated by solving

a system of nonlinear equations that also take into account uncertainties regarding actual departures. The extension of this framework to a partially double-tracked line was later presented by HARKER and HONG (1990).

Recently, HALLOWELL and HARKER (1996) described a model used to predict on-time arrival performance of trains on a partially double-tracked line with scheduled traffic. This model is an interesting alternative to simulation methods for estimating the lateness of delayed trains and can be used in tactical train scheduling or in train dispatching applications. In particular, it can be calibrated to generate target arrival times that can be achieved under an optimal planning of meets and overtakes.

The problem of track time use can also be seen from a game-theoretic standpoint. For example, HARKER and HONG (1994) presented an equilibrium model of an internal market for track time allocation. The generalized Nash equilibrium of the resulting model can be obtained by solving a quasi-variational inequality problem.

Most line delay models assume a fixed track configuration. However, PETERSEN and TAYLOR (1987) presented a method for finding the optimal location and length of sidings for a single-track line with high-speed passenger trains. The solution is derived under the hypothesis of ideal train performance, but an analysis of robustness to small and large delays is also presented. Simulation experiments were performed using a methodology, introduced by PETERSEN and TAYLOR (1982), which is a framework for modeling train movements over single-track and multiple-track lines.

Finally, ÖZEKICI and ŞENGÖR (1994) analyzed the problem of train dispatching with the emphasis on suburban passenger rail transport systems. They considered a train station in which passenger arrivals, although random, are related to train departures through the published timetable (ÖZEKICI, 1987). The model is used for

evaluating the performance, as measured by the service delay and the average waiting time of passengers, of different train dispatching strategies.

1.4.2 Train Dispatching Models

The train dispatching problem has received increased attention lately as several railroads are now developing and implementing advanced train control systems that provide real-time information on train position and velocity, as well as decisions to assist operations. These systems should help to reduce energy consumption and increase railroad lines capacity and service reliability with improved train dispatching. An introduction to computerized train dispatching was written by PETERSEN *et al.* (1986). SMITH (1990) exposed the general guidelines that should be followed in designing a module for meet/pass planning. JOVANOVIĆ and HARKER (1990) also presented some analysis on the proper elaboration of computer-aided train dispatching systems. Then, HARKER (1989, 1995) reviewed some models and algorithms developed for such systems, and discussed the importance of advanced train control in the context of the current restructuring of technology and management practices that is taking place in the railroad industry.

Although most optimization models for train dispatching have appeared in the last decade, other enumerative approaches have also been in use. In particular, SZPIGEL (1973) described a method for train dispatching on a single-track line with meets and overtakes. SAUDER and WESTERMAN (1983) proposed a decision support system for train dispatching that implicitly enumerates all feasible meet locations and selects the one minimizing delays. KRAFT (1987) presented a branch-and-bound approach for resolving train conflicts to minimize a weighted sum of delays.

Computerized tools have also been developed to assist planners in constructing feasible dispatch plans. Such systems were described, for example, by RIVIER and TZIEROPOULOS (1984, 1987) and CHURCHOD and EMERY (1987).

We now discuss the recent optimization models for train dispatching. We first review models that assume all trains are operating at their maximum velocity, followed by models for the case in which velocity is variable. A summary of these models is presented in Table 1.3.

Table 1.3: Characteristics of train dispatching models

Authors	Problem type	Planning horizon	Objective function	Model structure	Solution method
JOVANOVIC and HARKER (1991)	Fixed velocity	Tactical	Max reliability	MIP	Branch-and-bound
CAREY and LOCKWOOD (1995)	Fixed velocity	Operational	Min schedule deviation	Linear MIP	Heuristic decomposition
CAREY (1994)	Fixed velocity	Operational	Min schedule deviation	Linear MIP	Heuristic decomposition
CAREY (1994)	Fixed velocity	Operational	Min schedule deviation	Linear MIP	Heuristic decomposition
KRAY and HARKER (1995)	Fixed velocity	Tactical	Min schedule deviation	Nonlinear MIP	Heuristic decomposition
BRÄNNLUND et al. (1996)	Fixed velocity	Tactical	Min schedule deviation	Linear IP	Lagrangian relaxation
NÖU (1997)	Fixed velocity	Tactical	Min schedule deviation	Linear IP	Lagrangian relaxation
KRAAY et al. (1991)	Variable velocity	Tactical	Min train delays and fuel costs	Nonlinear MIP	Heuristic
HIGGINS et al. (1996)	Variable velocity	Operational	Min train delays and operating costs	Nonlinear MIP	Branch-and-bound
HIGGINS et al. (1997)	Variable velocity	Strategic	Min conflict delay and risk of delay	Nonlinear MIP	Heuristic decomposition

Fixed velocity models

The aim of train dispatching models is to determine where trains will meet and pass so as to minimize train delays or deviations from the planned schedule while satisfying a set of operational constraints. Because the meeting and passing of trains is intimately related to their operating speed, a complete model should treat velocity as a decision

variable. However, most dispatching models use a sequential approach and assume that trains will operate at maximum velocity whenever possible. A velocity profile is later determined for each train individually.

JOVANOVIĆ and HARKER (1991) proposed the SCAN system for the tactical scheduling of trains and maintenance operations. The main goal of their approach is to help in the design of reliable schedules in the sense that they are robust under stochastic operating conditions. The time horizon considered is a single day. The system, which can deal with single and double track segments, starts with a proposed schedule and first verifies its feasibility by separately analyzing each line of the network. To verify feasibility over a given line, a mixed integer programming problem with no explicit objective is solved with a branch-and-bound procedure to generate a feasible plan of meets and overtakes. This procedure incorporates a simulation method to model train movements and interactions. An automatic update procedure also helps in modifying an infeasible schedule into a feasible one. The mixed-integer programming problem has binary variables that indicate the ordering of the trains and continuous variables that represent departure and arrival times of trains at meetpoints. A complex set of constraints impose logical conditions concerning the meeting, passing and following of trains. Time window constraints on the arrival and departure of each train are also present. The system performed well on a real-life network with 24 lines and schedules for 100 freight and passenger trains. The authors also report an implementation at a major U.S. railroad.

CAREY and LOCKWOOD (1995) described a model for the train dispatching problem on a line composed of several links connected by stations where overtaking can take place. The line is dedicated to traffic in one direction but trains operate at different speeds. Their model is a 0-1 mixed integer program that incorporates several headway constraints, bounds on departure and arrival times, and additional constraints used to strengthen the model. The *headway* is the time or distance

separating two trains on the same link. The objective function to be minimized is rather general and takes deviations from the preferred schedule into account. The authors proposed to solve the model using a heuristic approach that first dispatches trains one at a time to obtain an initial solution, and then possibly redispatches individual trains to improve this solution. The subproblem of dispatching a single train has a reduced number of binary variables because the sequence order of the already dispatched trains is held fixed while the timings are allowed to vary. This problem is solved using a branch-and-bound procedure with branching decisions made on the link variables that specify the sequence order of the trains. Various strategies were proposed to accelerate the solution of the subproblem. In particular, branching in a depth-first search on the variables associated with the links in the same order as they are traversed by the trains seems to dramatically reduce the computing times. Of course, this method does not guarantee the optimality of the produced solution, nor does it ensure that a feasible solution will be found even if one exists. Good results are reported for computational experiments on small instances with 10 trains and 10 links.

In a follow-up paper, CAREY (1994a) extended the original model to introduce choices among multiple lines in each direction and choices of platforms to use for departures, arrivals, and stops at stations. This is done by introducing a more general type of link with two special cases representing train links and stations. Again, the model is solved with a heuristic decomposition approach that dispatches trains one at a time and redispatches individual trains until no further improvement in the solution is possible. Finally, the extension from one-way to two-way tracks was done by CAREY (1994b) who showed that the same solution methodology still applies in that case.

A model for optimizing freight train schedules was proposed by KRAAY and HARKER (1995). The goal of their approach is to provide a link between tactical

train scheduling and actual operations by generating target times to be used in dispatching models such as the SCAN system (JOVANOVIC AND HARKER, 1991). The model, which is a large nonlinear, mixed-integer program, directly considers the current position and relative importance of each train. Its solution indicates the target time for each train at each important point in its itinerary. For given values of the integer variables that determine the meeting and passing of trains, the model reduces to a continuous variable subproblem that is solved with an algorithm combining restricted simplicial decomposition and network flows. A simple heuristic approach and local search methods can be used to determine feasible values for the integer variables. Comparisons on a large set of real-life instances showed that the local search heuristics produced better results than the simple heuristic but required excessive computing time.

BRÄNNLUND *et al.* (1998) proposed a model to determine a profit maximizing schedule in which profit is measured by estimates of the value of running different types of services at specified times. The problem is formulated as a large integer programming problem and is solved with a Lagrangian relaxation approach in which track capacity constraints are dualized. The relaxed problem thus decomposes into a shortest path problem in a space-time network for each individual train. Feasible solutions are obtained with a heuristic that sequentially dispatches each train according to a priority list given the current dual prices associated with track capacity constraints. Various dual optimization schemes were compared on instances with 26 and 30 trains on a single-track line connecting 17 stations. Computational experiments indicated that feasible solutions within a few percent of the lower bound were found in rather short computing times. According to this computational experience, the duality gap appears to increase as the line becomes more congested. Even though the approach is described for a single-track line, it easily extends to a double-tracked one.

In a follow-up paper, NÕU (1997) suggested and compared alternative approaches for generating feasible solutions. The author first extended the priority list heuristic described previously by BRÄNNLUND et al. In particular, a tabu search heuristic was proposed for the problem of finding the best possible permutation of the trains. Then, a conflict resolution heuristic which treats conflicts in order of occurrence was described. Finally, a greedy local improvement heuristic was introduced. This heuristic considers a feasible solution and tries to improve it by performing changes that maintain feasibility while improving the overall profit associated with the schedule. Computational experiments were performed on the same data that were used by BRÄNNLUND et al. Solution quality improvements in the order of 1% were obtained while computation times remained rather similar. The author concluded that the most effective approach is an enhanced priority list heuristic with a tabu search procedure to update the list.

Variable velocity models

Models that treat velocity as a decision variable are not very common even though they represent a significant improvement over fixed velocity models. Indeed, by treating operating speed endogenously, such models not only minimize deviations from the schedule but also quantify and minimize fuel consumption.

KRAAY *et al.* (1991) treated a train pacing problem in which train velocity and meeting and passing schedules are determined together to minimize fuel consumption and delays while satisfying time windows on the departure and arrival of each train. Their formulation is a nonlinear mixed integer program with a convex objective function. First, the authors proposed a branch-and-bound algorithm in which the initial relaxation is obtained by linearizing the objective function and by ignoring train interactions. This relaxation decomposes into simple linear programs solvable

with a sorting routine and a line-search procedure. At each node of the branch-and-bound tree, cutting planes are added to gradually impose the relaxed constraints. When the relaxation solved at a node of the tree yields a feasible meet/pass plan, a feasible solution for the global problem can be computed by solving a nonlinear program in which the integer variables are held fixed. An alternative approach, based on the generation of feasible plans for the meeting and passing of trains, was also proposed. For each plan, the optimal velocity profiles are also computed by solving the nonlinear program with the integer variables being fixed. This approach is very convenient because it can use an oracle to generate plans that obey very complex constraints that do not even possess a mathematical representation. This approach can also be used to evaluate and rank different scenarios. Feasible meet/pass plans are generated using the logic of the SCAN system (JOVANOVIĆ and HARKER, 1991). Finally, the authors proposed a rounding heuristic to filter out meet/pass plans and retain only those closest to the optimal solution obtained when ignoring train interactions. Results on instances of a major railroad produced fuel savings in the order of 5% while the standard deviation in train arrival times decreased by more than 19%. A theoretical analysis shows that, as the number of sidings goes to infinity, the probability that the heuristic will give an optimal solution goes to one.

HIGGINS *et al.* (1996) proposed a model and a solution method for the dispatching of trains on a single-track line. Their model mainly addresses the operational problem of dispatching trains in realtime but can also serve at the strategic level to evaluate the impacts of timetable or infrastructure changes on train arrival times and train delays. The formulation is a complex nonlinear mixed integer program that incorporates lower and upper limits on train velocities for each train on each segment. The objective function seeks to minimize a combination of total train tardiness and fuel consumption. When a train will be delayed in a conflict at the next siding or has slack time, it will be paced to reduce fuel costs. The problem is solved using a branch-and-

bound algorithm with lower bounds computed by using an estimate of the remaining delay cost, based on the calculation of the least cost path for each train. Comparisons with both an enumerative procedure that computes lower bounds by relaxing the remaining conflict constraints and a tabu search heuristic showed that the proposed method is very effective at finding the optimal solution. A real-life instance with 31 trains and 14 sidings was solved in less than one minute. Other experiments are reported on instances of similar size.

In a follow-up paper, HIGGINS *et al.* (1997) extended their solution methodology for simultaneously deciding the number and location of sidings and the optimal train schedule for a single-track line. This strategic problem is again modeled as a nonlinear mixed integer program. It is solved with a heuristic decomposition scheme that iterates between two subproblems until no further improvement is possible. The first subproblem chooses the positions of the sidings and the departure and arrival times for a given fixed schedule; the second subproblem chooses a train schedule, considering fixed siding locations. The method also considers an initial set of sidings that are held at fixed position. Because maximum train velocity on a given segment depends on the sidings location, velocity is determined endogenously. The objective function minimizes a weighted combination of conflict delay and risk of delay. The risk of delay represents the likely delay caused by unexpected events. Computational experiments on instances with up to 30 trains indicated that the algorithm converges very quickly.

1.4.3 Locomotive Assignment Models

Given a planned train schedule, the locomotive assignment problem consists of assigning a set of locomotives to the scheduled trains to satisfy requirements expressed as a number of locomotives or as a measure of the pulling power needed (i.e.,

horsepower and tonnage). At a strategic planning level, the objective followed is usually to minimize the required fleet size. At the tactical and operational levels, the available rolling stock is given and one usually wants to minimize costs incurred by light running. Light running or *deadheading* occurs when an engine must be repositioned between two successive trips.

Early research on the problem of assigning engines to trains was conducted by CHARNES and MILLER (1957) who used linear programming for the assignment of crew-engine pairings to a set of potential trips to provide each train in a given schedule with sufficient resources. BARTLETT (1957) gave an algorithm for minimizing fleet size based on the idea that, for a fixed time horizon, this objective is tantamount to minimizing total idle time. An algorithm for finding an assignment that satisfies maintenance constraints while minimizing deviations from a target mileage between successive maintenance stops was proposed in related work by BARTLETT and CHARNES (1957).

Over the years, many railways have developed decision support systems to assist planners in making locomotive assignment and scheduling decisions. Although early systems relied in large part on simulation techniques and decision rules dictated by experience, some of them also used optimization methods. For example, GOHRING (1971) and MCGAUGHEY *et al.* (1973) described a periodic network flow model, solved with the out-of-kilter algorithm (FORD AND FULKERSON, 1962), to minimize fleet size at Southern Railway. Also, HOLT (1973) mentioned the use of branch-and-bound procedures and decomposition approaches for locomotive distribution at British Railways.

We now review the more recent optimization models for locomotive assignment. We first discuss the case in which each train needs a single engine, followed by models

for the multiple engine case. The simultaneous assignment of both engines and cars to passenger trains is treated last. Table 1.4 provides a summary of these models.

Table 1.4: Characteristics of locomotive assignment models

Authors	Problem type	Planning level	Objective function	Model structure	Solution method
FORBES et al. (1991)	Single engine	Tactical	Min operating costs	Assignment problems	Branch-and-bound
FISCHETTI and TOTH (1997)	Single engine	Tactical	Min fleet size and deadheading	Assignment problems	Lagrangian relaxation
FLORIAN et al. (1976)	Multiple engines	Strategic	Min investment and maintenance	Multi-commodity	Benders decomposition
SMITH and SHEFFI (1988)	Multiple engines	Strategic	Min operating costs	Multi-commodity	Heuristic
CHIH et al. (1990)	Multiple engines	Operational	Max expected profit	Multi-commodity	Heuristic decomposition
ZIARATI et al. (1997)	Multiple engines	Operational	Min operating costs	Multi-commodity	Dantzig-Wolfe decomposition
NÔU et al. (1997)	Multiple engines	Tactical	Min operating costs	Multi-commodity	Dantzig-Wolfe decomposition
ZIARATI et al. (1997)	Multiple engines	Operational	Min delays	Multi-commodity	Dantzig-Wolfe decomposition
CORDEAU et al. (1998)	Engines and cars	Tactical	Min operating costs	Multi-commodity	Benders decomposition

Single locomotive models

Most models for the problem in which multiple engine types are available but each train needs a single locomotive have a multi-commodity network flow structure with linking constraints that ensure that each train is covered exactly once. For example, BOOLER (1980) proposed a heuristic algorithm that starts with a feasible allocation of locomotive types to the trains and iteratively updates this allocation using the dual information gathered when solving the resulting assignment problems. A Lagrangian relaxation approach, that dualizes the linking constraints in the objective function, was later proposed by the same author (BOOLER, 1995). WRIGHT (1989) compared stochastic algorithms based on the solution of assignment problems and the update of the locomotive types assigned to the trains.

An exact algorithm for a model with a similar structure was proposed by FORBES *et al.* (1991). The objective function takes into consideration fixed costs and operating costs. The solution technique consists of solving the LP relaxation of an integer programming formulation before applying a branch-and-bound procedure to obtain an integer solution. To solve the continuous relaxation, a further relaxation is obtained by removing the locomotive type restrictions. The solution to that problem is then converted into a dual feasible solution to the original problem and the dual simplex method is used to obtain the optimal solution to the LP relaxation. Branching is first performed on the number of locomotives used. Additional branching is performed on the successors of the trains and on the locomotive types assigned to the trains. The data sets used for testing purposes did not impose constraints on the number of available locomotives of each type, but the authors mentioned how these can be enforced in their formulation. They reported very small integrality gaps, in particular when the objective function does not include preferences for locomotive types.

Very recently, a heuristic method for the weekly problem was proposed by FISCHETTI and TOTH (1997). Engines are distributed across a number of depots that are associated with stations of the network. Each depot has a maximum number of engines available and each engine must go through its depot every week to allow for maintenance. In addition, engine trips must satisfy a set of operational constraints. By relaxing the maintenance and operational constraints, one obtains an assignment problem whose solution provides a very good lower bound on the optimal solution. The objective function is a weighted combination of the number of engines needed, the number of deadheading trips performed, and the distance covered by deadheading trips. Real-life instances with up to 10,000 trains are solved in less than one hour on a workstation computer. Cost savings in the order of 10-20% are typically obtained over the solution in use by the Italian Railways.

Multiple locomotive models

When each train may require more than one locomotive but these requirements are given as a number of engines, the problem can still be formulated as a multi-commodity network flow problem with rather simple linking constraints. The most difficult version of the problem occurs when multiple locomotive types are available and each train may require more than one locomotive to satisfy its requirements expressed in terms of motive power.

One of the first models dealing with this version of the problem was proposed by FLORIAN *et al.* (1976). The strategic problem considered is to select the mix of engine types that gives the lowest capital investment and maintenance costs over a long planning horizon, while providing each train with sufficient engines to meet its motive power requirements. In this model, the motive power requirements of each train are determined according to its weight and length in terms of cars, and to the route on which it must travel. The model used is defined on a set of network flow circulation problems with some linking constraints that translate the motive power requirements. The solution approach is based on Benders decomposition (BENDERS, 1962) and takes advantage of the particular structure of the problem. Variables in the integer programming master problem impose lower bounds on the arcs of network flow subproblems. To speed up the solution of the master problem, a decomposition scheme is coupled with a rounding heuristic. Upper bounds on the number of engines of each type are not treated. Computational results were reported on problems with a few hundred trains and the convergence was deemed slow on the larger instances. However, it should be emphasized that the algorithm was stopped after less than 30 iterations were performed. Hence, given the performance of today's computers, it is very likely that the conclusions would now be different.

A model that incorporates the uncertainty in locomotive requirements was suggested by SMITH and SHEFFI (1988). This model has a multi-commodity network flow structure with linking constraints that enforce locomotive requirements expressed as a lower bound on the horsepowers supplied to each train. These constraints are relaxed in the objective function by using a penalty function that permits deviations from the requirements at a cost. Also, the lower bounds on horsepower are replaced by random variables with known distributions. The resulting model has a convex nonlinear cost function and is solved with a two-phase heuristic. In the first phase, a feasible solution is obtained by incremental flow assignments along shortest paths. In the second phase, interchanges are performed to improve the solution by identifying cycles with a negative marginal cost. The major advantage of this heuristic procedure is that it maintains integrality throughout. To evaluate its performance, lower bounds were computed with two approaches. The first one relaxes integrality constraints and solves the resulting problem with a Frank-Wolfe method (FRANK and WOLFE, 1956). The second one uses a piecewise linearization of the cost function to obtain a pure network flow problem. Computational experiments on instances with up to 102 trains produced feasible solutions with short computing times and costs within a few percent of the best lower bound.

CHIH *et al.* (1990) described the implementation of an operational planning model for locomotive assignment. The model, which seeks to maximize the difference between expected revenue and operational costs, is based on a time-space network representing all possible locomotive movements during the planning horizon. To obtain a first approximation of motive power assignments to a set of weekly scheduled trains, a multi-commodity network flow problem is solved with a resource-directive decomposition approach. Locomotives that must be directed to a shop for maintenance are then routed individually by solving a shortest path problem and horsepower requirements are lowered to reflect these assignments. Given the

solution to the multi-commodity network flow problem and the residual requirements, locomotive consists are finally built for each train by an exhaustive enumeration process. The approach was tested on actual data from the Union Pacific Railroad. On an instance with 15 types of locomotives and a network for each type containing more than 25,000 arcs, a solution was found within 30 minutes.

More recently, a Dantzig-Wolfe decomposition approach (DANTZIG and WOLFE, 1960) was developed for the operational version of this problem by ZIARATI *et al.* (1997b). Train requirements are determined as above but some engines must also be dispatched to special stations at which they have to perform local work. A list of preferred locomotives is considered for each train and care is also taken of the locomotives that must be routed to a shop for maintenance. The objective considered is the minimization of the total operational costs. The problem is modeled as a multi-commodity flow problem with supplementary variables and constraints. The time horizon considered is a week, but, to solve very large instances, the problem is divided on a temporal basis into a set of overlapping slices involving fewer trains. Once the problem for a slice is solved, the problem for the next one is solved with initial conditions determined by the solution of the preceding slice. The problem for each slice is solved using a branch-and-bound procedure in which the linear relaxations are solved by column generation. Constrained and unconstrained shortest path problems must be solved on an acyclic network to generate columns for the master problem. A heuristic branching strategy is used, in which many path variables are fixed together. Branching decisions are made on the path variables with the largest fractional part, and the selected variables are rounded up to the next integer. Computational experiments carried out on real-life data involving approximately 2000 trains allowed an improvement of 7% over the solution in use by the company when taking slices of two days with a one-day overlap. This improvement goes to 7.5% when slices of three days are used, but the computations then take a few hours. ZIARATI *et al.*

(1998) introduced additional cuts, based on the enumeration of feasible assignments of locomotive combinations to trains, which strengthen the LP relaxation lower bound and improve solution quality. A day-to-day operational model was also proposed by ZIARATI (1997).

A similar approach was used by NÕU *et al.* (1997) for the tactical assignment problem at Swedish State Railways. In this problem, cyclic locomotive assignments are sought and maintenance constraints related to cumulated distance must be satisfied. Two approaches based on a branch-and-bound procedure and Dantzig-Wolfe decomposition are presented for solving the problem. In the first approach, the weekly problem is replaced by a series of smaller size problems with overlapping horizons. In the second one, maintenance constraints are relaxed to obtain a smaller problem that is solved without being decomposed on a temporal basis. Tests performed with actual data from Swedish State Railways involving 2422 trains showed that the first approach failed to produce a feasible cyclic solution. The second approach produced a solution that violated maintenance constraints for a very limited number of locomotives.

In some cases, the operational locomotive assignment problem may be infeasible because not enough engines are available. One possibility to circumvent this difficulty is to allow train undercovering. Undercovering happens when the motive power requirements are not fully satisfied. This is easily achieved by introducing slack variables in the appropriate constraints. ZIARATI *et al.* (1997a) presented an alternative approach that consists of delaying trains. The basic idea of the method is to postpone the departure of an undercovered train until enough locomotives are available at the origin station. For the case of express trains, one can instead postpone the departure of a preceding train to assign the available engines to the express train. Using these strategies, one can often find a feasible solution in terms of the covering constraints. To determine a valid lower bound, the authors use an augmented network

that includes both penalty and delay costs, as well as fixed and routing costs. The solution to an instance with almost 2000 train segments was obtained in less than 10 minutes and had a cost within 5% of the lower bound.

Locomotive and passenger car models

Very little work has been accomplished concerning the assignment of locomotives and cars in the context of passenger transportation. A decision support system was developed by RAMANI and MANDAL (1992) for the planning of passenger trains at Indian Railways. However, the assignment of locomotives and cars is dealt with separately and the system uses a simple local improvement procedure that generates optimal train connections by examining the departures and arrivals at individual stations. This procedure is reminiscent of the algorithm given by BARTLETT (1957) for fleet size minimization. The work of RAMANI and MANDAL extends an information system for car assignment developed by RAMANI (1981).

Recently, an optimization model for the assignment of both locomotives and passenger cars was proposed by CORDEAU *et al.* (1998b). The tactical periodic problem is formulated as an integer programming problem based on a time-space network. As in the work of FLORIAN *et al.* (1976) for locomotive assignment, this model possesses an interesting variable decomposition: for given values of the binary variables that represent the assignment of equipment combinations to trains, the problem decomposes into one network flow problem for each type of equipment. The formulation incorporates compatibility constraints between the different types of equipment that may be combined to form valid train consists. Equipment availability for each type is also enforced. However, the model does not directly impose maintenance constraints. Comparisons between primal and dual

decomposition methods and a simplex-based branch-and-bound approach showed that the formulation was best solved using Benders decomposition (BENDERS, 1962). Algorithmic refinements were suggested to improve the performance of the algorithm. In computational experiments performed on real-life data, the algorithm found optimal solutions within short computation times. The largest instance solved contained six types of equipment and 348 trains over a period of one week.

1.5 Conclusions

This paper has presented a review of the recent optimization models proposed for solving routing and scheduling problems in rail transportation. The field is clearly receiving increased attention as measured by the number of contributions in the last few years. The nature and scope of the research conducted is also gaining in diversity as nearly every domain of rail transport planning has been the object of some recent research.

There also appears to be a constant refinement and diversification of the modeling and solution methods proposed and used. Early models were usually built to have a structure that made them solvable by linear programming or network optimization. One then witnessed a gradual introduction of integer programs with simple underlying structures. Although some recent models are solved with more sophisticated mathematical programming techniques, others still are solved using meta-heuristics that have proven to be very effective for several classes of discrete optimization problems. Of course, this progression is also made possible by the increased power of computers and information systems.

As mentioned in the introduction, optimization models for train routing and scheduling have advanced tremendously in the last few years. Whereas early models were often based on very crude approximations of reality, recent applications demonstrate an important effort to deal with complex yet important characteristics of the actual functioning of railway systems. As a result, problems which, in the past, were only approachable by simulation can now be solved, at least approximately, using mathematical optimization. Nevertheless, simulation techniques have also made considerable progress in the last decade and remain a very useful tool of analysis and support to decision making. The recent work of POWELL (1995) is an illustrative example of this progress.

Also, despite the increasing realism of optimization models, considerable work remains to be accomplished to make the railways benefit from this wealth of knowledge. Even though most proposed models are tested on realistic data instances, very few are actually implemented and used in railway operations. Hence, efforts must be made to bridge the gap between theory and practice. MARTLAND and SUSSMAN (1995) presented an interesting discussion of factors that explain the success or failure of different approaches.

Future research paths in rail transportation planning are oriented toward models that address the integration of various policies. Because rail activities are generally complex and involve large-scale systems, the traditional approach in the industry has been to separate planning activities into several components. This natural tendency yields more manageable subsystems but also presents several limitations. In particular, there is a strong incentive to simultaneously treat routing and scheduling problems because of the important interactions linking these two categories of decisions. Hence, models that integrate several aspects and levels of planning should be increasingly common in upcoming years.

Acknowledgments

We sincerely thank Dr. Edwin R. Kraft of Amtrak. His suggestions helped us to improve both the contents and presentation of this paper. Thanks are also due to Professor Jacques Desrosiers for his comments on preliminary versions of the text. This work was supported by the Québec Government (Fonds pour la Formation de Chercheurs et l'Aide à la Recherche) and by the Natural Sciences and Engineering Research Council of Canada.

Chapitre 2

Simultaneous Locomotive and Car Assignment at VIA Rail Canada

Article écrit par Jean-François Cordeau, Guy Desaulniers, Norbert Lingaya, François Soumis et Jacques Desrosiers; soumis pour publication à *Transportation Research B*.

Comme en témoigne la revue de littérature du chapitre précédent, aucun modèle n'avait précédemment été proposé pour résoudre le problème de l'affectation simultanée des locomotives et des wagons aux trains de passagers. Dans cet article, nous proposons un premier modèle développé dans le contexte d'une application pratique chez le transporteur canadien VIA Rail. En développant ce modèle, nous avons néanmoins tenté de conserver un niveau de généralité suffisant pour que l'approche puisse par la suite être adaptée aux problèmes d'entreprises différentes.

Après avoir défini le problème étudié, nous décrivons en détail les réseaux espace-temps utilisés pour représenter l'ensemble des mouvements possibles pour les différents types d'équipement à l'intérieur de la période de planification. La définition de ces réseaux sert en outre à imposer certaines contraintes telles que des temps de connexion qui varient en fonction de l'orientation des trains.

Nous donnons ensuite une formulation mathématique du problème qui est basée sur ces réseaux espace-temps mais qui contient également de nombreuses

contraintes liantes entre les différentes pièces d'équipement. En plus des contraintes de demande et des contraintes de capacité des locomotives, le modèle comprend des contraintes d'entretien, des contraintes de disponibilité d'équipement ainsi que des contraintes d'espace d'entreposage. La première formulation suppose que tous les trains opèrent durant le jour et que l'entretien s'effectue exclusivement pendant la nuit à l'unique centre d'entretien disponible. Elle suppose également que la combinaison d'équipement utilisée sur chaque train est choisie à l'avance. Nous expliquons par la suite comment généraliser le modèle afin de relâcher ces hypothèses.

Une méthode de résolution par séparation et évaluation progressive est présentée pour résoudre ce modèle. À chaque noeud de l'arbre de branchement, une relaxation linéaire est résolue à l'aide d'une approche de génération de colonnes. Chaque colonne générée correspond en fait à un itinéraire débutant au centre d'entretien, couvrant un certain nombre de trains, et se terminant au centre d'entretien à l'intérieur de la durée maximale permise entre deux entretiens successifs. Afin d'imposer les contraintes d'intégrité, des décisions de branchement heuristiques de plusieurs types sont utilisées.

Ce modèle étant très difficile à résoudre en raison du grand nombre de contraintes liantes qui apparaissent dans le problème maître de la décomposition de Dantzig-Wolfe, différentes stratégies sont utilisées afin d'en réduire la taille. Une de ces stratégies consiste à définir des équipements de base comprenant une locomotive et un certain nombre de wagons nécessaires pour former un train minimal. Ces équipements de base permettent de réduire le nombre de contraintes de demande et de contraintes de capacité, allégeant ainsi considérablement le problème maître.

Puisque l'entreprise désire non seulement minimiser les coûts d'opération mais également réduire le nombre d'opérations de couplage et de découplage des wagons, une approche de résolution en deux phases est utilisée. Dans la première phase, le problème est résolu sans tenir compte de ces opérations mais l'intégrité n'est exigée

que sur les variables de flot associées aux locomotives. Dans la seconde phase, un problème réduit est résolu en fixant les chemins de locomotives obtenus durant la première phase et en imposant des pénalités pour le couplage et le découplage des wagons. Ce processus de résolution séquentiel est clairement heuristique mais réduit de manière très importante la difficulté du problème.

Les résultats numériques montrent que la méthode peut résoudre des problèmes réels avec tout l'éventail des contraintes en quelques heures de calcul sur une station de travail. De plus, les comparaisons avec les solutions produites manuellement par les employés de planification de VIA Rail indiquent que notre approche permet très souvent de réduire à la fois les coûts d'opération et le nombre d'opérations de couplage et de découplage des wagons.

La principale contribution de cet article est de présenter le premier véritable modèle pour l'affectation simultanée de locomotives et de wagons aux trains de passagers. Le niveau de détail considéré dans ce modèle et les résultats obtenus confirment qu'il est possible de développer des modèles relativement complets pour ce type de problèmes, et que ces modèles peuvent être résolus de manière approximative en des temps de calcul raisonnables compte tenu du fait que la planification tactique n'est revue que quelques fois par année.

Remarquons enfin que cet article fait référence à l'article présenté au prochain chapitre car, bien que le travail présenté ici ait débuté antérieurement, la rédaction du texte ne fut réalisée qu'après l'implantation du logiciel chez VIA Rail.

Simultaneous Locomotive and Car Assignment at VIA Rail Canada

JEAN-FRANÇOIS CORDEAU[1], GUY DESAULNIERS[2],
NORBERT LINGAYA[1], FRANÇOIS SOUMIS[1]
and JACQUES DESROSIERS[3]

November 1998

[1] *École Polytechnique de Montréal*
[2] *Université Laval*
[3] *École des Hautes Études Commerciales de Montréal*

Abstract

An important aspect of railway planning concerns the distribution of locomotives and cars in the network and their assignment to the scheduled trains. In this paper, we present a sophisticated model and a heuristic solution approach based on mathematical optimization for the assignment of locomotives and cars to passenger trains. Given a periodic schedule and a fleet composed of several types of locomotives and cars, our approach determines a set of equipment cycles that cover all scheduled trains while satisfying a set of operational constraints. We first present a basic formulation that translates maintenance requirements and other fundamental difficulties of the problem. We then discuss several extensions, such as substitution possibilities and the minimization of switching operations, which are required in a real-life application. The resulting model is optimized with a branch-and-bound method in which the linear relaxations are solved by a Dantzig-Wolfe decomposition. The model and solution strategy were tested on data from VIA Rail in Canada and a complete system based on this approach is now implemented at the company.

Keywords: Rail passenger transportation; multi-commodity network flow model; Dantzig-Wolfe decomposition.

2.1 Introduction

A major concern to every railway is to optimize the distribution and the use of the available stock of locomotives and cars. In the context of rail freight transportation, the specification of train formation plans and the assignment of engines to trains are usually dealt with separately. Once freight cars have been assigned to a set of trains according to operating policies, the requirements of each train in terms of motive power can be computed and a locomotive assignment problem can be solved. In practice, the operating policies are usually updated every few months whereas the locomotive assignment problem must be solved more frequently to account for daily or weekly variations in the demand for transportation.

Separating the formation of trains from the assignment of locomotives may certainly yield suboptimal decisions. However, this is a very natural approach which significantly reduces the size of the resulting problem. Simultaneously planning the assignment of freight cars and locomotives to trains would lead to very large problems even for small railways. Also, since demand varies continually, cyclic solutions are seldom applicable. Traditionally, a similar sequential planning approach has also been very common in the context of passenger transportation where the assignment of locomotives and cars to trains are often treated separately despite the fact that a simultaneous approach could be used.

However, rail passenger transportation differs from freight in one important respect: the same trains are usually run each week with more or less the same number of cars. Indeed, passenger trains generally adhere closely to a published schedule which is revised on a seasonal basis to account for changes in the demand. Also, the number of passengers wishing to travel from one city to another at a given moment varies only slightly from week to week. Hence, there is a strong incentive to

treat the cars and locomotives together so as to obtain a global equipment assignment plan that either maximizes fleet utilization or minimizes operating costs. While this would be extremely difficult in freight transportation given the large number of cars that make up each train, it is a reasonable goal in passenger transportation. Since the same schedule is to be repeated cyclically for a certain period of time, important savings can thus be obtained by treating both locomotives and cars in the same model as opposed to optimizing their use separately.

Given a periodic train schedule and a fleet composed of several types of equipment, the simultaneous locomotive and car assignment problem is to determine a set of minimum cost equipment cycles such that every train is assigned appropriate equipment and some side constraints are satisfied. A large variety of side constraints must often be considered and most are dictated by operating policies or the characteristics of the physical network. For example, each unit of equipment must usually be inspected at regular intervals to comply with safety regulations and perform minor repairs. Also, the maximum number of cars which may remain idle in a given station is limited by track capacity.

The simultaneous locomotive and car assignment problem may be further complicated by the fact that combining different units of equipment has an effect on operating speed which, in turn, impacts on the arrival times of the trains. Since schedule adherence is of prime importance in passenger transportation, operating speeds must then be considered explicitly. In addition, when equipment units are combined together to form *train consists*, compatibility restrictions must be taken into account: while several types of locomotives and cars may be allowed on a given train, some of these types may be pairwise incompatible. Finally, when the contents of a train consist can be modified during its trip through the network by *switching* cars on or off the train, a particular modeling approach must be adopted to appropriately translate the fact that these modifications have an impact on connection times.

Literature Review. Very few references can be found in the Operations Research literature regarding the simultaneous assignment of locomotives and cars to passenger trains. One of the first known efforts in this direction is a decision support system developed by RAMANI and MANDAL (1992) for the maximization of equipment utilization on passenger trains at Indian Railways. Improvements over a current solution are obtained by using a simple local exchange procedure generating optimal train connections in each station by matching compatible departures and arrivals. This approach, which is clearly heuristic as it fails to consider the network as a whole, produced significant savings on the large instances on which it was tested. A system was also developed by SABRE for the French Railways SNCF (BEN-KHEDER *et al.*, 1997). This system optimizes the assignment of equipment modules containing both locomotives and cars. However, these modules are already formed and there only remains to assign a certain number of modules to each train. Also, all modules allowed to cover a given train are compatible and their coupling does not affect operating speed. Finally, modules can be coupled and decoupled in just a few minutes.

Very recently, CORDEAU *et al.* (1998b) proposed a basic modeling and solution approach for the simultaneous assignment of locomotives and cars. Their model is based on a set of time-space networks associated with the different equipment types available. The definition of these networks captures several characteristics of the problem such as restrictions on train modifications and orientation-dependent connection times. The networks are linked by demand and capacity constraints as well as compatibility restrictions. These restrictions are modeled by defining a set of possible train consist types representing valid combinations of equipment. Each of these combinations contains a locomotive type and some compatible car types, and its operating speed is determined by the slowest of its components. The proposed model possesses an interesting variable partitioning which makes it well suited for a Benders decomposition approach: for a given assignment of consist types to trains,

the problem decomposes into a set of network flow subproblems with one additional constraint per subproblem. However, although it was tested on real-life data and produced optimal solutions in reasonable computing times, the model is probably not sophisticated enough to be used in practice. In particular, it does not deal with maintenance constraints. The model introduced in the present paper incorporates a much larger set of constraints and possibilities which are required in a commercial application. Hence, our modeling approach borrows some ideas from the work of CORDEAU *et al.* but is clearly differentiated by the broader range of refinements captured by the formulation.

Whereas the simultaneous assignment of locomotives and cars to passenger trains has received very little attention in the literature, the problem of locomotive assignment in the context of freight transportation has been the object of much more work. In the most simple version of the problem, several locomotive types are available but each train requires a single engine. Models and algorithms for this version of the problem were first proposed by BOOLER (1980, 1995), WRIGHT (1989) and FORBES *et al.* (1991). Recently, FISCHETTI and TOTH (1997) developed a heuristic algorithm based on the solution of assignment problems for the weekly cyclic problem. Their approach takes maintenance and refueling constraints into consideration and was able to produce near-optimal solutions to very large instances from the Italian Railways FS.

A more complex problem occurs when each train may require several locomotives. One of the first models for this case was developed by FLORIAN *et al.* (1976). The authors considered the strategic problem of locomotive acquisition and proposed a multi-commodity network flow (MCNF) formulation solved with an algorithm based on Benders decomposition. More recently, an MCNF model incorporating uncertainty in locomotive requirements into the objective function was developed by SMITH and SHEFFI (1988). The model is solved with a two-phase heuristic

approach that produced good results on small instances from a railroad. Then, CHIH *et al.* (1990) reported the implementation of a planning system based on mathematical decomposition at the Union Pacific Railroad. The computational results obtained on some large instances suggested a significant reduction in the number of locomotives needed, in the operating costs, and in train delays. Finally, a Dantzig-Wolfe decomposition approach was proposed for the operational version of the problem by ZIARATI *et al.* (1997b). The problem is modeled as an MCNF problem with supplementary variables and constraints. A weekly horizon is considered but in order to solve very large instances, the problem is decomposed on a temporal basis into a set of overlapping slices involving fewer trains. The problem for each slice is optimized using a branch-and-bound procedure in which the linear relaxations are solved by column generation. Computational experiments carried out on real-life data from CN North America yielded an improvement of more than 7% over the solution used by the company.

A review of recent discrete optimization models for public rail transport planning with an emphasis on line planning and train scheduling was prepared by BUSSIECK *et al.* (1997). Also, CORDEAU *et al.* (1998c) provide a more comprehensive but less technical survey of optimization models for train routing and scheduling.

Contribution. In this paper, we describe the model and the heuristic solution approach based on mathematical optimization that we implemented at VIA Rail in Canada to solve the equipment assignment problem. Although they were developed with the specific needs of this railway in mind, they have a certain degree of generality and could certainly be adapted to several other railways. Given a weekly train schedule, a description of the physical network and a list of the available stock of locomotives and cars, our method determines a near-optimal assignment of equipment to trains in the form of a set of cycles which satisfy a large variety of operational constraints. The approach is based on a multi-commodity network flow formulation

which is optimized through a branch-and-bound method in which the relaxations are solved with a Dantzig-Wolfe decomposition. This approach is flexible and facilitates the introduction of maintenance constraints which represent a major difficulty of the problem. In CORDEAU *et al.* (1998b), the authors argued that a straightforward implementation of Dantzig-Wolfe decomposition was not appropriate to solve their formulation because of the large size of the resulting master problem. Here, we propose several refinements which make the problem more tractable, and show that column generation can indeed be an effective solution approach.

Overview. The rest of the paper is organized as follows. In the next section, we describe the general equipment assignment problem in a general context similar to that of VIA Rail. We also introduce several concepts which are then used to formulate a basic mathematical model of the problem in Section 2.3. Section 2.4 presents a solution approach based on column generation for this basic model, while Section 2.5 introduces various extensions to the model and the required adaptations of the solution approach. Computational experiments are reported in Section 2.6 and conclusions are given in the last section.

2.2 Problem Description

The equipment assignment problem treated in this paper is usually solved every few months when the train schedule is updated and it thus belongs to the tactical level of planning. However, its optimization horizon is normally shorter and corresponds to the period of the train schedule which is often a week. Since we are looking for a cyclic solution that will repeat period after period, the problem can be appropriately called the *tactical periodic equipment assignment problem.*

We consider a railway that operates locomotives and cars of different types. While locomotives all serve the same purpose, passenger cars come in different flavors: railways typically use a mix of club (first-class) and coach (second-class) cars. Besides its nature, the most important characteristics which distinguish an equipment type from another are its capacity and its operating speed. For a locomotive, the capacity is measured by the number of cars it can pull, whereas for a car, it is measured by its seating capacity. Since we are considering a tactical planning problem with an horizon of a few months, the fleet can be considered as fixed. Hence, the number of units of each type which are available is assumed to be known. This number can nevertheless vary from day to day to account for maintenance activities and other restrictions. Also, for each type of equipment, we are given per mile costs associated with fuel and maintenance. These operating costs are variable since they are related to mileage and not to equipment ownership. The cost of using one unit of equipment on a given train can then be computed as the distance between the origin and the destination stations times the total operating cost per mile.

The equipment types which are available to the railway can be combined in various ways to form train consists. Generally, a train consist contains one or two locomotives and a certain number of club and coach cars. Occasionally, additional baggage cars can also be part of a consist. The set of possible consist types is specified originally and the operating speed of each consist type is set to match that of its slowest component.

One of the basic input of the equipment assignment problem is a periodic schedule that specifies, for each train operated during the period, resource requirements and possible pairs of departure and arrival times. For a given train leg, these quantities are not unique but depend instead on the type of consist that will be used to ensure service on that leg. Indeed, different consist types may have different operating speeds and the railway must take this into consideration. Demand on each train leg

is usually given in terms of the number of first-class and second-class passengers. It can alternatively be given as a number of club and coach cars. In the latter case, the requirements may vary depending on the consist type used since cars of different types do not necessarily have the same seating capacity.

Since the solution to the equipment assignment problem is a set of cyclic equipment trips, its feasibility is in part determined by the connection possibilities at the various stations of the network. Hence, for each station, different durations must be known to determine the possible connections. In most cases, these durations will be dependent upon the respective orientation of the two successive train legs. Here, we assume that all trains belong to one of two orientations although this assumption can be easily relaxed. If one considers eastbound and westbound trains, then the *run-thru* time represents the minimum time needed to make a connection between two train legs that have the same orientation, while the *turn-around* time is the time needed to make a connection between two train legs that have opposite directions. These durations apply only when the train consist used on the first leg is also used unmodified on the second leg. If cars must be switched on or off the train at the intermediate station, a longer connection time is required. The necessary duration is given by the *switching time* which may depend on the respective orientation of the two trains. In some stations located at the end of a line, run-thrus may not be feasible. Also, switching may be restricted to some period of the day and may even be completely forbidden in certain stations. Finally, each station has a limited storage capacity determined by the available tracks and this capacity cannot be exceeded.

Normally, operating rules set forth by transport authorities stipulate that maintenance and inspection must be performed on each unit of equipment at a regular interval. Hence, every equipment cycle must include periodic stops at one of the stations associated with maintenance centers. Furthermore, these stops must be long

enough to allow for maintenance and minor repairs to be performed. When all trains operate during the day, maintenance can sometimes be restricted to be performed exclusively at night. In that case, the duration of the stop at the maintenance center will always be sufficient to permit maintenance.

In short, the basic equipment assignment problem consists in finding a minimum-cost set of equipment cycles which ensure that sufficient seating capacity (per class) is supplied on each train while satisfying constraints on minimum connection times, locomotive pulling capacity, equipment availability, storage capacity and maintenance requirements.

2.3 Mathematical Model

We now describe a basic mathematical model that integrates the most essential ingredients of the locomotive and car assignment problem. This model, which is based on a multi-commodity network flow structure with linking constraints, is a special case of the unified framework for deterministic time constrained vehicle routing and crew scheduling problems proposed by DESAULNIERS *et al.* (1998).

To simplify the notation and the statement of the model, it is first assumed that each train leg can be covered by a unique consist type. Also, all trains operate during the day and there is a single maintenance center where all equipment trips must start and end. Extensions to more complex situations will be discussed in Section 2.5.

Railways operate equipment units of different natures which can be grouped into a set of classes according to their respective roles. For example, locomotives, club

cars and coach cars are typical classes used by most railways. In each class, different makes or models of equipment can also be operated. Given this partitioning, let K be the set of *equipment types* where each type $k \in K$ corresponds to a particular class and a make available to the railway. If different equipment makes in the same class can be considered as identical with regard to their operating characteristics, they can be treated as a single equipment type. Let R be the set of *consist types*. Each consist type $r \in R$ is a set $\{k_1^r, k_2^r, \dots\}$ of compatible equipment types containing at least a locomotive type and a car type. Let L be the set of *train legs*. Each train leg $l \in L$ is defined by origin and destination stations, departure and arrival times, and resource requirements. Since it is assumed that each train leg l can be covered by a single consist type, let $r_l \in R$ denote this type. Then, for each equipment type $k \in r_l$, resource requirements can be specified as the minimum number of units of equipment, n_l^k, which are required on leg l.

Consider an ordered pair of train legs (l_i, l_j). These two legs can be successively covered by the same physical train consist if *(i)* $r_{l_i} = r_{l_j}$; *(ii)* the destination station of leg l_i is the origin station of leg l_j; and *(iii)* the connection time between the two legs is sufficient. A modeling difficulty appears when the minimum connection time in a station depends on whether the physical train consist is to be modified between the two consecutive legs. Since switching cars on or off a train consist requires a certain amount of time, the minimum connection time is normally greater when such work has to be performed. To take this into consideration in our model, we define a train *sequence* as an ordered set of train legs such that these train legs can be covered by the same train consist only if the consist is not modified at any intermediate station. For notational convenience, a sequence may contain a single train leg. Let S be the set of train sequences. The importance of this concept, which was introduced by CORDEAU *et al.* (1998b), becomes more apparent when considering the network representation.

2.3.1 Network Representation

For each equipment type $k \in K$, we define a time-space network structure $G^k = (N^k, A^k)$ where N^k is the node set and A^k is the arc set. Since a periodic solution is sought, all networks are cyclic. A small portion of such a network is presented in Figures 2.1 and 2.2. The graphical representation has been separated in two parts because of the large number of different arc types: both figures show the same set of nodes but a different subset of arcs. The figures represent the train departures and arrivals taking place in three stations of the network during day 2 of the planning period.

Nodes. The set N^k is composed of nine types of nodes. The first four types serve to represent the start and the end of each day in the planning period. At the station associated with the maintenance center (designated by Montréal in the figures), *source* and *sink* nodes represent, respectively, the start and the end of an equipment trip on that day. For all other stations of the network, *start-of-day* (SOD) and *end-of-day* (EOD) nodes are used to represent the corresponding moment of each day.

The next five types of nodes are associated with actual train movements. For each train sequence on which equipment of type k must be used, *departure opportunity* (OPP), *departure* (DEP) and *arrival* (ARR) nodes are defined. In addition, *run-thru* (RT) and *turn-around* (TA) nodes may be defined to represent the end of the corresponding activity after the arrival of the train. For every sequence, at least one of these last two nodes must be defined if the sequence can be followed by another sequence in the same day. However, if the destination station of the sequence is located at the end of a line in the network, run-thrus may be impossible. As shall be explained later, the role of OPP nodes is to simplify the introduction of station storage capacity constraints in the model.

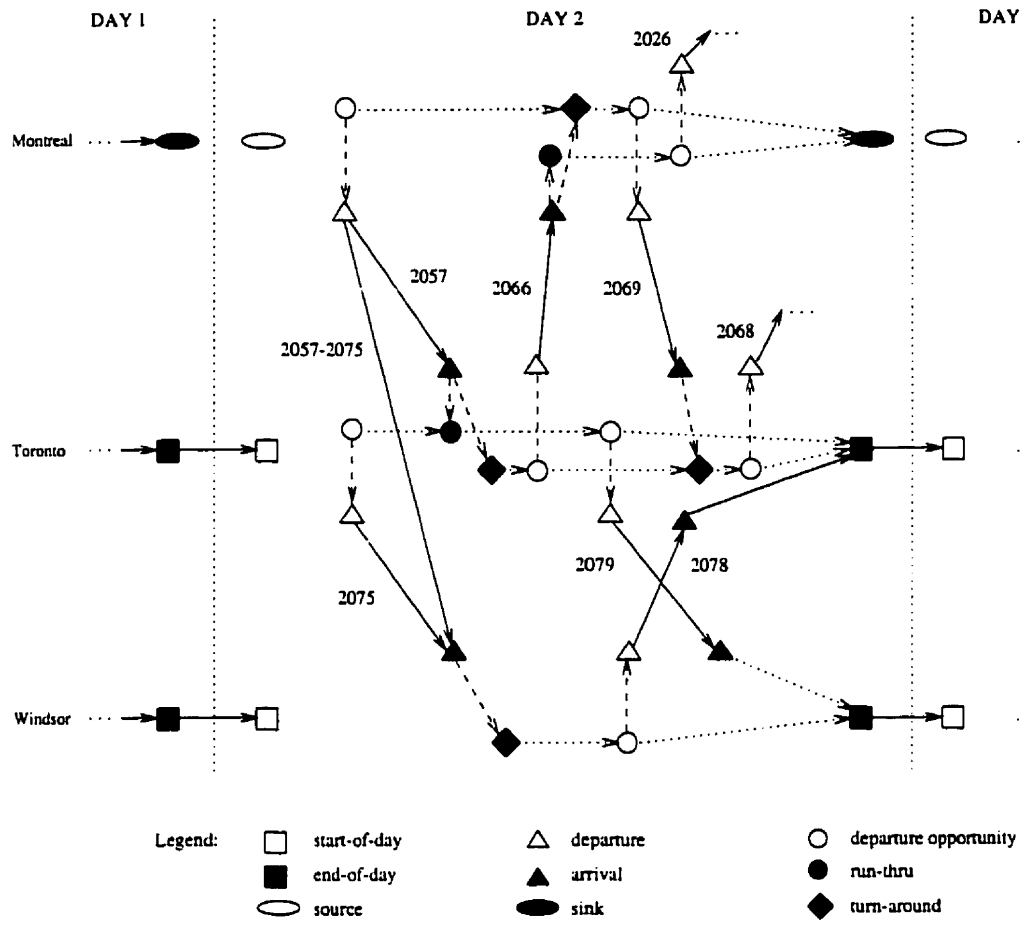

Figure 2.1: Portion of network G^k for equipment type k (part 1)

The time associated with an OPP node is the actual departure time of the first train leg in the sequence. On the other hand, the time associated with a RT node or TA node is the arrival time plus the corresponding switching time. This is where the notion of sequence comes into play. Consider train legs 2057 and 2075 represented in Figure 2.1. These two legs can be covered by the same train consist since the destination station of the first is the origin station of the second and the connection time between the two is sufficient. However, since the time between the arrival of the first leg and the departure of the second is small, it is not possible to modify the consist between the two legs. This is represented in the figure by the fact that

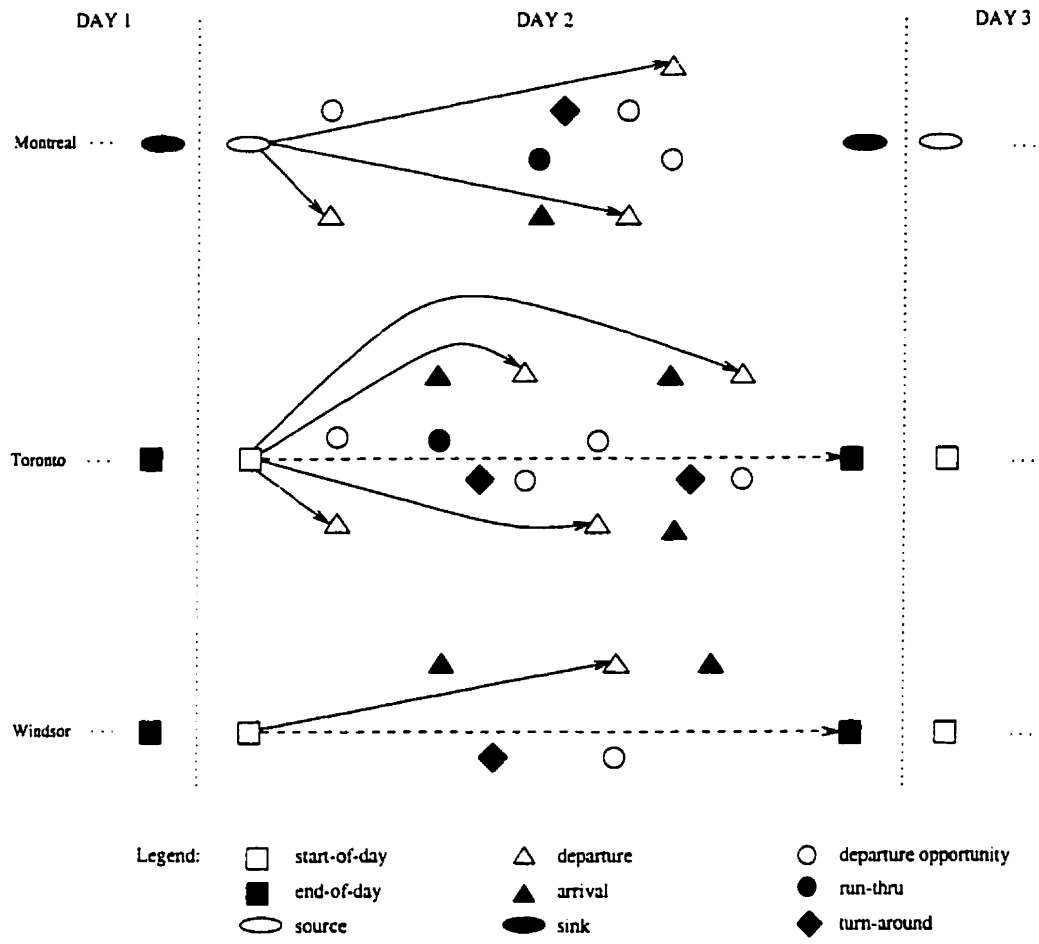

Figure 2.2: Portion of network G^k for equipment type k (part 2)

the RT and TA nodes for leg 2057 are located on the right of the OPP node for leg 2075. Hence, if one wishes to switch a car on or off the train consist used on leg 2057, the equipment will not be available to cover leg 2075. However, the two legs can be covered by the same consist if the equipment goes directly from the DEP node of leg 2057 to the ARR node of leg 2075. Without the notion of train sequence, it would be very difficult to impose the condition that a car cannot be switched on or off the train consist if the same locomotive covers both train legs. Indeed, since locomotives and cars have their own individual network, it would be extremely complicated to make sure that a locomotive does not cover both legs 2057 and 2075 if one of the cars used

on leg 2057 is not used on leg 2075, or vice-versa. Using sequences, this condition is easily imposed by using constraints stating that either legs 2057 and 2075 are covered by completely different units of equipment or they are covered by exactly the same units.

At each station and for each day of the period, OPP, RT and TA nodes are divided into two groups as follows. The first group contains OPP nodes associated with eastbound trains, RT nodes associated with eastbound trains and TA nodes associated with westbound trains. The second group contains all other nodes. Nodes within the same group are then sorted in chronological order and aggregated in case of equality.

Arcs. The arc set A^k is composed of eight types of arcs. For each station except the one associated with the maintenance center, there exists a *night* arc between each pair of consecutive EOD and SOD nodes. This arc represents a night stop at that station. For each station and each day, there is also a *wait* arc between each pair of consecutive OPP, RT and TA nodes that belong to the same group. In addition, wait arcs exist between the last node of each group and the EOD or sink node associated with the corresponding station and day. Such arcs are also present between ARR nodes associated with sequences that cannot be followed by another sequence on the same day and EOD or sink nodes.

For each train sequence on which equipment of type k can be used, there is a *sequence* arc that links the corresponding DEP and ARR nodes. Also, *run-thru* and *turn-around* arcs are defined between ARR nodes and associated RT and TA nodes. These arcs exist if the train consist covering the sequence can connect to another sequence on the same day. A *switching-on* arc is also defined between the corresponding OPP node and the DEP node. This arc represents the possibility of

switching cars before the departure. The additional time required for switching is however included in the RT and TA arcs.

For each train sequence, there is a *first-sequence* arc linking the SOD or source node of the day to the DEP node. The purpose of these arcs is to allow the proper computation of the number of units which stay idle in a station during a complete day. Since the SOD nodes are not linked by any *wait* arcs, idle units will have to flow on the *storage* arcs that link the SOD and EOD nodes of the same day in every station except at the maintenance center. These two types of arcs are illustrated in Figure 2.2. The purpose of OPP nodes is to allow train sequences to be covered by equipment that was not present in the station at the start of the day but has instead finished a sequence at the origin station during that day.

Finally, to appropriately impose maintenance constraints, the network structure just described is replicated to generate a set of overlapping subnetworks. Suppose that there are t days in the planning period and that every unit of equipment must be inspected at the unique maintenance center at least once every t_m days. Let $P = \{0, \ldots, t-1\}$ denote the set of days in the planning period. To impose these constraints, a subnetwork is created for each day $p \in P$. Subnetwork for day $p \in P$ has a single source node, which represents the beginning of an equipment trip on day p, but several sink nodes representing the end of an equipment trip on days $p, \ldots, (p + t_m) \mod t$. Therefore, $G^{kp} = (N^{kp}, A^{kp})$ will denote the subnetwork associated with equipment type k and day p. All feasible paths in this network will correspond to equipment trips leaving the maintenance center on day p and returning to the maintenance center at most t_m days later. The solution approach presented in Section 2.4 takes advantage of this subnetwork definition to implicitly impose maintenance constraints.

2.3.2 Mathematical Formulation

In order to give an integer programming formulation of the problem, some additional notation must be introduced. First, it is convenient to partition the set K of all equipment types into two subsets as follows. Let $K^C \subset K$ and $K^L = K \backslash K^C$ represent the subsets of equipment types corresponding to car types and to locomotive types, respectively. Let also V be the set of all stations represented in the network and denote the station associated with the maintenance center by the element $m \in V$. The set V is not required to contain all stations present in the physical rail network but only those at which car switching is allowed. It is assumed here that switching is allowed at the station m associated with the maintenance center.

For every equipment type $k \in K$ and every day $p \in P$, let $T^{kp} \subseteq A^{kp}$ be the set of arcs associated with train sequences and let $T_l^{kp} \subseteq T^{kp}$ be the subset of arcs associated with train sequences containing train leg $l \in L$. Let $E_q^{kp} \subseteq A^{kp}$ and $B_q^{kp} \subseteq A^{kp}$ be the sets of arcs directed into the sink node for day q (ending arcs) and out of the source node for day q (beginning arcs), respectively. Let also C_q^{kp} denote the set of arcs that are either directed out of the source node on day q or into a SOD node on day q. For given $k \in K$ and $p \in P$, let $\tilde{N}^{kp} \subset N^{kp}$ be the subset of nodes that excludes only source and sink nodes from the set N^{kp}. Then, for every node $n \in \tilde{N}^{kp}$, let the sets $I_n^{kp} \subseteq A^{kp}$ and $O_n^{kp} \subseteq A^{kp}$ contain all arcs that are directed in and out of node n, respectively.

Two types of decision variables are used in the formulation. For every equipment type $k \in K$, every day $p \in P$, and every arc $a \in A^{kp}$, let X_a be a non-negative integer variable representing the flow on arc a. For every equipment type $k \in K$, and every day $q \in P$, let Y_q^k be a positive integer variable representing the flow of equipment of

type k on the storage arc for day q at the maintenance center. The basic locomotive and car assignment problem can then be written as follows:

$$\text{Minimize} \quad \sum_{k\in K}\sum_{p\in P}\sum_{a\in A^{kp}} c_a X_a \tag{2.1}$$

subject to

$$\sum_{p\in P}\sum_{a\in T_l^{kp}} X_a \geq n_l^k \qquad (k\in K; l\in L) \tag{2.2}$$

$$\sum_{k\in K^C}\sum_{p\in P}\sum_{a\in T^{kp}} f_a^s X_a - \sum_{k\in K^L}\sum_{p\in P}\sum_{a\in T^{kp}} z_s f_a^s X_a \leq 0 \qquad (s\in S) \tag{2.3}$$

$$\sum_{p\in P}\sum_{a\in B_{q+1}^{kp}} X_a + Y_q^k - Y_{q+1}^k = 0 \qquad (k\in K; q\in P) \tag{2.4}$$

$$\sum_{p\in P}\sum_{a\in E_q^{kp}} X_a - \sum_{p\in P}\sum_{a\in C_0^{kp}} X_a + Y_0^k \leq w^k \qquad (k\in K) \tag{2.5}$$

$$Y_q^k \geq w^k - d_q^k \qquad (k\in K; q\in P) \tag{2.6}$$

$$\sum_{k\in K^C} Y_q^k \leq h_m \qquad (q\in P) \tag{2.7}$$

$$\sum_{k\in K^C}\sum_{p\in P}\sum_{a\in A^{kp}} g_a^{qv} X_a \leq h_v \qquad (v\in V\setminus\{m\}; q\in P) \tag{2.8}$$

$$\sum_{a\in I_n^{kp}} X_a - \sum_{a\in O_n^{kp}} X_a = 0 \qquad (k\in K; p\in P; n\in \bar{N}^{kp}) \tag{2.9}$$

$$X_a \geq 0, \text{ integer} \qquad (k\in K; p\in P; a\in A^{kp}) \tag{2.10}$$

$$Y_q^k \geq 0, \text{ integer} \qquad (k\in K; q\in P). \tag{2.11}$$

If c_a is the cost of using one unit of equipment on arc a, then the objective function (2.1) minimizes the sum of all operational costs. Given that n_l^k is the number of units of equipment k needed on leg l, constraints (2.2) ensure that enough units of each type of equipment are supplied on each leg. Constraints (2.3) translate locomotive pulling capacity limits, where f_a^s is a binary constant equal to 1 if and only if arc a is associated with sequence s, and z_s is the maximum number of cars which can be pulled by one locomotive over sequence s. Flow conservation between equipment trips is enforced by (2.4). Constraints (2.5) and (2.6) impose weekly and daily equipment availability, respectively. In these constraints, w^k is the maximum number of units of equipment k available at any time in the period, while d_q^k is the difference between w^k and the number of units available on day q. By letting h_v denote the car storage capacity in station v, storage capacity at the maintenance center is satisfied with constraints (2.7). Constraints (2.8) serve the same purpose for all other stations of the network; here, g_a^{qv} is a binary constant equal to 1 if arc a is the storage arc for station v on day q. Flow conservation along equipment trips is satisfied through constraints (2.9). Finally, constraints (2.10) and (2.11) require that each variable take a non-negative integer value.

2.4 Solution Methodology

The integer programming model (2.1)-(2.11) can be solved by a branch-and-bound algorithm where lower bounds are computed through a Dantzig-Wolfe decomposition (DANTZIG and WOLFE, 1960). In Section 2.4.1, we describe the application of this decomposition approach to our model, followed by an explanation of branching rules in Section 2.4.2.

2.4.1 Dantzig-Wolfe Decomposition

Model (2.1)-(2.11) has a block angular structure with linking constraints. Indeed, the objective function (2.1) and constraints (2.9) and (2.10) are separable by equipment type k and day p. One can thus take advantage of this structure by decomposing the model into a master problem and a set of subproblems. For any $k \in K$ and $p \in P$, consider the polyhedron defined by

$$\sum_{a \in I_n^{kp}} X_a - \sum_{a \in O_n^{kp}} X_a = 0 \qquad (n \in \tilde{N}^{kp}) \tag{2.12}$$

$$X_a \geq 0 \qquad (a \in A^{kp}). \tag{2.13}$$

This polyhedron has a unique extreme point, the null vector $\mathbf{0}$, but a potentially large number of extreme rays. Let Ω^{kp} represent the set of extreme rays of the polyhedron. Each extreme ray corresponds to a path in the graph G^{kp} from the unique source to a sink. Hence, any solution to constraints (2.9) and (2.10) for given k and p can be expressed as a non-negative combination of extreme rays chosen from Ω^{kp}. For every extreme ray $\omega \in \Omega^{kp}$ and for every arc $a \in A^{kp}$, let $x_{a\omega}$ be a binary constant equal to 1 if arc a is part of the path associated with extreme ray ω. For every $\omega \in \Omega^{kp}$, let also θ_ω be a non-negative variable. Then, for any $k \in K$, $p \in P$ and $a \in A^{kp}$, one can write

$$X_a = \sum_{\omega \in \Omega^{kp}} x_{a\omega} \theta_\omega.$$

By substituting the last expression in the original formulation (2.1)-(2.11), one obtains the following master problem:

$$\text{Minimize} \quad \sum_{k\in K}\sum_{p\in P}\sum_{a\in A^{kp}} c_a \sum_{\omega\in\Omega^{kp}} x_{a\omega}\theta_\omega \tag{2.14}$$

subject to

$$\sum_{p\in P}\sum_{a\in T_l^{kp}}\sum_{\omega\in\Omega^{kp}} x_{a\omega}\theta_\omega \geq n_l^k \qquad (k\in K; l\in L) \tag{2.15}$$

$$\sum_{k\in K^C}\sum_{p\in P}\sum_{a\in T^{kp}} f_a^s \sum_{\omega\in\Omega^{kp}} x_{a\omega}\theta_\omega - \sum_{k\in K^L}\sum_{p\in P}\sum_{a\in T^{kp}} z_s f_a^s \sum_{\omega\in\Omega^{kp}} x_{a\omega}\theta_\omega \leq 0 \qquad (s\in S) \tag{2.16}$$

$$\sum_{p\in P}\sum_{a\in E_q^{kp}}\sum_{\omega\in\Omega^{kp}} x_{a\omega}\theta_\omega + Y_q^k - \sum_{p\in P}\sum_{a\in B_{q+1}^{kp}}\sum_{\omega\in\Omega^{kp}} x_{a\omega}\theta_\omega - Y_{q+1}^k = 0 \qquad (k\in K; q\in P) \tag{2.17}$$

$$\sum_{p\in P}\sum_{a\in C_0^{kp}}\sum_{\omega\in\Omega^{kp}} x_{a\omega}\theta_\omega + Y_0^k \leq w^k \qquad (k\in K) \tag{2.18}$$

$$Y_q^k \geq w^k - d_q^k \qquad (k\in K; q\in P) \tag{2.19}$$

$$\sum_{k\in K^C} Y_q^k \leq h_m \qquad (q\in P) \tag{2.20}$$

$$\sum_{k\in K^C}\sum_{p\in P}\sum_{a\in A^{kp}} g_a^{qv} \sum_{\omega\in\Omega^{kp}} x_{a\omega}\theta_\omega \leq h_v \qquad (v\in V\setminus\{m\}; q\in P) \tag{2.21}$$

$$\sum_{\omega\in\Omega^{kp}} x_{a\omega}\theta_\omega \geq 0, \text{ integer} \qquad (k\in K; p\in P; a\in A^{kp}) \tag{2.22}$$

$$Y_q^k \geq 0, \text{ integer} \qquad (k\in K; q\in P) \tag{2.23}$$

$$\theta_\omega \geq 0 \qquad (k\in K; p\in P; \omega\in\Omega^{kp}). \tag{2.24}$$

The last model is obtained by applying the Dantzig-Wolfe decomposition principle to (2.1)-(2.11) while keeping constraints (2.9) and (2.10) in the subproblem. Since the only extreme point of the subproblem polyhedron is the null vector, removing the usual convexity constraint from the master problem does not affect its feasible region.

Given the potentially large size of the set Ω^{kp} for each equipment type $k \in K$ and each day $p \in P$, model (2.14)-(2.24) can be solved by a branch-and-bound method in which the linear relaxation lower bounds are computed by using a column generation approach. To this purpose, a relaxed master problem is obtained by replacing the set Ω^{kp} by the subset $\Omega^{kp}_\tau \subseteq \Omega^{kp}$ $(\tau = 0, 1, \dots)$ of extreme rays available at iteration τ of the column generation process. Since the subproblem decomposes into a set of flow conservation constraints for each k and p, new columns (i.e., extreme rays) are then generated for the relaxed master problem by solving a shortest-path problem in each network G^{kp}. In fact, extreme rays can be characterized by sending one unit of flow from the source to any of the sinks in the network. Arc costs are modified from one iteration to the next to reflect the new values of the dual variables associated with the constraints of the relaxed master problem. This process continues until no further negative-cost path can be identified in any of the networks.

Let $\boldsymbol{\alpha} = (\alpha^k_l \geq 0 \,|\, k \in K; l \in L)$, $\boldsymbol{\beta} = (\beta_s \leq 0 \,|\, s \in S)$, $\boldsymbol{\gamma} = (\gamma^k_q \,|\, k \in K; q \in P)$, $\boldsymbol{\delta} = (\delta^k \leq 0 \,|\, k \in K)$ and $\boldsymbol{\phi} = (\phi^v_q \leq 0 \,|\, v \in V \setminus \{m\}; q \in P)$ be the dual variables associated respectively with constraints (2.15)-(2.18) and (2.21) of the LP relaxation of (2.14)-(2.24). Constraints (2.19) and (2.20) need not be dualized since they involve only the Y^k_q variables. For given k and p, the objective function of the subproblem becomes

$$\sum_{a \in A^{kp}} c_a X_a -$$

$$\sum_{l \in L} \sum_{a \in T^{kp}_l} \alpha^k_l X_a - I(k \in K^C) \sum_{s \in S} \sum_{a \in T^{kp}} f^s_a \beta_s X_a + I(k \in K^L) \sum_{s \in S} \sum_{a \in T^{kp}} z_s f^s_a \beta_s X_a -$$

$$\sum_{q \in P} \sum_{a \in E^{kp}_q} \gamma^k_q X_a + \sum_{q \in P} \sum_{a \in B^{kp}_{q+1}} \gamma^k_q X_a - \sum_{a \in C^{kp}_0} \delta^k X_a - I(k \in K^C) \sum_{v \in V \setminus \{m\}} \sum_{q \in P} \sum_{a \in A^{kp}} g^{qv}_a \phi^v_q X_a$$

where $I(\cdot)$ denotes the indicator function taking the value 1 if its argument is true, and the value 0 otherwise. By summing over all k and p, one obtains the objective

function derived by applying Lagrangian relaxation to model (2.1)-(2.11) and relaxing all constraints but (2.9) and (2.10). Since the subproblem has the integrality property, the optimal value of the Lagrangian dual problem is equal to the LP relaxation bound of both (2.1)-(2.11) and (2.14)-(2.24) (see, e.g., GEOFFRION, 1974).

2.4.2 Branching Rules

In order to determine a feasible integer solution to model (2.14)-(2.24), a heuristic branch-and-bound method is used. This method consists in a depth-first search with very limited backtracking possibilities: if at a given node of the tree the LP relaxation is feasible but its optimal solution contains variables with fractional values, one or two child nodes are created by applying one of the following branching criteria.

First, to accelerate the solution of the master problem, the locomotive pulling capacity constraints (2.16) are relaxed and generated dynamically when they are not satisfied by the current solution. At a given node of the branch-and-bound tree, all violated constraints are added at once to the LP relaxation to create the child node. Constraint generation has the greatest priority and is always applied first.

The second branching rule involves fixing the number of locomotives covering a given train leg. Without this type of branching, the integrality gap may increase rapidly and the solution may become infeasible in the last few levels of the branching tree because of insufficient equipment availability. Different criteria can be used to choose the train leg on which branching is to be performed. For the case where every train leg requires at most two locomotives, the following rules have proven to be quite effective. Let α and β be two thresholds such that $1 < \alpha < \beta < 2$. If there are train legs covered by a fractional number of locomotives lying between 1 and

α, then branching is performed sequentially on all of these. In addition, branching is also performed on the train leg with the largest fractional number of locomotives between α and β, if any. In all cases, branching is performed by adding the constraint that the number of locomotives be equal to 1 on the corresponding train leg. This branching rule is applied when all relaxed constraints are satisfied and at least one train leg is covered with a fractional number of locomotives between 1 and β. This type of branching is also the only one for which backtracking is allowed. If the relaxation becomes infeasible before any other type of branching is applied (excluding the dynamic generation of constraints), the search backtracks to the node at which the number of locomotives was fixed and explores the alternative branch obtained by increasing this number by one.

The third rule for branching consists in choosing a fractional path variable θ_ω associated with a locomotive type and setting the value of this variable equal to 1. When branching on a locomotive path variable, arcs can sometimes be eliminated from several networks. In fact, one can remove all arcs that represent a train sequence containing a train leg which is present in the locomotive path but in a different train sequence. This is not only true for the locomotive network but also for all networks associated with equipment types that are required on any of the train legs which are part of the fixed path. When applying this rule, the path with the largest fractional value is chosen.

The fourth type of decision involves specifying that a succession of two train legs (l_i, l_j) must be covered by the same locomotive. This decision can be treated directly in the subproblem by using a shortest-path algorithm that adds an additional dimension to each label (see, e.g., DUMAS *et al.*, 1991). Again, when such a rule is applied, all arcs that are incompatible with the decision can be removed from the corresponding networks. In this case, all arcs associated with sequences that cover

leg l_i followed by leg $l_k \neq l_j$ or cover leg $l_k \neq l_i$ followed by leg l_j can be eliminated as they contradict the decision. The last two methods are used whenever the first two cannot yield any decision. In that case, scores are used to determine which method should be applied and several decisions can be made at once. Also, branching can be performed on several path variables or leg successions at the same time provided that their flow is larger than a preset threshold.

The last branching rule consists in choosing a fractional path variable θ_ω associated with a car type and rounding up its value to the next integer. This type of branching is applied only when none of the four preceding rules can provide a decision.

2.5 Extensions

We now describe several extensions that were necessary to adapt the model of Section 2.3 to the actual problem at VIA Rail. These extensions are somewhat general and are likely to be required by other railways as well. Some of them are treated directly by adapting the networks or the objective function whereas others also require a modification of the solution approach. For reasons of clarity, all extensions are presented in an individual and independent manner although their combination poses no difficulty.

2.5.1 Substitutions Between Equipment Types

A simple yet important extension to the basic model is the ability to take substitution possibilities into account. For example, a club car can usually be used in place of

a coach car since the service level provided by the former type is superior to that of the latter. The reverse can also be allowed but a large penalty cost should be imposed to avoid such substitutions. Although not very frequent, substitutions help the railway to reduce unnecessary empty car movements and are particularly useful when equipment availability is restrictive.

For each type of equipment $k \in K$, let $J^k \subset K$ be the set of equipment types which can be substituted for type k. For notational convenience, assume that $k \notin J^k$. Also assume that a locomotive cannot be substituted for a car and vice-versa. Then, for each sequence arc $a \in T^{kp}$ and each $j \in J^k$, an additional arc must be added to A^{jp}. This arc will represent the substitution of equipment j for equipment k on the sequence associated with arc a. The cost associated with this arc should take into account not only operational expenses but also possible penalty costs.

Let $T^{jkp} \subset A^{jp}$ be the set of arcs corresponding to the substitution of equipment j for equipment k on train sequences. Let also $T_l^{jkp} \subseteq T^{jkp}$ be the subset of these arcs that are associated with sequences covering leg l. Let finally $\tilde{T}^{kp}$ be the set of all sequence arcs in network G^{kp}, including those corresponding to substitutions.

Given these definitions, only constraint sets (2.2) and (2.3) need to be modified to allow for substitutions. The rest of the model as well as the solution approach are not affected by the introduction of substitution possibilities. Constraints (2.2) and (2.3) now become

$$\sum_{p \in P} \sum_{a \in T_l^{kp}} X_a + \sum_{j \in J^k} \sum_{p \in P} \sum_{a \in T_l^{jkp}} X_a \geq n_l^k \qquad (k \in K; l \in L) \qquad (2.25)$$

$$\sum_{k \in K^C} \sum_{p \in P} \sum_{a \in \tilde{T}^{kp}} f_a^s X_a - \sum_{k \in K^L} \sum_{p \in P} \sum_{a \in \tilde{T}^{kp}} z_s f_a^s X_a \leq 0 \qquad (s \in S). \qquad (2.26)$$

2.5.2 Basic Consists

A distinguishing characteristic of the VIA Rail application is that some units of equipment are grouped together to form *basic consists*. For each consist type $r \in R$, a basic consist is formed by assembling a certain number of units of each required equipment type $k_1^r, k_2^r, \ldots$. The number of units of equipment k_i^r is usually chosen so as to match the minimum number required on any leg which must be covered by a consist of type r. For example, if all legs must be supplied with at least one locomotive, one club car and two coach cars, then a basic consist containing the corresponding number of units of each type may be defined. For each consist type $r \in R$ and each equipment type $k \in K$, let b_r^k be the number of units of equipment k needed in a basic consist of type r.

Using these compound equipment types has two main advantages from an operational standpoint. First, it helps to reduce the amount of car switching performed since all units in the same basic consist always remain together between two stops at a maintenance center. Second, it simplifies maintenance activities since all units in the same consist can be inspected at the same time. From an optimization point of view, basic consists are also very appealing since they help to reduce the size and the difficulty of the problem. Indeed, when the resource requirements of a train leg coincide with the units supplied by a basic consist, a single constraint can replace the set of all individual demand constraints per equipment type.

A disadvantage of using basic equipment consists is that more dead-heading movements may be performed if it is necessary to use a complete basic consist where a single extra car is needed. However, this will not happen if the demand for transportation is balanced as it is the case in our application. Finally, while substitutions between elementary (disaggregated) equipment types can still be

modeled as explained in the previous section, substitutions between equipment units in the same basic consist would require using a very large number of variables to take all possibilities into consideration.

To incorporate basic consists into formulation (2.1)-(2.11), we must first define a network G^{rp} for each $r \in R$ and $p \in P$. These networks are very similar to the networks G^{kp} defined previously for the elementary equipment types. For each of these networks, let T_l^{rp} be the subset of arcs associated with sequences covering train leg l. Constraints (2.2) can then be replaced by the following two groups of constraints:

$$\sum_{p \in P} \sum_{a \in T_l^{rlp}} X_a \geq 1 \qquad (l \in L) \tag{2.27}$$

$$\sum_{p \in P} \sum_{a \in T_l^{kp}} X_a + \sum_{p \in P} \sum_{a \in T_l^{rlp}} b_{r_l}^k X_a \geq n_l^k \qquad (k \in K; l \in L). \tag{2.28}$$

Constraints (2.27) ensure that at least one basic consist is supplied on each train leg while constraints (2.28) make sure that the total number of units of each type supplied to each leg (including the units of the basic consist) satisfies the demand. Then, whenever $n_l^k - b_{r_l}^k \leq 0$ for given $k \in K$ and $l \in L$, the corresponding constraint (2.28) can be removed from the model since the demand constraint is automatically satisfied by the units of the basic consist. Finally, constraints (2.3), (2.4), (2.5) and (2.8) must be modified to take into account the number of units of each equipment type which are present in a basic consist. For example, constraints (2.3) become

$$\sum_{k \in K^C} \left[\sum_{p \in P} \sum_{a \in T^{kp}} f_a^s X_a + \sum_{r \in R} \sum_{p \in P} \sum_{a \in T^{rp}} f_a^s b_r^k X_a \right] - \sum_{k \in K^L} \left[\sum_{p \in P} \sum_{a \in T^{kp}} z_s f_a^s X_a + \sum_{r \in R} \sum_{p \in P} \sum_{a \in T^{rp}} z_s f_a^s b_r^k X_a \right] \leq 0 \qquad (s \in S). \tag{2.29}$$

2.5.3 Daytime Maintenance

In our basic model, we assumed that all trains were run during the day and that maintenance operations were performed at night in the unique maintenance center. In fact, the model can easily be adapted to deal with many scenarios regarding maintenance constraints.

We first consider the possibility of performing maintenance during the day if sufficient time is available during a connection at the station associated with the maintenance center. To incorporate this additional possibility into our model, we define supplementary arcs that represent the maintenance activity. These additions are illustrated in Figure 2.3. Several types of arcs have been omitted from the figure for reasons of clarity.

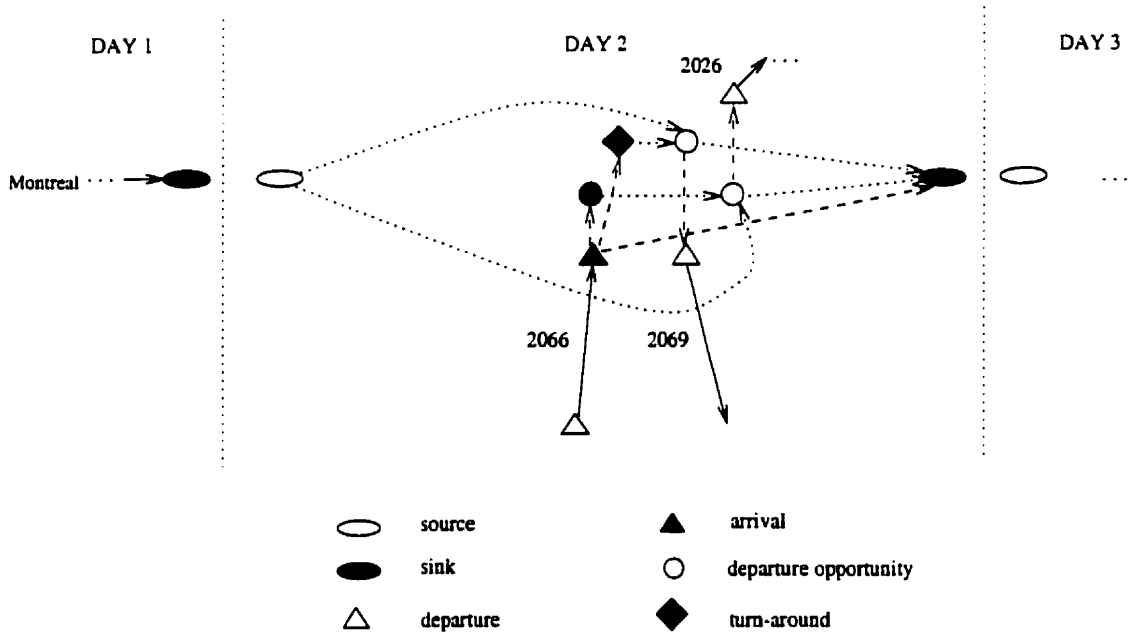

Figure 2.3: Additional arcs used for daytime maintenance

After the arrival of a train at the station associated with the maintenance center, the equipment can go directly to maintenance. This activity is represented by a special arc whose head is the sink node for the corresponding day. This arc is defined only when the time between the train arrival and the latest departure on the same day is greater than or equal to the minimum time needed for maintenance. Also, an arc links the source node on the same day to the departure opportunity node corresponding to the first possible departure in each direction after the maintenance. These arcs do not however contribute to constraints (2.4). Instead, an additional flow conservation constraint is added to the model to make sure that the flow on the first arc is equal to the sum of the flows on the other two arcs. Such a constraint is needed for each train leg after which daytime maintenance can be performed and for each type of equipment that is required on that leg.

In Figure 2.3 for example, daytime maintenance can be performed after the arrival of train 2066 since there will be enough time to connect with either train 2026 or train 2069. Hence, there is an arc that goes directly from the arrival node associated with that leg to the sink node for day 2. Also, there are arcs that link the source node for day 2 to the departure opportunity nodes for the two trains 2026 and 2069.

This approach also provides a way to relax the assumption that all trains operate during the day. For any train leg that has the maintenance center as its destination, arcs would link the arrival node to the sink node of the day during which the arrival takes place. Also, arcs would link the source node for the day on which the equipment will become available again after having been maintained to the departure opportunity node for the first train leg in each orientation that can be covered after maintenance. Finally, a flow conservation constraint would be needed for each such group of arcs to make sure that units of equipment leaving the source node on these arcs have indeed been maintained.

Obviously, this approach introduces an error in enforcing maintenance constraints since a unit of equipment which leaves the maintenance center at 11:00 at night will be considered to have been used during the complete day. To get a better approximation, source and sink nodes can be replicated and associated with shorter time periods. For example, instead of using sources and sinks for each day, one could use equivalent nodes for periods of 6 hours. The drawback is that 28 subnetworks per equipment type would be necessary in a one-week period instead of 7.

Finally, if more than one maintenance center is available to the railway, the model can be adapted with little effort to deal with this situation. Source and sink nodes must be defined for each station where maintenance can be performed. In each network, equipment trips can then begin and terminate at any of the source and sink nodes, respectively. Also, (2.4) must be replicated for each maintenance center.

2.5.4 Minimizing Switching Operations

Another important extension which must be considered is the minimization of switching operations. This objective often conflicts with the minimization of equipment circulation costs since reducing the total number of switchings generally has the effect of increasing empty car movements.

Recall that a train sequence is a series of consecutive train legs such that if these train legs are to be covered by the same train consist, then this consist cannot be modified at any point from the origin of the sequence to its destination. Hence, no switching occurs between legs that are covered in the same sequence. On the other hand, consider two legs such that there is sufficient time to perform switching between the arrival of the first leg and the departure of the second leg. Given the definition of train sequences, these two legs cannot be part of the same sequence. This is not to

say that switching will necessarily be performed during the connection. It is indeed possible that the best solution is to use exactly the same equipment on both legs. The difficulty here is to be able to determine whether switching is taking place or not. One way to circumvent this difficulty would be to broaden the definition of a sequence to include series of consecutive train legs with no switching, even if switching is possible at certain places. This way, switching would be performed at both ends of a sequence or else the corresponding previous and next legs would be covered in an even longer sequence. The drawback of this approach is obvious: the number of possible sequences would grow out of proportion.

The approach that we propose consists in using a two-phase method. In the first phase, the model given previously is solved with the integrality requirements imposed only on locomotive flows. In the second phase, the locomotive equipment cycles determined in the first phase are held fixed and a modified network is used to minimize a weighted combination of circulation costs and switching costs. Since the locomotive trips are known, it is possible to determine easily if car switching does take place. For example, if a car used on a leg l_i is next used on leg l_j while the locomotive used on leg l_i is next used on leg l_k, then switching has occurred.

To introduce switching costs in the model, additional connection arcs are defined in the networks associated with car types. Suppose that two train sequences s_i and s_j are covered with the same locomotive in the solution of the first phase. Then, a connection arc with a cost of zero will link the arrival node of sequence s_i to the departure node of sequence s_j. In addition, the run-thru and turn-around arcs originating from the arrival node of sequence s_i will be given positive costs representing the cost of switching one unit of equipment. If any of these arcs is used, switching has necessarily occurred and will be accounted for in the objective function. Connection arcs are represented by bold dashed lines in Figure 2.4.

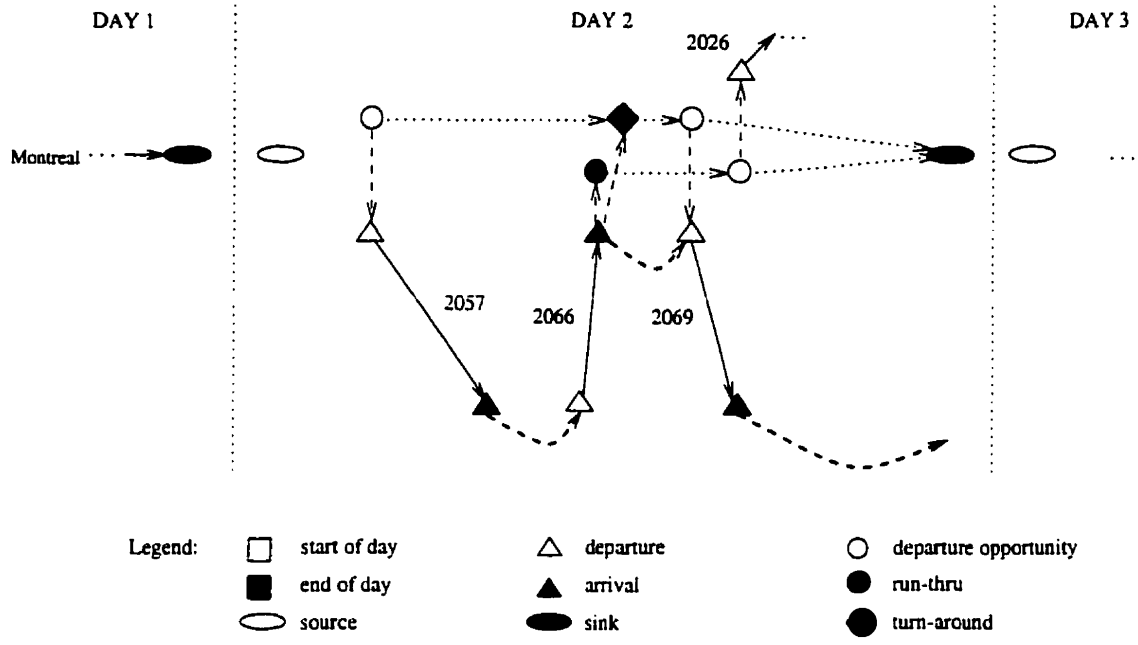

Figure 2.4: Additional arcs used for switching minimization

The optimization model for the second phase is very similar to the one used in the first phase. However, networks are no longer needed for locomotive types. Also, demand constraints (2.2) as well as equipment availability constraints (2.6) and (2.5), and flow conservation constraints (2.4) and (2.9) are only necessary for car types. Finally, pulling constraints (2.3) which previously linked car and locomotive types now only involve car types since the number of locomotives used on each sequence is known. The resulting model is then solved with the integrality requirements satisfied for all equipment types.

It is worth mentioning that the connection arcs could have been introduced directly in the original network representation for each type of equipment. By using constraints stating that a given arc in a car network cannot be used unless the corresponding arc is used in the locomotive network, switching costs could have been taken into account in the original formulation. However, there would be an

enormous number of such arcs and constraints. Indeed, one arc would be needed for each possible connection between two legs during the planning period. Since the additional constraints would appear in the master problem of the Dantzig-Wolfe decomposition, the model would become intractable. Hence, although it provides only a heuristic method, the two-phase approach seems to be an acceptable alternative in practice.

2.5.5 Choosing Between Consist Types

In a tactical planning model, the aim is to optimize the utilization of the stock of locomotives and cars given a fixed train schedule. Hence, it is reasonable to assume that the consist type used on each leg is held fixed. In long term planning however, it may be interesting to consider the possibility of choosing the combination of equipment used on certain train legs.

Assume that for train leg $l \in L$, one must choose a consist type from the subset $R_l \subseteq R$. To incorporate this possibility into our formulation, several steps are necessary. First, the networks associated with each type of equipment must be augmented. Instead of having a single arc for each possible sequence, we must now introduce one arc for each possible sequence and each possible consist type that can be used on the corresponding sequence. Thus, let T_{rl}^{kp} be the set of arcs associated with sequences covering leg l with a consist of type r. These different arcs are needed for two main reasons. First, different consist types may have different operating speeds which affect the arrival times. Second, we must distinguish between consist types so as to impose compatibility constraints.

For each equipment type $k \in K$ and each consist type $r \in R$, let e_r^k be a binary constant equal to 1 if and only if equipment k is included in a consist of type r. For

each leg $l \in L$ and each consist type $r \in R$, let Z_l^r be a binary variable equal to 1 if leg l is to be covered with a consist of type r. Finally, let u_l^k be an upper bound on the number of units of equipment k that may be used on leg l. Then, demand constraints (2.2) must be replaced with the following sets of constraints:

$$\sum_{r \in R_l} Z_l^r = 1 \qquad (l \in L) \tag{2.30}$$

$$\sum_{p \in P} \sum_{a \in T_{rl}^{kp}} X_a - n_l^k e_r^k Z_l^r \geq 0 \qquad (k \in K; l \in L; r \in R_l) \tag{2.31}$$

$$\sum_{p \in P} \sum_{a \in T_{rl}^{kp}} X_a - u_l^k e_r^k Z_l^r \leq 0 \qquad (k \in K; l \in L; r \in R_l). \tag{2.32}$$

Constraints (2.30) state that a single consist type must be chosen for each train leg while constraints (2.31) and (2.32) ensure that the number of units of equipment k used on leg l lies between n_l^k and u_l^k if equipment k is included in the chosen consist type, and is equal to zero otherwise. Since there may be a large number of constraints (2.32), these constraints may be originally relaxed and generated dynamically during the branch-and-bound search. Also, branching should first be performed on the Z_l^r variables.

2.6 Computational Experiments

Since the primary objective of this paper was to describe the development of a model and solution strategy for a real-life application, we chose to restrict computational experiments to the data available from VIA Rail. Nevertheless, we performed a rather large selection of tests to measure the performance of the algorithm and tune its parameters. We now describe the data used in the computational experiments,

followed by a brief summary of results and a comparison with the solutions used by VIA Rail.

2.6.1 Description of Data

The data used in all computational experiments concern the trains operated by VIA Rail in the Québec-Windsor corridor. More than 325 trains are run weekly in accordance with a schedule that is revised on a seasonal basis to reflect changes in the demand. These trains, which link together the most important cities in the provinces of Québec and Ontario, all operate during the day.

The physical rail network considered in this application is composed of nine primary stations and each train leg originates and terminates in one of these stations. The physical network also has a large number of secondary stations at which passengers can get on or off a train but where no train consist modifications can take place. Hence, only primary stations need to be considered in the model. For each primary station, minimum run-thru, turn-around, and switching times are known. These durations generally vary from 30 minutes to a few hours. In each instance solved, car switching is permitted in either two or three stations. For each station, a storage limit also specifies the maximum number of cars that may be stored at any time in that station.

Six types of equipment are used by the company to ensure service on the corridor: two types of engines (LRC and F40) and two types of club and coach cars (LRC and HEP). These equipments can be combined in three different ways to create consists types with different operating speeds. Combining an LRC locomotive with LRC cars yields a consist with a maximum operating speed of 100 mph while combining an F40

locomotive with LRC and HEP cars yield consist types with maximum operating speeds of 95 mph and 90 mph, respectively. For each type of equipment, daily availabilities are known and must be strictly respected. These availabilities, which vary only slightly from one day to the next, are rather restrictive in the case of locomotives but do not lead to particularly tight constraints in the case of cars. The complete fleet is composed of over 130 units of equipment.

For each train leg, demand is expressed as the minimum number of club cars and coach cars needed on that leg. This demand must be satisfied and can be exceeded provided that locomotive pulling constraints are satisfied. The total number of cars needed on a train normally lies between 3 and 8 whereas the pulling capacity of an engine varies from 5 to 8 cars depending on different factors such as locomotive make and the physical characteristics of the train segment. Most trains require a single locomotive but a few exceptions may require two. Basic consist types are defined for each of the three consist types enumerated above. For example, a basic consist for F40 locomotives and LRC cars contains one locomotive, one club car and two coach cars. This basic consist contains the minimum requirements for any train to be covered with a consist of that type.

A single maintenance center is used by VIA Rail. This maintenance center is located in Montréal and is of course associated with a station in which car switching is allowed. Maintenance must be performed at least once a week on every unit of equipment. Since all trains are operated during the day, most maintenance activities take place at night. However, daytime maintenance is also allowed provided that the time for connection is at least 5 hours.

The primary objective considered in all computational experiments is to minimize the sum of operational costs associated with total car miles. However, a secondary

objective of minimizing the number of switchings is also taken into account by introducing penalties in a modified objective function (as explained in Section 2.5.4).

Three different data sets were provided by VIA Rail, each corresponding to the schedule used in a different season. From these three data sets, six instances were derived: in the first group of instances (instances 1 to 3), the type of consist used on each train leg is chosen to match the assignment that was used by VIA Rail. The last group of instances (instances 4 to 6) is similar to the first group but incorporates the possibility of choosing the consist type to be used on certain train legs. In this case, a choice must be made between two possible consist types for approximately 30% of all train legs. The number of train legs in each instance varies from 326 to 348. Finally, three scenarios were considered for each instance by varying the importance given to the minimization of switching operations. In the first scenario, switching is barely penalized whereas in the second, a moderate penalty is imposed. In the third scenario, switching penalties are calibrated so as to produce a solution with a number of switchings smaller than or equal to the number of switchings in the solution used by VIA Rail.

2.6.2 Computational Results

All instances were solved with the approach presented in Section 2.4 using an adaptation of the GENCOL[1] software. This adaptation was required to incorporate the various extensions and to implement the branching strategies described in Section 2.4.2. All experiments were performed on a Sun Ultra 2 computer (300 MHz).

[1]GENCOL is an optimization software based on column generation that was developed at GERAD in Montréal.

Phase I results

As explained in Section 2.5.4, a two-phase approach is used to minimize switching. If two instances differ only by the importance given to the minimization of switching, the solution to the first phase is thus common to the two instances. Table 2.1 reports the number of branch-and-bound nodes, the number of relaxed constraints generated dynamically during the search, the total number of iterations of the master problem for column generation, and the phase I gap for each instance. Recall that in the first phase, integrality constraints are imposed only on locomotive variables. Hence, the gap is the relative difference between the cost of the (partially) integer solution and the cost of the initial relaxation. This gap may be larger than the actual integrality gap since the initial relaxation does not contain all the constraints of the LP relaxation of the problem (some of them being generated dynamically).

Table 2.1: Results of Phase I Optimization

Instance	Train legs	BB nodes	Dynamic constraints	MP iterations	Phase I gap	CPU time (hours)
1	326	249	370	3 006	1.8%	1.30
2	348	308	522	5 313	2.7%	3.33
3	347	155	359	3 481	1.3%	2.72
4	326	291	404	5 450	1.6%	4.73
5	348	412	740	10 499	3.2%	14.60
6	347	225	355	4 343	1.5%	4.32

For example, the search tree for the first instance contained 249 nodes and 370 constraints were generated during the exploration of the tree. Solving the linear relaxations at these nodes with column generation required a total of 3006 iterations of the column generation process. The problem was solved in 1.3 hours and the relative integrality gap was 1.8%. As expected, solving each of the last three instances required more efforts than did solving their corresponding counterpart in which the consist type assignment was fixed. In particular, instance 5 required more than 14

hours of computation. However, it is worth recalling that this is a planning problem that is solved only once every two or three months.

Phase II results

Solving the second phase model is, however, much faster. This process rarely requires more than 5 seconds of CPU as very few nodes must be explored before an optimal integer solution is found. Table 2.2 reports the cost of the solutions obtained by varying the weight assigned to switching minimization. For confidentiality reasons, we do not directly report the cost of the solution but rather express it as a percentage of the cost of the solution used by VIA Rail. Hence, the total cost column indicates the ratio of the cost of the solution produced by the algorithm over the cost of the solution produced manually by VIA personnel.

Table 2.2: Comparisons with solutions from VIA (fixed consist types)

Instance	Total cost	Variable cost	Number of switchings	Total gap
1a	97.3%	55.1%	36 (138.5%)	4.6%
1b	98.2%	69.6%	25 (96.2%)	5.5%
1c	99.1%	84.4%	19 (73.1%)	6.3%
2a	98.3%	71.7%	40 (181.8%)	5.0%
2b	99.4%	90.5%	29 (131.8%)	6.0%
2c	99.9%	99.1%	21 (95.5%)	6.5%
3a	97.3%	56.7%	35 (166.7%)	4.3%
3b	99.1%	85.5%	27 (128.6%)	5.9%
3c	99.8%	96.5%	21 (100.0%)	6.5%

The cost reduction may appear to be small at first sight. However, it must be emphasized that a large portion of the total cost is in fact a fixed cost for supplying each train with the minimum number of units of each type of equipment. Since demand constraints must be satisfied, this cost cannot be reduced. Thus, the part

of the cost that can actually be reduced by an improved planning is the variable cost associated with non-productive (or dead-heading) movements of cars supplied in excess of the minimum requirements of each train. The variable cost column of Table 2.2 expresses the variable cost in the computed solution as a percentage of the variable cost in the solution of VIA Rail. For example, while the total cost decreased by only 2.7% in scenario 1a, the variable cost decreased by 44.9%.

Of course, one way to decrease the total number of miles traveled by inactive cars is to detach unnecessary cars before each train leg when this is possible and to re-attach them as needed. Hence, we also report the total number of times that switching was performed. In scenario 1a, cars were switched on or off a train 36 times whereas this number was only 26 in the solution used by VIA Rail (an increase of 38.5%). On the other hand, the solutions obtained for scenarios 1b and 1c dominate the solution of VIA since they reduce both the variable cost and the number of switchings.

Because our approach is aimed at minimizing operating costs and considers switching minimization as a secondary (and less important) objective, it is sometimes difficult to obtain solutions that improve both objectives concurrently. For example, the solution for instance 3 which had the lowest number of switchings could only reduce variable costs by 3.5%. However, considering the very significant savings which are associated with variable cost reduction, a good strategy is to choose a solution with the least total cost given that the number of switchings does not exceed a chosen threshold. Since reducing the variable costs by a significant margin translates into annual savings of hundreds of thousands of dollars, these savings more than offset the cost associated with performing a few additional switchings.

Finally, the last column of Table 2.2 indicates the relative difference between the cost of the final phase II integer solution and the cost of the initial phase I

relaxation. One observes that this gap (which does not take switching penalties into consideration) grows moderately as the weight attributed to switching minimization increases. Again, this is a result of the fact that no consideration is given to the concept of switching in the first phase. Hence, as the importance given to switching minimization increases, more unproductive movements are required, leading to larger variable costs.

In the last group of computational experiments, we used the same scenarios but applied them to the instances in which a choice must be made between two consist types for certain train legs. Table 2.3 summarizes the results of these tests. Since we did not have comparable solutions produced by VIA Rail, we could no longer compare the variable costs as we did for the first three instances. Instead, we computed the reduction in the fixed cost that resulted from a better assignment of consist types to the train legs.

Table 2.3: Comparisons with solutions from VIA (variable consist types)

Instance	Total cost	Fixed cost	Number of switchings	Total gap
4a	96.2%	98.7%	33 (126.9%)	4.8%
4b	97.6%	98.7%	21 (80.8%)	6.1%
4c	98.2%	98.7%	19 (73.1%)	6.7%
5a	98.2%	98.8%	39 (177.3%)	5.9%
5b	99.4%	98.8%	28 (127.3%)	6.9%
5c	100.1%	98.8%	22 (100.0%)	7.6%
6a	96.8%	98.9%	34 (161.9%)	4.3%
6b	98.1%	98.9%	26 (118.2%)	5.6%
6c	99.5%	98.9%	21 (100.0%)	6.9%

For example, the solution for scenario 4a yielded a 3.8% reduction of the total cost. This reduction, which is larger than the reduction of 2.7% yielded by the solution to instance 1a, is in part possible because of a reduction of 1.3% in the fixed cost. As

before, we also tried to produce a solution that matched or improved the number of switchings used by VIA Rail with a smaller cost. Whereas this was possible for instances 4 and 6, we could not obtain such a solution for instance 5. The best solution that we could obtain had a cost that exceeded that of the VIA Rail solution by 0.1%. This result, which is somewhat surprising given that instance 5 is a relaxation of instance 2, can be explained by the fact that we are using a two-phase method with heuristic branching.

2.7 Conclusions

We have proposed a formulation and a solution method for a real-life application of the locomotive and car assignment problem in the context of rail passenger transportation in North America. The basic model captures the fundamental difficulties of the problem and is also flexible in the sense that it can be customized to deal with many additional situations. Several extensions that are needed to make the model useful in practice have been discussed. The algorithm, which has been successfully implemented at VIA Rail, finds good quality solutions in a few hours of computing time. This performance is satisfactory given the fact that the model need only be solved once every few months. The model can also be used to evaluate and compare different scenarios. For example, VIA Rail could find a solution that required one less locomotive on one of the data sets, thus realizing potential savings of 400,000 $ annually. To obtain a valuable tool for performing "what-if" analysis, a faster solution approach would however be required. Also, a different model and solution approach would be necessary to deal with the daily operations problem in which more details, such as car positioning and orientation, must be considered explicitly. These areas of research will be addressed in subsequent papers.

Acknowledgments

We wish to thank Mrs. Francine Hébert, Mr. Alain Vigeant and Mrs. Natali Gagnon of VIA Rail for their valuable collaboration from the initial problem definition to the final software implementation. This work was supported by the Québec Government (Fonds pour la Formation de Chercheurs et l'Aide à la Recherche) and by a joint research project of the Natural Sciences and Engineering Research Council of Canada and AD OPT Inc.

Chapitre 3

A Benders Decomposition Approach for the Locomotive and Car Assignment Problem

Article écrit par Jean-François Cordeau, François Soumis et Jacques Desrosiers; accepté pour publication dans *Transportation Science* en 1999.

La principale faiblesse du modèle présenté au chapitre précédent est que le temps nécessaire à sa résolution est très fortement lié au nombre de contraintes dans le problème maître de la décomposition de Dantzig-Wolfe. Lorsque ce nombre excède 1500 ou 2000, le temps de calcul devient rapidement excessif. Or, c'est ce qui se produit si l'instance à résoudre comporte plus de 500 trains ou encore si la combinaison d'équipement utilisée sur chaque train n'est pas fixée *a priori*. C'est également ce qui se produit si l'instance à résoudre comporte plusieurs centres d'entretien et que l'entretien peut être effectué en tout temps après l'arrivée d'un train à l'un de ces centres.

Pour plusieurs transporteurs, il est important de pouvoir considérer plusieurs possibilités quant à la combinaison d'équipement utilisée sur chaque train. Compte tenu des coûts d'acquisition et d'entretien élevés des locomotives et des wagons, cette

flexibilité accrue permet souvent de réaliser des économies substantielles en obtenant une meilleure utilisation de l'équipement. De plus, de nombreuses entreprises opèrent des trains à la fois pendant le jour et la nuit et utilisent plusieurs centres d'entretien répartis dans le réseau. Afin de résoudre des instances de grande taille dans un tel contexte, il est clair qu'une approche différente est nécessaire.

Cet article présente un modèle simplifié mais plus général pour l'affectation simultanée des locomotives et des wagons. Le modèle permet de choisir la combinaison d'équipement utilisée sur chaque train et impose les contraintes de demande, de capacité, et de disponibilité de l'équipement. Bien qu'il ne tienne pas compte des contraintes d'entretien, des possibilités de substitution ou des pénalités pour le couplage et le découplage de wagons, il possède néanmoins une structure permettant de traiter simplement ces extensions. Comme le précédent, ce modèle est basé sur un ensemble de problèmes de flot dans un réseau qui sont cette fois reliés par des variables et des contraintes exprimant les restrictions relatives au choix des combinaisons d'équipement et aux séquences de train.

Afin de résoudre le problème, plusieurs approches exactes sont comparées. Nous considérons d'abord la relaxation lagrangienne et la décomposition de Dantzig-Wolfe. En ne conservant que les contraintes de conservation de flot dans le sous-problème, on obtient un modèle très facile à résoudre. Malheureusement, le trop grand nombre de contraintes liantes relaxées ou traitées au niveau du problème maître fait en sorte que les temps de calcul sont très élevés pour ces deux approches. Par contre, le modèle peut être résolu très rapidement à l'aide d'une approche basée sur la décomposition de Benders. Le modèle proposé admet en effet une décomposition primale au niveau des variables: pour une affectation réalisable de combinaisons d'équipement aux séquences de trains, le problème se résume à un problème de flot pour chaque type d'équipement.

Nous proposons par ailleurs plusieurs raffinements afin d'accélérer l'algorithme de décomposition de Benders appliqué à ce problème. Par exemple, la génération *a priori* d'un petit nombre de coupes d'optimalité et l'ajout de contraintes valides au problème maître permettent des gains de vitesse substantiels. En fait, cette approche permet de résoudre en quelques minutes seulement des instances semblables à celles décrites au chapitre précédent mais comportant toutefois moins de contraintes.

La contribution de cet article est donc de présenter un modèle simplifié mais général qui servira de point de départ pour le développement de modèles plus complexes incorporant tout la gamme des contraintes traitées dans l'application précédente. La structure de ce modèle fait en sorte qu'il peut être résolu très rapidement, même lorsque le nombre de trains augmente ou que chaque train peut être couvert par plusieurs combinaisons d'équipements. Le prochain chapitre présente différentes extensions à ce modèle.

A Benders Decomposition Approach for the Locomotive and Car Assignment Problem

JEAN-FRANÇOIS CORDEAU and FRANÇOIS SOUMIS

École Polytechnique de Montréal

JACQUES DESROSIERS

École des Hautes Études Commerciales de Montréal

July 1998

Abstract

One of the many problems faced by rail transportation companies is to optimize the utilization of the available stock of locomotives and cars. In this paper, we describe a decomposition method for the simultaneous assignment of locomotives and cars in the context of passenger transportation. Given a list of train legs and a fleet composed of several types of equipment, the problem is to determine a set of minimum cost equipment cycles such that every leg is covered using appropriate equipment. Linking constraints, which appear when both locomotives and cars are treated simultaneously, lead to a large integer programming formulation. We propose an exact algorithm, based on the Benders decomposition approach, that exploits the separability of the problem. Computational experiments carried on a number of real-life instances indicate that the method finds optimal solutions within short computing times. It also outperforms other approaches based on Lagrangian relaxation or Dantzig-Wolfe decomposition, as well as a simplex-based branch-and-bound method.

Keywords: Rail transportation; integer programming; multi-commodity network flow models; Benders decomposition.

3.1 Introduction

In most countries, passenger trains operate according to schedules which are revised every few months according to anticipated or observed variations in the demand. These schedules depend on evaluations of passenger traffic and on the availability of the resources required to operate the trains. Given a proposed train schedule, the railway must plan the utilization of the available equipment so as to ensure service on all scheduled trains while minimizing operational costs. The traditional planning approach consists in separating the assignment of locomotives to trains from the assignment of cars. In this paper, we propose a method for simultaneously assigning both locomotives and cars. Because of the high degree of inter-dependence between these decisions, our approach can generate very significant savings for most railways.

Railways usually use locomotives and cars of different types which are combined together to form train consists. A *train consist* is a group of compatible units of equipment that travel along on some part of the physical rail network. In the context of passenger transportation, a train consist is typically formed by attaching to one or two locomotives a certain number of first-class and second-class cars. Occasionally, additional restaurant or baggage cars can also be part of it. When multiple types of locomotives and cars are available, attention must be paid to combine together compatible units of equipment: some units may not be coupled together for technical reasons while others should not be combined for the sake of homogeneity. Normally, each equipment also has an associated (maximum) operating speed and the operating speed of the consist will be limited by the slowest of its components. While delays are often tolerated in the case of freight transportation, they are a critical issue in the case of passenger transportation.

The equipment assignment plan specifies the composition of the train consist that will be used on each scheduled train, and indicates which trains will be covered by the same units of equipment. In a medium-term planning horizon (i.e., a few months), the fleet of equipment is fixed and the objective followed in making these decisions is usually to minimize some measure of operational costs. The problem is generally defined over a given planning horizon that corresponds to the period length of the schedule. In most cases, this period is equal to a week. However, since the schedule is updated only every few months, a periodic solution that will repeat cyclically is very desirable. Hence, the problem can be referred to as the *tactical periodic equipment assignment problem*. The model and solution methods introduced next could also be applied to the strategic problem in which resource acquisition is taken into account.

Besides these fundamental aspects of the problem studied, many complicating constraints must be considered. First, railway equipment is very costly and resources are generally limited. Hence, planners must deal with upper bounds on the number of units of equipment of each type they may use. Next, a large variety of additional constraints come from the specific characteristics of the physical rail network. For example, reversing the orientation of a train or detaching a car from a train consist during a stop in a station may require the presence of special equipment or personnel. Finally, to comply with safety regulations and perform minor repairs, each unit must usually be inspected at regular intervals.

The simultaneous assignment of both locomotives and cars has received a rather limited attention in the operations research literature. Decision support systems for improving the utilization of locomotives and cars at Indian Railways were developed by RAMANI (1981) and RAMANI and MANDAL (1992). Very little optimization is present in their system which basically helps planners to perform local improvements to the solution by analyzing train connections. A system developed by SABRE for SNCF (BEN-KHEDER *et al.*, 1997) treats equipment modules containing both

locomotives and cars. However, these modules are already built and there only remains to decide which modules will be assigned to each train. Also, all modules allowed to cover a given train are compatible. The problem treated in the present paper is more complex since there are separate requirements for each type of equipment and one must consider compatibility constraints between these equipment types.

The problem of locomotive assignment has, however, been the subject of much more research. In particular, BOOLER (1980, 1995), WRIGHT (1989), FORBES *et al.* (1991) and FISCHETTI and TOTH (1997) have studied the version in which each train requires a single locomotive but multiple locomotive types are available. The more complex problem where each train may require many locomotives was first studied by FLORIAN *et al.* (1976). The authors proposed a multi-commodity network flow formulation and an algorithm based on Benders decomposition for the strategic problem of engine acquisition. Later, SMITH and SHEFFI (1988) described a heuristic for a model that incorporates uncertainty in locomotive requirements through the definition of the objective-function. Also, the implementation of a planning model at the Union Pacific Railroad was described by CHIH *et al.* (1990). Very recently, ZIARATI *et al.* (1997b) modeled the operational version of the problem as a multi-commodity network flow problem with supplementary variables and constraints. A weekly horizon is considered but in order to solve large instances, the problem is split on a temporal basis into a set of overlapping slices of two or three days each. The problem for each time slice is optimized using a branch-and-bound procedure in which the LP relaxations are solved with a Dantzig-Wolfe decomposition. Computational experiments performed on data from the Canadian National Railroad generated very significant savings over the solution used by the company. However, models for the assignment of locomotives do not generalize easily to the simultaneous assignment of locomotives and cars since they consider only train-locomotive compatibility and

neglect the effect of equipment combinations on operating speed. Additionally, most of these models do not consider connection times that depend on whether the locomotives are uncoupled after the arrival of the train.

For a recent survey of optimization models for train routing and scheduling, the reader is referred to the work of CORDEAU *et al.* (1998c).

The rest of the paper is organized as follows. In Section 3.2, the notation that is used throughout the text is introduced and a mathematical formulation of the problem based on a space-time network is presented. An exact algorithm based on the Benders decomposition approach is then described in Section 3.3, while refinements and implementation considerations are detailed to some extent in Section 3.4. In order to investigate the efficiency of the method, computational experiments were performed using data from VIA Rail Canada. The results of these experiments are reported in Section 3.5. Conclusions and paths for future research are presented in the last section.

3.2 Mathematical Model

Let K be the set of equipment types. An *equipment type* $k \in K$ is usually defined for each make of locomotive or car operated by a railway, and specifies the common characteristics and availability of a group of units that are considered identical. Let R be the set of consist types. A *consist type* $r \in R$ identifies a collection of compatible equipment types containing one locomotive type and some car types that may be used to form a train consist with a given maximum operating speed. Set R is used to impose compatibility constraints. Let L be the set of train legs. Each *train leg* $l \in L$ is defined by origin and destination stations, resource requirements, and possible pairs

of departure and arrival times that depend upon the speed of the consist used on the leg. An ordered set of train legs $\{l_{i_1}, l_{i_2}, \ldots, l_{i_m}\}$ is said to be feasible for a given consist type if for every pair of consecutive legs $(l_{i_j}, l_{i_{j+1}})$, the destination station of the first leg is the origin station of the second leg, and the connection time between the two legs is sufficient. The feasibility of a set of train legs depends on the consist type since its operating speed affects the departure and arrival times.

In some cases, even though a pair of legs is feasible, it may be impossible to modify the consist at the intermediate station, either because the connection time is too short or because the necessary installations are not available. To take this into consideration in our model, we define a *train sequence* as a feasible ordered set of train legs such that if these legs are covered by the same consist, then the consist may not be modified at any intermediate point. Let S^r $(r \in R)$ represent the set of train sequences on which a consist of type r can be used. For notational convenience, the set S^r also contains sequences composed of a single train leg which can be covered by a consist of type r.

3.2.1 Network Representation

For each equipment type $k \in K$, we define a space-time network $G^k = (N^k, A^k)$ where N^k is the node set and A^k is the arc set. A portion of such a network is presented in Figure 3.1.

The node set N^k contains departure, arrival and repositioning nodes for every train leg on which equipment of type k may be used. While the departure node corresponds to the exact departure time for the given leg, the arrival node represents the moment defined by the arrival time plus an additional duration, called the *thru-turn time*, which corresponds to the time needed to inspect the train consist after

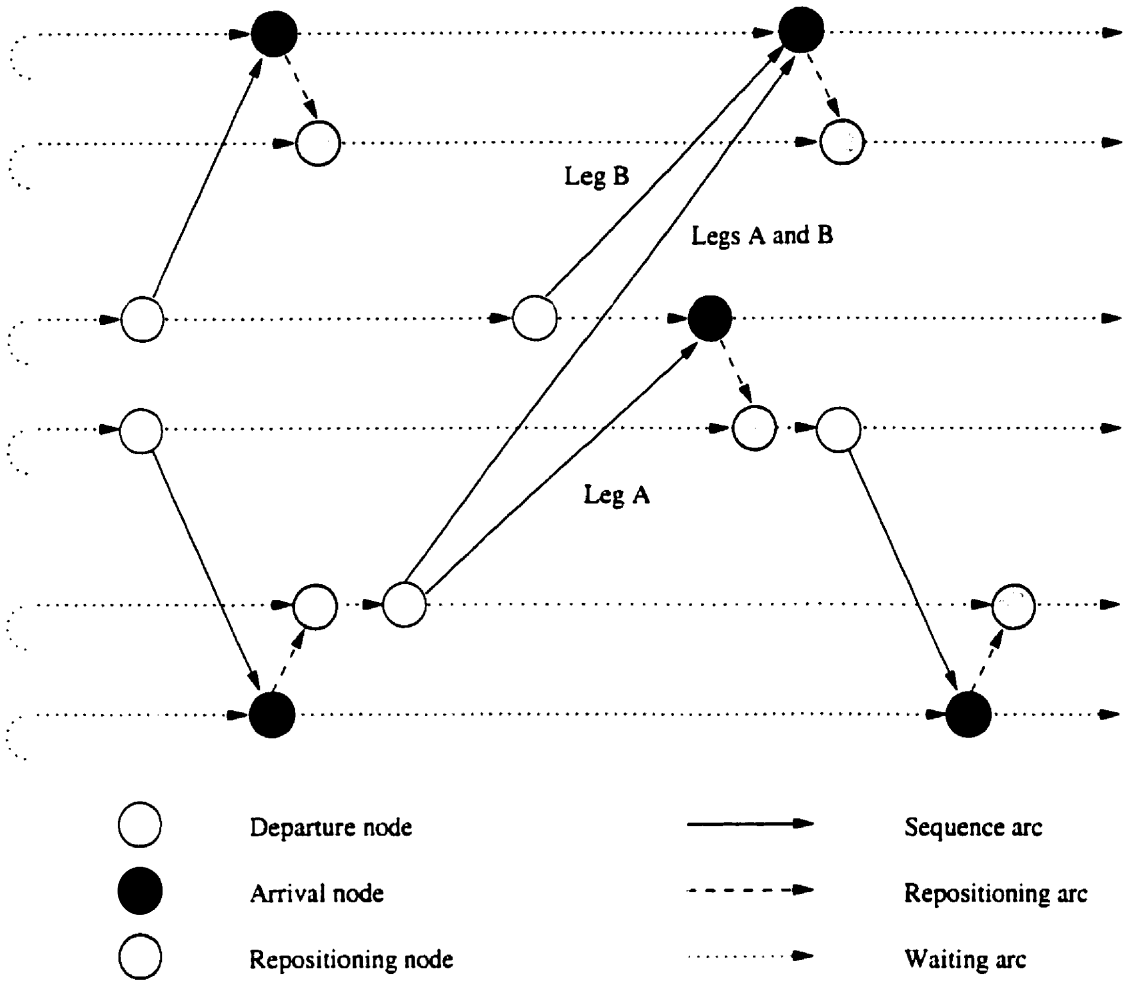

Figure 3.1: Portion of network G^k for equipment type k

its arrival and let passengers get on or off the train. This moment thus represents the time after which the train consist can continue its trip through the network. Additional nodes are used to represent the repositioning of a unit within the same station after its arrival. For example, if a station is located along an east-west track, then a train consist entering the station from the east will need an extra amount of time, called the *turn-around time*, to reposition itself for an eastbound leg. When the train consist can be modified at the end of a leg, the thru-turn and turn-around times include the time necessary to perform the modifications, called the *switching time*.

If switching is completely forbidden in a certain station, that station need not be represented in the space-time network. Legs that have the corresponding station as an origin or a destination will necessarily be covered as part of a sequence containing two or more legs.

The arc set A^k contains a train sequence arc for every sequence on which equipment of type k may be used. Such an arc goes from the origin node of the first leg to the arrival node of the last leg in the sequence. Define $T^k \subseteq A^k$ as the set of arcs in the graph G^k that are associated with train sequences.

To illustrate the purpose of train sequences, consider legs A and B in Figure 3.1. These two legs constitute a feasible ordered sequence of legs since the arrival time of the first leg plus the connection time (thru-turn time) is less than the departure time of the second leg. However, these legs cannot be covered by the same piece of equipment if modifications are to be made to the consist at the intermediate station. In the space-time network, the arrival node for leg A is thus located after the departure node for leg B. Hence, if it is desired to modify the consist covering leg A after its arrival in the destination station, this equipment cannot be used to cover leg B. If however, one accepts to use exactly the same equipment on leg B as on leg A, then the two legs can be covered by the same consist as part of a sequence. In some cases, it will be preferable to avoid modifying the consist even though unnecessary equipment can possibly be hauled on one of the two legs. The additional units of equipment that are present on a consist but not needed for its operation are called *dead-heading* units. Dead-heading units usually give rise to unnecessary costs that railways try to minimize.

The set A^k also contains a repositioning arc for every possible movement within a station. For example, if a station is divided between eastbound and westbound trains, such arcs would be used to represent the change of orientation of a physical

train consist. Generally, one repositioning arc is needed for each train leg which can occur last in a sequence.

Finally, a waiting arc is defined for every pair of nodes that represent consecutive events (departure, arrival or repositioning) involving trains with the same orientation. Again, if a station is divided between eastbound and westbound trains, then waiting arcs exist between departure, arrival or repositioning nodes that involve the trains oriented accordingly. Since a periodic solution over a given horizon is sought, waiting arcs are also present between the nodes that represent the last event and the first event of the period in each station.

3.2.2 A Multi-Commodity Network Flow Formulation

For every consist type $r \in R$ and for every sequence $s \in S^r$, let y_{rs} be a binary variable equal to 1 if and only if train sequence s is covered using a consist of type r. For every equipment type $k \in K$ and every arc $a \in A^k$, let x_a be a non-negative integer variable representing the flow on arc a and let c_a represent the operational cost of using one unit of equipment on that arc. For sequence arcs, this cost depends on the length (mileage) of the arc and on the type of equipment. It usually concerns fuel consumption, maintenance and minor repairs, but can also include a depreciation cost associated with equipment ownership. For waiting and repositioning arcs, this cost can include a penalty for minimizing their utilization.

For every equipment type $k \in K$ and every sequence arc $a \in T^k$, define $r_a \in R$ and $s_a \in S^{r_a}$ as the consist type and the sequence associated with the arc a, respectively. Since a given sequence and a given consist type usually have several arcs associated with them (one for each type of equipment used in the consist), it is convenient to be able to refer to the collection of all arcs associated with this sequence and this consist

type. Hence, define $T_{rs} = \bigcup_{k \in K} \{a \in T^k | r_a = r, s_a = s\}$. For any arc a, let also k_a represent the equipment type associated with this arc.

For any train leg l and any train sequence s, define the binary constant d_{ls} equal to 1 if and only if train leg l is part of sequence s. For every $k \in K$ and $a \in T^k$, define ℓ_a as the minimum number of units of equipment k needed to cover train sequence s_a, and u_a as the maximum number of units of equipment k allowed on train sequence s_a. These numbers, which serve to impose resource requirements and locomotive pulling capacity, have a meaning only when the corresponding sequence is covered with a consist using the given equipment. Otherwise, no unit of equipment k will be allowed on the arc associated with the corresponding sequence.

Finally, for every node $n \in N^k$ $(k \in K)$, the sets $I_n \subseteq A^k$ and $O_n \subseteq A^k$ contain all arcs that are directed in and out of the node n, respectively.

The tactical periodic equipment assignment problem can now be stated as follows:

$$\text{Minimize} \quad \sum_{k \in K} \sum_{a \in A^k} c_a x_a \tag{3.1}$$

subject to

$$\sum_{r \in R} \sum_{s \in S^r} d_{ls} y_{rs} = 1 \qquad (l \in L) \tag{3.2}$$

$$x_a - \ell_a y_{rs} \geq 0 \qquad (r \in R; s \in S^r; a \in T_{rs}) \tag{3.3}$$

$$x_a - u_a y_{rs} \leq 0 \qquad (r \in R; s \in S^r; a \in T_{rs}) \tag{3.4}$$

$$\sum_{a \in I_n} x_a - \sum_{a \in O_n} x_a = 0 \qquad (k \in K; n \in N^k) \tag{3.5}$$

$$x_a \geq 0 \text{ and integer} \qquad (k \in K; a \in A^k) \tag{3.6}$$

$$y_{rs} \in \{0, 1\} \qquad (r \in R; s \in S^r). \tag{3.7}$$

In this model, the binary y_{rs} variables indicate the assignment of consist types to train sequences while the integer x_a variables represent the actual routing of the locomotives and cars. The objective function (3.1) minimizes the sum of all operational costs. Constraints (3.2) require that each train leg be part of exactly one sequence covered by an appropriate consist. Constraints (3.3) and (3.4) impose lower and upper bounds on sequence arcs of all networks depending on the choice of sequences and consist types. Flow conservation at every node for each equipment type is imposed by constraints (3.5).

This formulation is rather general and does not take into account the specific details of any particular application. It can however be customized to deal with many additional situations. First, equipment availability constraints can be incorporated easily. Let $C^k \subseteq A^k$ be a set of pairwise incompatible arcs in G^k such that the removal of these arcs makes the network acyclic. For example, this cut can contain all arcs that traverse a given moment in time. It is easy to verify that when flow conservation equations are satisfied throughout the network, it suffices to impose an upper bound on the sum of the flows on all arcs of the cut C^k to ensure that equipment availability will be satisfied at any time. If the number of available units of equipment k is denoted by e^k, then the following constraints can be added to the original formulation:

$$\sum_{a \in C^k} x_a \leq e^k \qquad (k \in K). \tag{3.8}$$

Next, it is assumed here that the cost of using one unit of equipment is the same whether the unit is active or inactive. If this is not the case, then for every sequence arc $a \in T^k$, c_a may represent the cost of a dead-heading unit on sequence s_a and an additional cost c_{rs} may then be associated with y_{rs} to represent the supplementary expenses, over the dead-heading costs, incurred by all active units required. One

would then add the term $\sum_{r \in R} \sum_{s \in S^r} c_{rs} y_{rs}$ to the objective function (3.1). Also, by setting $c_a = M^k$ for each cut arc $a \in C^k$ where M^k is a large positive constant representing the fixed cost of owning and maintaining one unit of equipment k over the considered planning horizon, one will minimize a weighted combination of the number of units used and the operating costs.

Finally, the inclusion of maintenance constraints is more involved. If the constraints are expressed in terms of a maximum number of days, say p, between successive maintenances, these constraints can be incorporated to the formulation by replacing the network for each type of equipment by a multi-commodity network. For each day of the planning horizon, a commodity would then be associated with equipment trips starting at a maintenance station on that day and finishing at most p days later. Additional flow conservation constraints would also be needed to link these commodities at every station where maintenance can be performed. Maintenance constraints will not be treated in this paper but will be the object of subsequent research to address various extensions.

The formulation contains a large number of variables and constraints, even for moderate-size instances. The large size of the model is a direct result of the need to consider connection times that depend on whether switching is performed. When switching is allowed in any station and switching time is not larger than thru-turn and turn-around times, there is no need to define sequences containing more than one leg. However, this is very rarely the case and the resulting model usually has a large number of sequences. Solving it through a branch-and-bound method with bounds computed using the simplex algorithm may thus require a significant amount of computing time. However, the model has a nice block angular structure which is well suited for mathematical decomposition. We now consider the use of a primal decomposition method to solve this problem.

3.3 Benders Decomposition

For any feasible solution to constraints (3.2) and (3.7) which involve only the y_{rs} variables, problem (3.1)-(3.7) decomposes into $|K|$ network flow subproblems. Hence, for given values of the (complicating) y_{rs} variables which indicate the assignment of consist types to train sequences, the resulting subproblems are relatively easy to solve and involve only the flow variables x_a. This observation points to a method that would iteratively adjust the values of the y_{rs} variables until optimality is reached or a good solution is found. This is the motivation for using Benders decomposition (BENDERS, 1962). We now proceed to reformulate model (3.1)-(3.7) into a model with less variables but many more constraints. Fortunately, most of these constraints are inactive at optimality and need not be considered explicitly. Hence, we will then describe how an efficient algorithm can be derived from this reformulation.

3.3.1 Benders Reformulation

Let $\boldsymbol{Y}$ be the set of binary vectors for the y_{rs} variables that satisfy constraints (3.2) and (3.7). For any given vector $\bar{\boldsymbol{y}} \in \boldsymbol{Y}$, the resulting problem in the x_a variables, called the *primal subproblem*, is defined as follows:

$$v(\bar{\boldsymbol{y}}) = \text{Minimize} \quad \sum_{k \in K} \sum_{a \in A^k} c_a x_a \tag{3.9}$$

subject to

$$x_a \geq \ell_a \bar{y}_{rs} \qquad (r \in R; s \in S^r; a \in T_{rs}) \tag{3.10}$$

$$x_a \leq u_a \bar{y}_{rs} \qquad (r \in R; s \in S^r; a \in T_{rs}) \tag{3.11}$$

$$\sum_{a \in I_n} x_a - \sum_{a \in O_n} x_a = 0 \qquad (k \in K; n \in N^k) \tag{3.12}$$

$$x_a \geq 0 \text{ and integer} \qquad (k \in K; a \in A^k). \tag{3.13}$$

Since the values of the y_{rs} variables are fixed, constraints (3.10) and (3.11) become simple lower and upper bounds on the x_a variables: once the assignment of consist types to sequences is made, there only remains to determine the exact number of locomotives and cars used on the sequence arcs. These lower and upper bound constraints can be treated implicitly, thus considerably reducing the size of the problem. Also, since problem (3.9)-(3.13) decomposes into a set of pure network flow problems, integrality constraints can be discarded with no effect on the optimal solution. Hence, the optimal value of this problem is equal to the optimal value of the dual of its LP relaxation.

Let $\boldsymbol{\beta} = (\beta_a \geq 0 | r \in R, s \in S^r, a \in T_{rs})$, $\boldsymbol{\gamma} = (\gamma_a \leq 0 | r \in R, s \in S^r, a \in T_{rs})$ and $\boldsymbol{\pi} = (\pi_n | k \in K; n \in N^k)$ be the dual variables associated with constraints (3.10), (3.11) and (3.12), respectively. For every arc $a \in A^k$ $(k \in K)$, define i_a and j_a as the tail and head nodes, respectively. The dual of the LP relaxation of the primal subproblem, called the *dual subproblem*, is written as follows:

$$\text{Maximize} \quad \sum_{r \in R} \sum_{s \in S^r} \sum_{a \in T_{rs}} (\ell_a \bar{y}_{rs} \beta_a + u_a \bar{y}_{rs} \gamma_a) \tag{3.14}$$

subject to

$$\beta_a + \gamma_a - \pi_{i_a} + \pi_{j_a} \leq c_a \qquad (k \in K; a \in T^k) \tag{3.15}$$

$$-\pi_{i_a} + \pi_{j_a} \leq c_a \qquad (k \in K; a \in A^k \setminus T^k) \tag{3.16}$$

$$\beta_a \geq 0 \qquad (k \in K; a \in T^k) \tag{3.17}$$

$$\gamma_a \leq 0 \qquad (k \in K; a \in T^k). \tag{3.18}$$

Let $\boldsymbol{D}$ be the feasible region of the dual subproblem and let $P_{\boldsymbol{D}}$ and $Q_{\boldsymbol{D}}$ be the set of extreme points and extreme rays of $\boldsymbol{D}$ respectively. Note that $\boldsymbol{D}$ does not depend on $\bar{\boldsymbol{y}}$ and that $\boldsymbol{D} \neq \emptyset$ whenever $c_a \geq 0$ $(k \in K; a \in A^k)$ since the null vector $\mathbf{0}$ is a

feasible solution. Hence, by strong duality, either the primal subproblem is infeasible or it is feasible and bounded. The optimal value of the preceding pair of primal and dual subproblems can thus be characterized as follows. If

$$\sum_{r\in R}\sum_{s\in S^r}\sum_{a\in T_{rs}}(\ell_a\bar{y}_{rs}\beta_a + u_a\bar{y}_{rs}\gamma_a) \leq 0$$

for every extreme ray $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi}) \in Q_D$ then the dual subproblem is bounded and the primal subproblem is feasible. The optimal value of both problems is then given by

$$\max_{(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\pi})\in P_D}\sum_{r\in R}\sum_{s\in S^r}\sum_{a\in T_{rs}}(\ell_a\bar{y}_{rs}\beta_a + u_a\bar{y}_{rs}\gamma_a).$$

If, however, there exists an extreme ray $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi}) \in Q_D$ such that

$$\sum_{r\in R}\sum_{s\in S^r}\sum_{a\in T_{rs}}(\ell_a\bar{y}_{rs}\beta_a + u_a\bar{y}_{rs}\gamma_a) > 0,$$

then the dual subproblem is unbounded and the primal subproblem must be infeasible.

Since we are interested only in vectors $\bar{\boldsymbol{y}}$ such that the resulting primal subproblem in the x_a variables is feasible, we wish to make sure that we select only $\bar{\boldsymbol{y}}$ vectors that give rise to a bounded dual subproblem.

The original problem (3.1)-(3.7) can thus be restated as

$$\text{Minimize} \quad \max_{(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\pi})\in P_D}\sum_{r\in R}\sum_{s\in S^r}\sum_{a\in T_{rs}}(\ell_a\beta_a + u_a\gamma_a)y_{rs} \tag{3.19}$$

subject to

$$\sum_{r\in R}\sum_{s\in S^r}\sum_{a\in T_{rs}}(\ell_a\beta_a + u_a\gamma_a)y_{rs} \leq 0 \qquad ((\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi}) \in Q_D) \tag{3.20}$$

$$\boldsymbol{y} \in \boldsymbol{Y}. \tag{3.21}$$

Introducing the free variable z, we then obtain the *Benders reformulation:*

$$\text{Minimize} \quad z \tag{3.22}$$

subject to

$$z - \left(\sum_{r \in R} \sum_{s \in S^r} \sum_{a \in T_{rs}} (\ell_a \beta_a + u_a \gamma_a) y_{rs} \right) \geq 0 \qquad ((\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi}) \in P_D) \tag{3.23}$$

$$\sum_{r \in R} \sum_{s \in S^r} \sum_{a \in T_{rs}} (\ell_a \beta_a + u_a \gamma_a) y_{rs} \leq 0 \qquad ((\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi}) \in Q_D) \tag{3.24}$$

$$\boldsymbol{y} \in \boldsymbol{Y}. \tag{3.25}$$

Replacing set $\boldsymbol{Y}$ by its definition, one finally obtains the following reformulation, called the *master problem*:

$$\text{Minimize} \quad z \tag{3.26}$$

subject to

$$z - \left(\sum_{r \in R} \sum_{s \in S^r} \sum_{a \in T_{rs}} (\ell_a \beta_a + u_a \gamma_a) y_{rs} \right) \geq 0 \qquad ((\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi}) \in P_D) \tag{3.27}$$

$$\sum_{r \in R} \sum_{s \in S^r} \sum_{a \in T_{rs}} (\ell_a \beta_a + u_a \gamma_a) y_{rs} \leq 0 \qquad ((\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi}) \in Q_D) \tag{3.28}$$

$$\sum_{r \in R} \sum_{s \in S^r} d_{ls} y_{rs} = 1 \qquad (l \in L) \tag{3.29}$$

$$y_{rs} \in \{0, 1\} \qquad (r \in R; s \in S^r). \tag{3.30}$$

We have thus reformulated problem (3.1)-(3.7) as an equivalent problem with binary variables and one continuous variable. However, this model contains a huge number of constraints, most of which being inactive at optimality. A natural approach is then to solve a relaxation obtained by dropping the constraints associated with the

extreme points and extreme rays of the dual subproblem and generating them as needed by solving the subproblem itself. We first explain how model (3.1)-(3.7) can be solved to optimality, and then explain how the algorithm may be adapted to deal with equipment availability constraints (3.8).

3.3.2 Basic Algorithm

Let τ represent the iteration number and let P_D^τ and Q_D^τ represent, respectively, the restricted sets of extreme points and extreme rays of $\boldsymbol{D}$ available at iteration τ. Let the relaxed master problem be the problem obtained from the master problem (3.26)-(3.30) by replacing P_D by P_D^τ and Q_D by Q_D^τ. The algorithm may be summarized as follows.

1. Set $\tau := 1$, $P_D^1 := \emptyset$ and $Q_D^1 := \emptyset$.
2. Solve the relaxed master problem.
   (a) If the master problem is infeasible, then the original problem is infeasible, **stop**.
   (b) Otherwise, let $\bar{\boldsymbol{y}}^\tau$ be an optimal solution of value z^τ (a lower bound on the value of the original problem).
3. Solve the primal subproblem, taking $\bar{\boldsymbol{y}}^\tau$ as an input.
   (a) If the subproblem is finite, let $\boldsymbol{x}^\tau$ be a primal optimal solution and let $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi})^\tau$ be a dual optimal solution given as an extreme point.
      - If $v(\bar{\boldsymbol{y}}^\tau) = z^\tau$, then $(\boldsymbol{x}^\tau, \bar{\boldsymbol{y}}^\tau)$ is an optimal solution to the original problem, **stop**.
      - Otherwise, $v(\bar{\boldsymbol{y}}^\tau)$ yields an upper bound on the value of the original problem. Set $P_D^{\tau+1} := P_D^\tau \cup \{(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi})^\tau\}$ to generate an *optimality cut*; set $Q_D^{\tau+1} := Q_D^\tau$.

(b) If the subproblem is infeasible, let $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi})^\tau$ be a dual extreme ray such that

$$\sum_{r \in R} \sum_{s \in S^r} \sum_{a \in T_{rs}} (\ell_a \bar{y}^\tau_{rs} \beta_a + u_a \bar{y}^\tau_{rs} \gamma_a) > 0.$$

Set $Q_D^{\tau+1} := Q_D^\tau \cup \{(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi})^\tau\}$ to generate a *feasibility cut*; set $P_D^{\tau+1} := P_D^\tau$.

(c) Set $\tau := \tau + 1$ and return to step 2.

If problem (3.1)-(3.7) is feasible, the algorithm will stop with an optimal solution $(\boldsymbol{x}^\tau, \bar{\boldsymbol{y}}^\tau)$ in step 3 (a). In the worst case, all extreme points and extreme rays of $\boldsymbol{D}$ will be enumerated.

3.3.3 Equipment Availability Constraints

If equipment availability constraints (3.8) must be enforced, the primal subproblem for a given vector $\bar{\boldsymbol{y}} \in \boldsymbol{Y}$ becomes

$$\text{Minimize} \quad \sum_{k \in K} \sum_{a \in A^k} c_a x_a \tag{3.31}$$

subject to

$$x_a \geq \ell_a \bar{y}_{rs} \qquad (r \in R; s \in S; a \in T_{rs}) \tag{3.32}$$

$$x_a \leq u_a \bar{y}_{rs} \qquad (r \in R; s \in S; a \in T_{rs}) \tag{3.33}$$

$$\sum_{a \in I_n} x_a - \sum_{a \in O_n} x_a = 0 \qquad (k \in K; n \in N^k) \tag{3.34}$$

$$\sum_{a \in C^k} x_a \leq e^k \qquad (k \in K) \tag{3.35}$$

$$x_a \geq 0 \text{ and integer} \qquad (k \in K; a \in A^k). \tag{3.36}$$

Given the LP relaxation of formulation (3.31)-(3.36), let $\boldsymbol{\delta} = (\delta^k \leq 0 | k \in K)$ be the dual variables associated with constraints (3.35). The dual of the LP relaxation of the primal subproblem is a weak dual for the primal subproblem and is given by

$$\text{Maximize} \quad \sum_{r \in R} \sum_{s \in S^r} \sum_{a \in T_{rs}} (\ell_a \bar{y}_{rs} \beta_a + u_a \bar{y}_{rs} \gamma_a) + \sum_{k \in K} e^k \delta^k \tag{3.37}$$

subject to

$$\beta_a + \gamma_a - \pi_{i_a} + \pi_{j_a} + \delta^k \leq c_a \qquad (k \in K; a \in T^k \cap C^k) \tag{3.38}$$

$$\beta_a + \gamma_a - \pi_{i_a} + \pi_{j_a} \leq c_a \qquad (k \in K; a \in T^k \setminus C^k) \tag{3.39}$$

$$-\pi_{i_a} + \pi_{j_a} + \delta^k \leq c_a \qquad (k \in K; a \in C^k \setminus T^k) \tag{3.40}$$

$$-\pi_{i_a} + \pi_{j_a} \leq c_a \qquad (k \in K; a \in A^k \setminus (T^k \cup C^k)) \tag{3.41}$$

$$\beta_a \geq 0 \qquad (k \in K; a \in T^k) \tag{3.42}$$

$$\gamma_a \leq 0 \qquad (k \in K; a \in T^k) \tag{3.43}$$

$$\delta^k \leq 0 \qquad (k \in K). \tag{3.44}$$

Since the primal subproblem does not possess the integrality property, a duality gap may exist and it is then impossible to characterize the optimal value of these problems in terms of the extreme points and extreme rays of the dual subproblem polyhedron.

A situation where the optimal solution to the LP relaxation of the primal subproblem fails to be integer is illustrated in Figure 3.2. The problem contains a single equipment type and three mandatory train sequences (A, B and C) which are represented by arcs with lower and upper bounds equal to 1 and a zero cost. Other arcs with no bounds but a cost of 10 correspond to optional train sequences. Finally, horizontal waiting arcs have no bounds and a cost of zero. It is easy to

check that if the number of units available is unbounded, then the optimal solution to this cyclic problem has cost 0 and consists in using one unit of equipment to cover each mandatory sequence. If however, we impose the constraint that at most 2 units be used, then the optimal fractional solution has a cost of 15 and consists in using 0.5 units on each of the three cycles formed by the waiting arcs and one sequence arc, plus 0.5 units on the cycle that uses the arcs with cost 10. The optimal integer solutions (there are two) have a cost of 30. One of them is to use one unit to cover all mandatory and optional sequences.

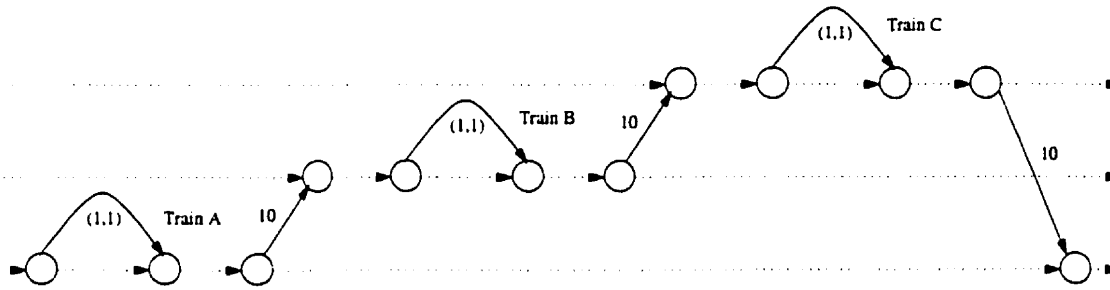

Figure 3.2: Example of a problem with a fractional optimal solution

Hence, problem (3.31)-(3.36) does not have the integrality property. It has, however, a nice property which is stated in the following proposition.

Proposition 1 *Problem (3.31)-(3.36) is feasible if its LP relaxation is feasible.*

Proof: To verify this proposition, first observe that problem (3.31)-(3.36) decomposes into $|K|$ independent problems. For every type of equipment k, assign a cost of 1 to the arcs in the cut C^k and a cost of 0 to all other arcs. Dropping constraints (3.35) and solving the resulting minimum cost circulation problems will determine the minimum number of units of equipment of each type needed to satisfy all requirements. If, for every equipment type k, this number is less than or equal to e^k, then the solution obtained constitutes a feasible solution to the original problem. Since it is also the solution of pure network flow problems, it must be integer. If, however, the minimum

number of units of type k is greater than e^k, then no feasible solution exists. Hence, it is impossible that the LP relaxation of problem (3.31)-(3.36) admit a feasible fractional solution but no integer feasible solution. □

If the integrality requirements on the x_a variables are relaxed in the primal subproblem, the duality gap vanishes. An optimal solution to problem (3.1)-(3.8) can then be computed by using a branch-and-bound procedure. At every node of the branch-and-bound tree, a lower bound is computed by solving the relaxation obtained by dropping the integrality requirements on the x_a variables. This relaxation is solved with the algorithm of section 3.3.2 in which the primal subproblem is replaced by the LP relaxation of (3.31)-(3.36).

The enumeration tree can be pruned at a given node if the relaxation is infeasible or if all variables assume integer values in the optimal solution. It can also be pruned if the cost of the optimal solution is greater than the best upper bound identified so far. Otherwise, child nodes are created by branching on a fractional x_a variable.

At a child node, the algorithm can be accelerated drastically by initializing the sets of extreme points and extreme rays with the elements available at the father node. The validity of the cuts generated at a parent node is justified as follows. When branching on a fractional x_a variable, one is restricting the primal subproblem and thus relaxing the dual subproblem. Hence, all extreme points and extreme rays enumerated previously represent valid, although not necessarily extreme, points and rays of the dual subproblem polyhedron. Therefore, all generated constraints must still be satisfied by the relaxation at a child node.

For any node of the branch-and-bound tree where the relaxation is feasible, an upper bound can also be computed. Since, by Proposition 1, the primal subproblem has a feasible integer solution whenever it has a fractional feasible solution, a feasible

solution to (3.1)-(3.8) is obtained by introducing the integrality requirements on the x_a variables and solving the primal subproblem to optimality. This observation also proves the following proposition.

Proposition 2 *If problem (3.1)-(3.8) is feasible, then a feasible solution can be computed at the first node of the branch-and-bound tree.*

A heuristic algorithm can thus be obtained by first solving the problem without the integrality constraints on the x_a variables, and solving the integer primal subproblem once. The quality of the solution produced by this approach can be arbitrarily poor since the difference between its value and the value of the optimal solution depends on the duality gap in the subproblem. Nevertheless, this solution is optimal if there is no integrality gap in the subproblem.

To avoid the burden of a branch-and-bound procedure, a different approach can be adopted to obtain an optimal solution when imposing equipment availability constraints. Instead of keeping only the y_{rs} variables in the master problem, one can also retain the x_a variables for all $a \in C^k$ and $k \in K$. This way, equipment availability constraints will be satisfied directly in the master problem and the primal subproblem will have the integrality property. However, this approach will likely result in slow convergence unless additional constraints are introduced to link the two sets of variables.

3.4 Algorithmic Refinements

Returning to the basic algorithm of Section 3.3.2, we now discuss refinements that help to improve its performance and stability. We first discuss theoretical aspects,

followed by practical implementation considerations. All ideas apply directly to the branch-and-bound adaptation of Section 3.3.3.

3.4.1 Improving Worst-Case Behaviour

Convergence of the basic algorithm follows from the fact that in the worst case, the number of cuts generated will be equal to the number of extreme points and extreme rays of the dual subproblem polyhedron. This number can be reduced considerably using the fact that the primal subproblem can be decomposed into $|K|$ subproblems, one for each equipment type. Hence, instead of considering the large polyhedron of (3.9)-(3.13), one can consider the individual dual polyhedra of the $|K|$ subproblems and generate cuts directly from these. Let $T^k_{rs} = T_{rs} \cap T^k$ be the set containing the arc of graph G^k associated with train sequence s and consist type r, if any. Let also P_{D^k} and Q_{D^k} be the sets of extreme points and extreme rays of the dual polyhedron D^k for subproblem k. The master problem is then written as

$$\text{Minimize} \quad \sum_{k \in K} z_k$$

subject to

$$z_k - \left(\sum_{r \in R} \sum_{s \in S^r} \sum_{a \in T^k_{rs}} (\ell_a \beta_a + u_a \gamma_a) y_{rs} \right) \geq 0 \qquad (k \in K; (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi}) \in P_{D^k})$$

$$\sum_{r \in R} \sum_{s \in S^r} \sum_{a \in T^k_{rs}} (\ell_a \beta_a + u_a \gamma_a) y_{rs} \leq 0 \qquad (k \in K; (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi}) \in Q_{D^k})$$

$$\sum_{r \in R} \sum_{s \in S^r} d_{ls} y_{rs} = 1 \qquad (l \in L)$$

$$y_{rs} \in \{0, 1\} \qquad (r \in R; s \in S^r).$$

At iteration τ of the algorithm, $|K|$ potential cuts are thus generated when solving the subproblems. Each feasible subproblem proposes an optimality cut from an extreme point and each infeasible one proposes a feasibility cut from an extreme ray. While any feasibility cut generated is certainly violated by the current solution $\bar{\boldsymbol{y}}^{\tau}$, this needs not be the case for an optimality cut. If the cut is generated from a feasible solution that has been obtained previously, it is already satisfied and should not be added to the master problem. It is a simple matter to verify whether the cut should be added. This approach is much more efficient computationally because it takes advantage of the separability of the subproblem.

3.4.2 Solving the Relaxed Master Problem

A major difficulty with this decomposition lies in the solution of the relaxed master problem which is a large 0–1 programming problem with one continuous variable. In fact, this problem needs not be solved to optimality at each iteration. It is possible to generate new cuts from any integer solution. In this case, however, the cost of the relaxed master problem does not necessarily provide a lower bound on the cost of the optimal solution. Hence, it is not possible to stop the algorithm when the cost of the optimal solution to the subproblem is equal to the value of the relaxed master problem. A valid lower bound is nonetheless provided by the linear programming relaxation of the master problem. The algorithm may then be stopped when $UB - LB_{LP} < \epsilon$ where $\epsilon > 0$ is a chosen gap, UB is the cost of the best feasible solution identified so far and LB_{LP} is the lower bound provided by the LP relaxation of the relaxed master problem.

To accelerate the solution process of the master problem, McDaniel and Devine (1977) suggested to relax the integrality constraints on the variables of the master

problem and generate cuts from fractional solutions. On our problem, this approach can be summarized as follows: (*i*) solve the LP relaxation of the problem using the algorithm of Section 3.3.2; (*ii*) add integrality constraints to the relaxed master problem; (*iii*) restart the algorithm to solve the integer programming problem to optimality. Since the relaxation of integrality constraints does not affect the dual subproblem polyhedron, all optimality and feasibility cuts generated in step (*i*) can be used to initialize the corresponding sets of cuts in step (*iii*).

3.4.3 Choosing an Initial Set of Cuts

Even though the algorithm may be initialized from empty sets of extreme points and extreme rays, the choice of these initial sets may greatly affect its convergence. We have found that a good strategy is to start with empty sets of extreme rays but to generate K optimality cuts as follows. For equipment type k, set $\beta_a = c_a$ for every sequence arc $a \in T^k$ and set to 0 all other dual variables. Assuming that $c_a \geq 0$ for all $a \in T^k$, this point is a feasible point (but not necessarily an extreme point) of the dual subproblem polyhedron. It can thus be used to obtain the cut $z \geq \sum_{r \in R} \sum_{s \in S^r} \sum_{a \in T_{rs}} \ell_a c_a y_{rs}$ which is certainly valid since the point is feasible. This constraint can also be generated directly from formulation (3.1)-(3.7). From this formulation, one obtains

$$z = \sum_{k \in K} \sum_{a \in A^k} c_a x_a \geq \sum_{k \in K} \sum_{a \in T^k} c_a x_a \geq \sum_{k \in K} \sum_{a \in T^k} c_a l_a y_{r_a s_a} = \sum_{r \in R} \sum_{s \in S^r} \sum_{a \in T_{rs}} c_a l_a y_{rs}.$$

Hence, the above cut is equivalent to adding the constraint that the cost of the optimal solution to (3.1)-(3.7) must be greater than or equal to the cost of the optimal solution when no dead-heading movement is needed (i.e., $x_a = l_a y_{r_a s_a}$ for all train sequence arcs) and $c_a = 0$ for all $a \in A^k \setminus T^k$ ($k \in K$).

3.4.4 Adding Valid Cuts to the Master Problem

Additional valid cuts can be added to the master problem to enforce part of the constraints that appear in the subproblem and accelerate convergence to the optimal solution. For example, by limiting the number of y_{rs} variables that may be set to 1 among those associated with train sequences occurring at the same time and using a common equipment type, it is possible to help satisfy the availability constraints for the corresponding equipment type. Simple constraints of this nature are

$$\sum_{a \in C_b^k} \ell_a y_{r_a s_a} \leq e^k,$$

where arc set C_b^k is composed of all arcs in A^k that are associated with train sequences that are active at time b. For any values of the y_{rs} variables, this constraint guarantees that the minimum number of units of equipment k needed does not exceed the number available.

Since time is continuous, there is a very large number of potential cuts of this type. Knowing that the number of units needed can only increase when a new train sequence begins, one can simply generate a cut for every moment at which there exists a departure node. Let t_d be the departure time corresponding to node d. For every departure node d, we may then define $C_{t_d}^k \subseteq T^k$ as the set of arcs associated with train sequences that are active at the time of the departure. The subset $C_{t_d}^k$ thus contains the arcs associated with train sequences that begin at node d plus all arcs associated with train sequences that have a departure time smaller than t_d but an arrival time greater than t_d. It should be emphasized that different cuts must be generated for each type of equipment since the associated networks may differ.

To help satisfy flow conservation constraints in the subproblems, one can proceed in a similar way. The exact flows that will take place on the sequence arcs of the subproblems are not known in the master problem but bounds are however provided by the ℓ_a and u_a constants. Let V^k be the set of stations represented in network k, and let N_v^k be the set of nodes associated with station $v \in V^k$ in network k. For each station v, one generates the constraints

$$\sum_{n \in N_v^k} \sum_{a \in I_n} u_a y_{r_a s_a} \geq \sum_{n \in N_v^k} \sum_{a \in O_n} \ell_a y_{r_a s_a}$$

and

$$\sum_{n \in N_v^k} \sum_{a \in I_n} \ell_a y_{r_a s_a} \leq \sum_{n \in N_v^k} \sum_{a \in O_n} u_a y_{r_a s_a}.$$

Again, these valid constraints must be added independently for each type of equipment: even though they are defined on the y_{rs} variables, they depend on each particular type of equipment.

3.4.5 Implementation Considerations

To identify extreme points and extreme rays of the dual polyhedron, one may solve either the primal or the dual subproblem. If the primal subproblem is solved with a specialized network algorithm, the values of the dual variables β_a and γ_a are not directly available since constraints (3.10) and (3.11) are treated implicitly as bounds on the variables. They can however be computed easily using the following observations. For every arc $a \in T^k$, let $\bar{c}_a$ represent the reduced-cost of variable x_a in an optimal solution produced by the network algorithm. If $\bar{c}_a > 0$, then x_a must be at its lower bound and one sets $\beta_a = \bar{c}_a$ and $\gamma_a = 0$. On the other hand, if $\bar{c}_a < 0$,

then x_a must be at its upper bound and one sets $\beta_a = 0$ and $\gamma_a = \bar{c}_a$. Finally, if $\bar{c}_a = 0$, then $\beta_a = \gamma_a = 0$.

A difficulty lies in the identification of an extreme ray of the dual polyhedron $\boldsymbol{D}$ in the case where the primal subproblem is infeasible. To avoid generating cuts associated with extreme rays of the dual polyhedron, a natural alternative is to make the primal feasible for any choice of $\bar{\boldsymbol{y}} \in \boldsymbol{Y}$ by introducing artificial variables. When the primal is feasible, the dual must also be feasible and bounded since $\boldsymbol{D} \neq \emptyset$. However, this approach presents an important drawback: the addition of artificial variables with large costs introduces numerical instability in the solution of the master program and slows convergence. Using CPLEX (1997), the values of an extreme ray can be obtained directly by solving the primal subproblem with the primal simplex algorithm and disabling the pre-processor.

3.5 Computational Experiments

To measure the performance of the solution method described in Section 3.3 and evaluate the benefits of the refinements of Section 3.4, computational experiments were performed on a set of instances obtained from VIA Rail Canada. The data originate from the Québec-Windsor corridor, which accounts for the largest portion of all passenger trains operated in Canada. We first describe the test instances used. We then present an analysis of the improvements to the Benders decomposition approach, followed by comparisons with alternative solution methods and a discussion of subproblem integrality gaps.

3.5.1 Description of Data Sets

VIA Rail is the single most important passenger railway in Canada. Rail transportation is not as popular in North America as it is in Europe but the company operates more than 300 trains per week in the Québec-Windsor corridor which links the major cities in central Canada. The company uses six equipment types: two types of engines (LRC and F40) and two types of first-class and second-class cars (LRC and HEP), which can be combined in three different ways to yield train consists with different operating speeds. For example, combining an F40 locomotive with LRC cars yields a consist with an operating speed of 95 mph. Speed is an important issue and some train legs, such as those between Montréal and Québec, must be assigned the faster equipment. The complete equipment fleet is composed of more than 130 units.

The physical network, which is illustrated in Figure 3.3, has nine major stations, and every train leg originates and terminates in one of these stations. Secondary stations, where trains may stop during a leg to let passengers get on or off, need not be considered explicitly in the model. The minimum run-through, turn-around and switching times vary from station to station, and these values typically range from 30 minutes to a few hours. In particular, switching is allowed only in two stations (Montréal and Toronto). At this time, all trains operated by VIA Rail depart and arrive in the same day.

Three data sets were constructed, each corresponding to the weekly schedule used during a different season. Also, three variants of each instance were considered, leading to a total of nine instances. In the first scenario (instances 1a to 3a), the consist type used on each leg is set a priori to match the assignment used by the company and compound equipment types are used in order to reduce the amount

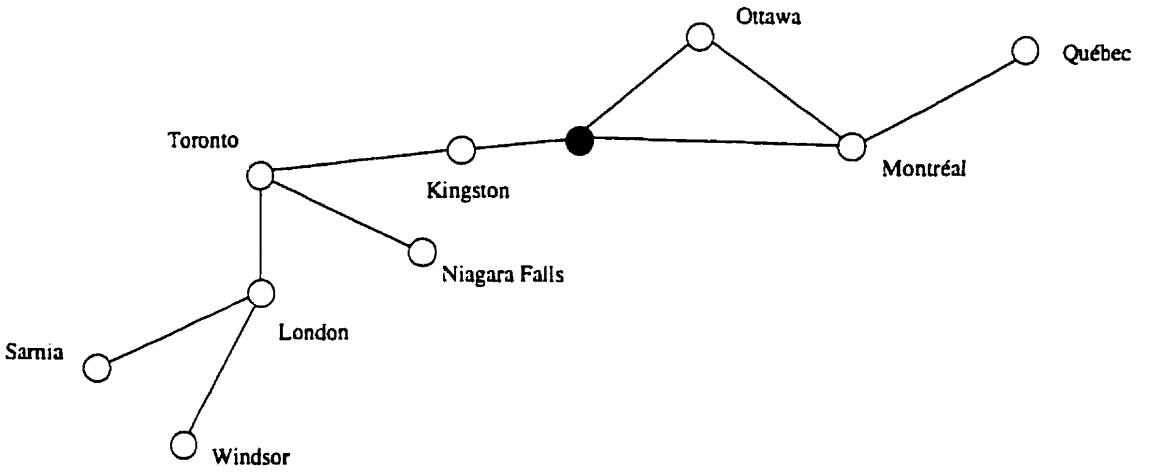

Figure 3.3: Physical network for VIA Rail

of switching performed. A compound equipment is a group of units of different types (e.g., a locomotive and two second-class cars) which are treated as a single module. Additional units of equipment are then assigned separately to train legs that require more than what is provided by a basic module. In the second scenario (instances 1b to 3b), the consist type used on each leg is still fixed but equipment units are disaggregated. In the third scenario (instances 1c to 3c), equipment is also disaggregated and more than half the legs can be covered by any of the three consist types. This is the most complex scenario since a choice must be made regarding the equipment combination that will be used on some train sequences. The first scenario is typically considered by VIA Rail in short-term planning, while the second and third scenarios allow for a greater flexibility and can be used to evaluate additional possibilities in long-term planning.

The objective considered in the experiments is to minimize the sum of operational costs. For each sequence arc, the unit cost is equal to the distance between the origin and the destination stations times the per-mile cost related to fuel, maintenance and minor repairs. Waiting and repositioning arcs have a cost of zero and there are no fixed

costs associated with equipment ownership. Since the consist type used on each leg is fixed in the first two scenarios, the objective then translates into the minimization of dead-heading costs.

Demand for each train leg is expressed as the number of first-class and second-class cars required. The demand for first-class cars is either 0 or 1 while the demand for second-class cars lies between 2 and 8 cars. Most trains require a single locomotive but a few exceptions require two. For locomotives types, the ℓ_a value is thus either 1 or 2 while the u_a value is 2. For car types, the ℓ_a values are determined so as to satisfy demand while the u_a values are set according to the pulling capacity of a locomotive. For example, if the capacity of a locomotive is 8 cars and the demand is 1 first-class car and 3 second-class cars, then the lower and upper bounds for the first-class cars are 1 and 2 respectively, while the corresponding bounds for the second-class cars are 3 and 6. Hence, if an extra locomotive is added to a train that requires only one, its power will not be available to pull additional cars.

In all instances solved, there is a limit on the number of units of each type of equipment that can be used at any time. Hence, constraints (3.8) are present and the approach of Section 3.3.3 must be used to solve the problem. For some equipment types, availability also varies from day to day as some units must be made available for trains outside the corridor while others must stay idle for major maintenance. This is taken into consideration by introducing fictitious train legs with a demand corresponding to the availability reduction.

The characteristics of the test instances are summarized in Table 1. For example, instance VIA1a has 326 train legs, leading to 18 027 possible sequences. The resulting model has 18 027 sequence variables, 38 291 arc variables and 74 964 constraints.

3.5.2 Analysis of Computational Refinements

The goal of these experiments was to evaluate the effects of the refinements proposed in the previous section. Since decomposing the subproblem into separate network flow problems can only improve performance, the algorithm was implemented as explained in Section 3.4.1

The first comparison involved solving the relaxed master problem directly with the added integrality requirements, and solving its LP relaxation followed by the reintroduction of the integrality constraints, as explained in Section 3.4.2. The results of our tests showed that there is a very significant reduction in computation time obtained by first solving the LP relaxation. CPU times were typically divided by a factor of ten on most instances. We do not report comparative statistics for solving the integer problem directly since the CPU time was simply prohibitive in most cases.

The first three columns of Table 3.2 report the number of Benders iterations, number of cuts (optimality cuts (3.27) over total cuts (3.27) and (3.28)), and CPU time needed to solve the LP relaxation. The additional effort needed to reach an optimal integer solution is reported in the next three columns. The gap corresponds

Table 3.1: Characteristics of test instances

Instance	Legs	Sequence variables	Arc variables	Constraints
VIA1a	326	18 027	38 291	74 964
VIA2a	348	18 981	40 328	78 976
VIA3a	348	19 022	40 378	79 112
VIA1b	326	18 027	56 916	111 468
VIA2b	348	18 981	59 943	117 412
VIA3b	348	19 022	60 018	117 618
VIA1c	326	26 752	86 373	167 346
VIA2c	348	30 546	98 070	190 510
VIA3c	348	32 981	105 327	205 080

to the relative difference between the value of the LP relaxation and the cost of the optimal integer solution. The algorithm was coded in C, and all tests were run on a Sun UltraSparc-1 computer (200 MHz).

Table 3.2: Computational results for two-phase method

	LP relaxation solution			**Optimal integer solution**			
Instance	Iter.	Cuts	CPU	Iter.	Cuts	CPU	Gap %
VIA1a	15	9/57	2.48	1	0	0.55	0.2383
VIA2a	11	10/51	1.93	1	0	0.75	0.0985
VIA3a	10	8/48	2.17	1	0	0.86	0.0765
VIA1b	12	8/55	2.89	1	0	0.03	0.0000
VIA2b	16	16/58	3.23	1	0	0.09	0.0000
VIA3b	10	6/49	3.94	1	0	1.93	0.1182
VIA1c	95	28/341	177.96	2	0/2	15.57	0.1245
VIA2c	107	34/357	159.74	1	0	2.18	0.0388
VIA3c	97	17/349	257.36	1	0	1.54	0.0112

These results indicate that the cuts generated when solving the LP relaxation constitute a very good approximation of the set of cuts that determine an optimal integer solution. Since the integrality gaps are also very small (less than 0.25%), one or two iterations of the relaxed integer master problem are usually sufficient before an optimal solution is reached. Also, only once were additional cuts generated. It is interesting to point out that the last three instances, although similar in size, are much harder to solve than the first six because of multiple possibilities with regard to the consist type that can be used on certain legs.

As pointed out in Section 3.3, the network flow subproblems do not possess the integrality property when equipment availability constraints are treated. In Section 3.3.3, we proposed to relax integrality constraints on the x_a variables and to embed the Benders decomposition approach in a branch-and-bound procedure. In fact, this has not been necessary in any of our experiments since there never was an integrality gap in any of the subproblems.

Our next experiments concerned the effect of using an initial set of cuts as described in Section 3.4.3. Using the two-phase approach just described, we solved each instance with and without these initial cuts. The results of these tests are reported in the left and middle portion of Table 3.3. Here, the number of iterations, number of cuts and CPU time refer to the total effort needed to find an optimal integer solution (see Table 3.2). The results show that the initial cuts have a very positive influence on convergence. In particular, computation times are reduced considerably on the last three instances. Although the number of iterations and number of generated cuts were not reduced for the first six instances, we observed that solving the relaxed master problem was much faster when the initial cuts were present. This is explained by the fact that the optimal solution to the master problem is less affected from one iteration to the next when the initial cuts are added.

Table 3.3: Effect of using initial and valid cuts

	Basic algorithm			**Initial cuts**			**Initial and valid cuts**		
Instance	Iter.	Cuts	CPU	Iter.	Cuts	CPU	Iter.	Cuts	CPU
VIA1a	16	57	3.03	15	63	2.11	10	30	2.20
VIA2a	12	51	2.68	13	49	2.26	7	23	1.64
VIA3a	11	48	3.03	11	51	1.56	11	22	1.08
VIA1b	13	55	2.92	9	42	1.86	6	24	1.45
VIA2b	17	58	3.32	11	55	3.03	7	24	1.79
VIA3b	11	49	5.87	11	53	3.74	11	22	1.62
VIA1c	97	343	193.53	66	223	21.27	54	139	13.39
VIA2c	108	357	161.92	60	249	23.08	60	181	15.81
VIA3c	98	349	258.90	44	196	14.94	46	136	11.22

The final step of our analysis was to evaluate the benefits associated with the valid cuts added to the relaxed master problem (Section 3.4.4). Since the first group of (availability) constraints did not produce significant improvements but slowed the solution of the master problem, we used only the second group of (flow conservation) constraints and added two cuts for each station and each equipment type. Using the two-phase approach and the initial cuts, we solved each instance with and without

these additional constraints. The results of these tests are reported in the right portion of Table 3.3. While the number of iterations was not really affected by the introduction of these constraints, the number of Benders cuts generated and CPU time were reduced considerably for most instances.

3.5.3 Comparisons with Alternative Solution Methods

In the second part of our experiments, we compared the performance of the proposed Benders decomposition approach to those of three other solution methods: Lagrangian relaxation (GEOFFRION, 1974), Dantzig-Wolfe decomposition (DANTZIG and WOLFE, 1960) and a simplex-based branch-and-bound algorithm.

By dualizing constraints (3.2), (3.3), (3.4) and (3.8) into the objective function, one obtains an easy problem that separates into a set of network flow subproblems in the x_a variables and a problem in the y_{rs} variables that can be solved by inspection. We have thus implemented this Lagrangian relaxation with a simple subgradient optimization process and a step-size that guarantees geometric convergence. Since the relaxed problem has the integrality property, the largest bound that can be obtained with this relaxation is equal to the value of the LP relaxation of (3.1)-(3.8). This approach must however be embedded in a branch-and-bound algorithm to obtain a feasible solution.

A similar solution method is obtained by applying the Dantzig-Wolfe decomposition principle to (3.1)-(3.8) and keeping constraint sets (3.5) and (3.6) in the subproblem. Again, the subproblem separates into one network flow problem for each type of equipment $k \in K$. This decomposition was implemented with a column generation approach that generates several independent columns from each network at

each iteration. To obtain an optimal integer solution, this approach must also be embedded in a branch-and-bound algorithm to impose integrality constraints on all variables.

The first step was to compare the time needed to compute the LP relaxation lower bound with each of these methods. Since the optimal value of the LP relaxation was known, both algorithms were stopped when the gap between the LP lower bound and the best bound found was less than or equal to 0.1%. For both Lagrangian relaxation and Dantzig-Wolfe decomposition, the CPU time required to solve the LP relaxation was clearly excessive. Even for the first six instances, these approaches required several hours of computation to only approximate the actual LP bound. The major difficulty with the Dantzig-Wolfe decomposition approach is that the master problem, which must be solved at each iteration, contains a very large number of lower and upper bound constraints (3.3) and (3.4). These constraints represent approximately 90% of the numbers reported in Table 3.1, and make each iteration of the column generation process very costly in terms of CPU time. For example, the master problem for instance VIA1a contained more than 70 000 constraints. Lagrangian relaxation was faster but the computing time still exceeded that required by Benders decomposition by at least a factor of 10 on all instances. Since we used straight implementations of Lagrangian relaxation and Dantzig-Wolfe decomposition, it is more than likely that the performance of these methods could be improved at least marginally by using more sophisticated techniques. However, we believe that these methods are not well suited for solving the model proposed in this paper.

It is generally admitted that when an optimization problem is small enough to be solved directly with an appropriate algorithm, using any kind of decomposition will result in longer computing times. Hence, to measure the performance of our

approach, we also compared it with the simplex algorithm. More specifically, we used CPLEX (1997) and solved the problem by first extracting and solving the network portion using the **netopt** module. The dual simplex algorithm was then used with steepest-edge partial pricing to obtain an optimal solution. In the second step of our experiments, we compared the Benders decomposition approach to the branch-and-bound procedure of CPLEX. At each node of the tree, the relaxation was solved with the dual simplex algorithm, except for node 0 where it was solved as explained above. Branching was first performed on y_{rs} variables and each variable was assigned a branching priority proportional to the number of train legs in the associated train sequence. Node selection was performed according to the best-bound criterion and strong branching was used. This strategy consistently gave the best results on all instances.

Table 3.4 summarizes the timing results obtained with Benders decomposition and the simplex-based branch-and-bound method of CPLEX. We report the time needed to solve the LP relaxation, and the total time to obtain an optimal integer solution. For the branch-and-bound method, we also report the number of nodes explored in the enumeration tree and the total number of simplex iterations performed. For Benders decomposition, the results were obtained by using the two-phase approach with both the initial and valid cuts. Surprisingly, the Benders decomposition algorithm was always faster than the branch-and-bound method. This is explained by the fact that constraints (3.3) and (3.4), which account for a large part of the total constraints, become simple bounds on the arcs when using Benders decomposition. The relaxed master problem solved at each iteration is thus reasonably small and the network flow subproblems are solved very quickly.

Table 3.4: CPU time needed to find an optimal solution

	Simplex-based branch-and-bound				Benders decomposition	
	BB	Simplex	CPU Time		CPU Time	
Instance	Nodes	iterations	LP	Integer	LP	Integer
VIA1a	12	20 816	36.77	42.51	0.48	2.20
VIA2a	12	21 285	39.57	46.71	0.59	1.64
VIA3a	5	24 633	53.18	54.43	0.55	1.08
VIA1b	1	54 101	136.98	137.00	0.64	1.45
VIA2b	3	47 847	116.35	118.04	0.85	1.79
VIA3b	6	54 235	145.19	148.79	0.74	1.62
VIA1c	30	128 429	378.60	769.61	12.04	13.39
VIA2c	18	114 792	460.50	630.59	14.55	15.81
VIA3c	8	111 975	567.03	608.71	10.63	11.22

3.5.4 A Discussion of Subproblem Integrality Gaps

In the last part of our experiments, we ran additional tests to evaluate the sensitivity of our approach to the tightness of equipment availability constraints (3.8), and to measure the effect of these constraints on subproblem integrality gaps.

We first solved each of the nine instances by using a modified objective function involving only fixed costs. These costs were chosen so as to first minimize the number of locomotives used, and then minimize the number of cars. The minimization of a weighted fleet size is more appropriate than the minimization of total fleet size given the large difference between the acquisition costs of locomotives and those of cars. The results of these tests are reported in the left part of Table 3.5. The cost column indicates the total fixed cost as a percentage of the fixed cost of the fleet used in the previous experiments. For example, minimizing the fleet size for instance VIA1a produced savings of more than 11%. On the other hand, the savings were less than 2% on all three variants of the third instance. This difference is explained by the fact that the same equipment availabilities were used in all experiments, while the three instances correspond to seasons with a varying level of demand.

We then returned to the original objective function of minimizing total operational costs but set the equipment availabilities according to the equipment availabilities determined in the preceding experiments. The corresponding results are reported in the right part of Table 3.5. Here, the cost column expresses the cost of the solution as a percentage of the cost of the solution obtained with actual availabilities in the experiments of Section 3.5.2. For example, the operational cost for instance VIA1a increased by 0.71% with the reduced equipment availabilities.

Table 3.5: Computational results for fixed cost and variable cost minimization

	Fixed cost minimization					**Variable cost minimization**				
Instance	Iter.	Cuts	CPU	Gap %	Cost %	Iter.	Cuts	CPU	Gap %	Cost %
VIA1a	14	41	0.39	0.1346	88.67	9	35	0.24	0.0000	100.71
VIA2a	12	38	0.48	0.0232	91.86	10	33	0.73	0.0000	100.88
VIA3a	13	43	0.54	0.0225	99.34	9	32	0.78	0.0000	100.15
VIA1b	10	42	0.85	0.1082	90.26	8	27	1.01	0.1558	100.81
VIA2b	9	32	0.61	0.0386	91.32	9	36	1.00	0.0000	101.50
VIA3b	13	43	0.76	0.0371	98.95	9	29	0.42	0.0000	100.15
VIA1c	117	363	67.89	0.0009	87.63	70	188	30.93	0.1516	100.76
VIA2c	96	306	16.06	0.0166	89.47	61	210	9.69	0.0032	101.19
VIA3c	89	300	19.45	0.0000	98.95	36	128	6.98	0.0000	100.00

As in the previous experiments, the integrality gaps were very small for all instances. In addition, no integrality gap was observed in any of the subproblem. This surprising result is in large part explained by the fact that an integrality gap can only appear if the marginal savings obtained by increasing equipment availability by one unit are not monotonically decreasing.

First observe that the subproblem for a given type of equipment is a pure network flow problem with one additional equipment availability constraint. Hence, an integrality gap in this problem can only be caused by the presence of the additional constraint. Consider Figure 3.4 that represents the optimal cost of the subproblem for one equipment type as a function of equipment availability given a fixed solution $\bar{y}$. The actual equipment availability is represented by the value e on the horizontal

axis. In this case, an integrality gap would be present since the (fractional) convex combination of the best solutions with one more or one less unit has a smaller cost than that of the best integer solution with e units available.

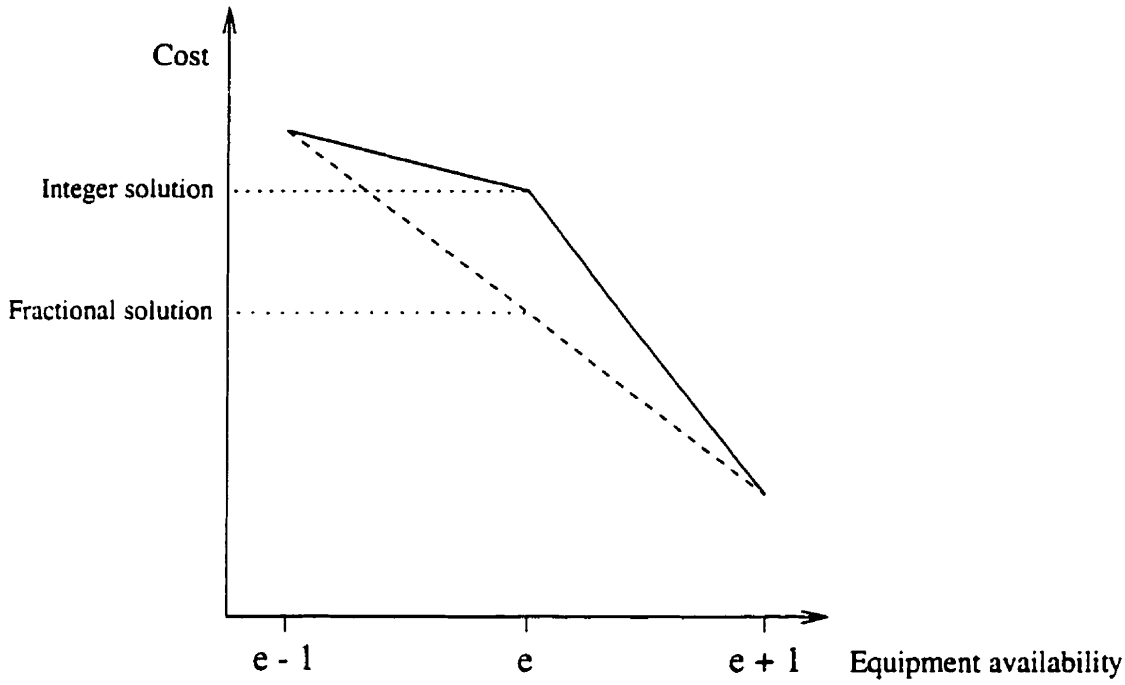

Figure 3.4: Subproblem with an integrality gap

This situation is however not very likely to appear in practice since the marginal savings obtained by increasing fleet size are normally decreasing as the total fleet size increases. In that case, the relation between the cost of the optimal integer solution and equipment availability is a convex function, and all fractional convex combinations have a cost that is greater than or equal to that of the optimal integer solution. It is also worth mentioning that if the objective function includes only fixed costs, then the subproblem cannot have an integrality gap since the availability constraint can be removed with no effect.

3.6 Conclusions

The aim of this paper was to present a basic modeling and solution approach for the problem of simultaneously assigning locomotives and cars to passenger trains. The proposed model captures the basic aspects of the problem and possesses a flexible structure which should facilitate the introduction of additional constraints and possibilities. The structure of the model also makes it well suited for a variable decomposition, leading to an efficient algorithm. The computational experiments performed show that even for instances of moderate size, the Benders decomposition algorithm is faster than solving the problem using a classical simplex-based branch-and-bound method. The superiority of the former method should be even greater on larger instances. In particular, as the number of equipment types increases, one should gain even more by decomposing the problem.

Despite the fact that these computational experiments were performed on real-life data from a railway, some extensions must be considered before the model can be used in practice. For example, maintenance constraints and substitution possibilities should be incorporated to the formulation. Other extensions which will be the object of subsequent research concern the introduction of switching costs or penalties to minimize the number of train consist modifications during connections in stations of the network. Considering the preliminary results obtained so far, we believe that the potential for cost reduction is very significant given the fact that equipment utilization planning is still performed manually by managers at most railways.

Acknowledgments

We wish to thank Mr. Alain Vigeant of VIA Rail who provided us with the data used in the computational experiments. This work was supported by the Québec Government (Fonds pour la Formation de Chercheurs et l'Aide à la Recherche) and by the Natural Sciences and Engineering Research Council of Canada.

Chapitre 4

Simultaneous Assignment of Locomotives and Cars to Passenger Trains

Article écrit par Jean-François Cordeau, François Soumis et Jacques Desrosiers; soumis pour publication à *Operations Research.*

Le chapitre précédent présentait un modèle simplifié ainsi qu'une méthode de résolution très efficace basée sur la décomposition de Benders. Bien qu'il tienne compte des principales caractéristiques du problème, ce modèle possède un niveau de détail insuffisant pour la plupart des applications pratiques du problème. Dans le présent article, nous introduisons trois extensions importantes du modèle du chapitre précédent: les contraintes d'entretien, les pénalités pour la modification des trains et les possibilités de substitution. Ce modèle complet peut encore être résolu très efficacement et représente donc une alternative intéressante à l'approche présentée au second chapitre.

Après avoir décrit brièvement le modèle de base, nous présentons une formulation qui introduit les contraintes d'entretien en remplaçant le modèle de flot associé à chaque type d'équipement par un modèle multi-flots. Cette approche est donc

similaire à celle utilisée pour imposer les contraintes d'entretien dans le modèle du chapitre 2. Nous proposons ensuite une approche permettant d'imposer une pénalité lorsqu'un wagon ou une locomotive est ajouté ou séparé d'un train durant une connexion dont la durée est inférieure à un certain seuil minimum. Contrairement au modèle du chapitre 2 qui pénalise tous les couplages et découplages de wagons, ce nouveau modèle ne pénalise donc que ceux qui risquent de causer le retard de certains trains si la station de connexion est congestionnée ou que le personnel nécessaire pour effectuer les opérations de couplage et de découplage n'est pas immédiatement disponible. En augmentant la valeur du seuil minimum, il est néanmoins possible de pénaliser toutes les modifications apportées aux trains. Nous expliquons finalement comment les possibilités de substitution peuvent être traitées en ajoutant des variables supplémentaires au problème maître.

Ces trois extensions affectent légèrement la structure du modèle mais une méthode de résolution basée sur la décomposition de Benders peut encore être utilisée. Dans ce cas, le sous-problème se décompose en un problème multi-flots pour chaque type d'équipement. La taille du sous-problème peut devenir considérable pour les grandes instances. Ainsi, nous considérons la possibilité de résoudre le sous-problème par une décomposition de Dantzig-Wolfe. L'algorithme de résolution consiste donc en une méthode de séparation et d'évaluation progressive qui résout, par une décomposition de Benders, un problème en variables mixtes à chaque noeud de l'arbre d'énumération. Dans ce problème en variables mixtes, le sous-problème en variables continues est résolu par l'algorithme du simplexe ou par une décomposition de Dantzig-Wolfe.

Deux améliorations sont proposées afin d'accélérer l'algorithme. La première consiste à résoudre une relaxation du problème obtenue en retirant les contraintes d'entretien. Tous les points et rayons extrêmes identifiés pendant la résolution de cette relaxation constituent des points et directions réalisables du sous-problème dual pour le modèle avec les contraintes d'entretien. Ils peuvent donc être utilisés pour

initialiser les ensembles de coupes correspondants. La seconde amélioration consiste à identifier une coupe non dominée dans le cas où le sous-problème dual possède plusieurs solutions optimales. La génération de coupes dites Pareto-optimales peut améliorer la convergence de façon très nette lorsque le sous-problème est fortement dégénéré comme c'est souvent le cas pour les problèmes multi-flots.

Les test effectués montrent que l'approche peut fournir des solutions optimales à des problèmes réalistes en moins d'une heure de calcul sur ordinateur. Cette bonne performance est en partie attribuable au fait que le saut d'intégrité est très faible pour ces instances. Dans tous les tests effectués, une méthode heuristique servant à générer des solutions entières réalisables a identifié une solution optimale du problème au premier noeud de l'arbre de branchement car le sous-problème n'avait aucun saut d'intégrité.

Cet article décrit donc un modèle général et complet pour l'affectation simultanée de locomotives et de wagons. Tout en incorporant trois facettes importantes du problème, il conserve une structure qui se prête bien à une décomposition de Benders. On obtient ainsi une approche permettant de résoudre à l'optimalité des problèmes réels en des temps de calcul très raisonnables. La méthode a par ailleurs été utilisée dans le cadre d'un mandat réalisé pour le compte de VIA Rail. Ce mandat consistait à déterminer la composition optimale de la flotte d'équipement nécessaire pour assurer le service étant donné un horaire hebdomadaire comportant plus de 500 trains.

Simultaneous Assignment of Locomotives and Cars to Passenger Trains

JEAN-FRANÇOIS CORDEAU and FRANÇOIS SOUMIS

École Polytechnique de Montréal

JACQUES DESROSIERS

École des Hautes Études Commerciales de Montréal

December 1998

Abstract

The problem of assigning locomotives and cars to trains is a complex task for most railways. In this paper, we propose a multi-commodity network flow based model for assigning locomotives and cars to trains in the context of passenger transportation. The model has a convenient structure that facilitates the introduction of maintenance constraints, car switching penalties, and substitutions possibilities. The large integer programming formulation is solved by a branch-and-bound method in which some of the integrality constraints are relaxed. At each node of the tree, a mixed-integer problem is solved by a Benders decomposition approach in which the LP relaxations of multi-commodity network flow problems are optimized either by the simplex algorithm or by a Dantzig-Wolfe decomposition. Some computational refinements, such as the generation of Pareto-optimal cuts, are proposed to improve the performance of the algorithm. Computational experiments performed on two sets of data from a railroad show that the approach can be used to produce optimal solutions to complex problems.

Keywords: Rail transportation; integer programming; multi-commodity network flow model; Benders decomposition; Dantzig-Wolfe decomposition.

4.1 Introduction

Planning the assignment of locomotives and cars to trains is a complex task for most railways. In freight transportation, the problem is very often separated into distinct components: freight routing policies first determine the assignment of cars to trains and a locomotive assignment problem is next solved to supply each scheduled train with enough power to pull the assigned cars. The need to resort to a sequential planning approach is a consequence of the large number of locomotives and cars that make up each train and of demand variability. In passenger transportation, however, both locomotives and cars can be assigned in parallel. Since the same set of trains is normally operated every week with a similar number of cars, a cyclic solution can be computed so as to optimize equipment utilization. In addition, the smaller number of units to assign makes it possible to treat both locomotives and cars simultaneously.

The locomotive and car assignment problem consists in finding a set of equipment cycles that cover a list of scheduled trains at minimum cost. Although the problem appears to be reasonably easy at first sight, planners must often deal with a large set of additional constraints that considerably complicate their work. For example, most trains can be covered using different types of equipment among which certain incompatibilities may exist. Also, the choice of equipment usually affects the operating speed of the train which, in turn, determines the arrival time and the set of possible connections. Finally, the assignment of locomotives and cars must satisfy a wide array of operational constraints such as those imposed by maintenance requirements. Hence, even for railways of small size, preparing an equipment assignment plan is a long and tedious task. Not only is it complicated to find a feasible solution given limited equipment availability, but it is also very difficult to evaluate its quality in terms of deviation from optimality. Furthermore, the adaptation of an existing solution to minor changes may require a considerable work.

Given the difficulties associated with the assignment of locomotives and cars to trains, the need to develop optimization tools is clearly apparent. However, while several models have been presented for the assignment of engines to freight trains, a recent survey of optimization models for train routing and scheduling (CORDEAU *et al.*, 1998c) indicates that very few have been developed for the simultaneous assignment of locomotives and cars to passenger trains. One of the first efforts in this direction belongs to RAMANI and MANDAL (1992) who developed decision support systems to assist planners at Indian Railways. Their approach seeks to minimize the required fleet size and is based on a simple exchange heuristic that proceeds by analyzing train connections in the stations of the rail network. More recently, BEN-KHEDER *et al.* (1997) described the development and implementation of a system for the assignment of locomotives and cars to passenger trains at SNCF. This system treats both types of equipment simultaneously but considers aggregated modules which are then assigned as a whole, thus avoiding to deal explicitly with compatibility constraints.

In previous papers, we proposed two modeling and solution approaches for the assignment of locomotives and cars to passenger trains. The first approach (CORDEAU *et al.*, 1998a) was developed by focusing on the specific needs of a particular railway and incorporates a wide range of possibilities and constraints such as substitutions between equipment types and maintenance requirements. The resulting model is based on a multi-commodity network flow structure with linking constraints and is optimized with a heuristic branch-and-bound method in which the linear relaxations are solved by column generation. This approach is the core of a system that has been successfully tested and implemented at VIA Rail Canada. However, the computing time needed to solve this model grows rapidly with the size of the problem and, because heuristic branching is used, the quality of the computed solutions is somewhat dependent upon problem characteristics.

The second approach (CORDEAU *et al.*, 1998b) is based on a more general framework that can be readily adapted to the characteristics of several different railways. For this framework, the authors proposed a basic model and a solution approach based on Benders decomposition. The model, which is described in the next section, captures the fundamental difficulties of the problem and has a structure which leads to a very efficient variable decomposition approach. In this paper, we describe some important extensions to this basic model that make it more appropriate for a real-life application.

The paper is organized as follows. In the next section, we briefly describe the problem and the basic mathematical formulation based on multi-commodity network flows. Then, three extensions are given in Section 4.3. First, we show how maintenance constraints expressed as a maximum number of days between successive stops in a maintenance center can be introduced in the formulation. Next, we propose a method to penalize car switchings so as to reduce the negative impact of such operations on schedule compliance. Finally, we indicate how locomotive and car substitution possibilities can be incorporated to the model. In Section 4.4, we present a branch-and-bound algorithm that solves, by Benders decomposition, a mixed-integer problem at each node of the tree. Within this decomposition scheme, the LP relaxations of multi-commodity network flow problems are solved either by the simplex algorithm or by a Dantzig-Wolfe decomposition. Computational refinements that improve the performance of the algorithm are then described in Section 4.5. Finally, the results of two sets of computational experiments are summarized in Section 4.6.

4.2 A Basic Model

Railways normally use locomotives and cars of different types which are combined in several ways to form *train consists*. Let K be the set of all *equipment types* available

to the railway. A different equipment type $k \in K$ may be defined for each make of locomotive or car operated by the railway. However, if two different makes have identical capacity, speed and compatibility characteristics, they can be aggregated and treated as a single type. Given the set K, let R denote the set of all *consist types* that can be defined using these types of equipment. Each consist type $r \in R$ is a subset $\{k_1^r, k_2^r, \dots\} \subseteq K$ of compatible equipment types that should contain at least one locomotive type and one car type. The definition of the set R serves to impose compatibility constraints: each train will be covered with a unique type of consist, and only the associated equipment types will be allowed on that train. The operating speed of a consist is determined by the slowest of its components.

Let L be the set of *train legs*. Each train leg $l \in L$ is defined by a pair of origin and destination stations together with a set of compatible consist types $\{r_1^l, r_2^l, \dots\} \subseteq R$ that may be used to cover the leg. In addition, for each compatible consist type r_i^l, one must specify the departure and arrival times of the train and, for each equipment type $k \in r_i^l$, the minimum and maximum number of units of that type to be used on the train leg if it is covered with a consist of type r_i^l.

An ordered set of train legs $(l_{i_1}, l_{i_2}, \dots, l_{i_m})$ is said to be feasible for a given consist type if, for every pair of consecutive legs $(l_{i_j}, l_{i_{j+1}})$, the destination station of the first leg is the origin station of the second leg, and the connection time between the two legs is sufficient to allow for passenger exchange and train consist repositioning. The feasibility of a set of train legs depends on the consist type used since its operating speed affects the arrival times.

In some cases, even though a pair of legs is feasible, it may be impossible to modify the consist at the intermediate station, either because the connection time is too short or because the necessary installations are not available. To take this into consideration in our model, we define a *train sequence* as a feasible ordered set of train legs such that if these legs are covered by the same physical train consist, then the

consist may not be modified at any intermediate station. Let S^r ($r \in R$) represent the set of train sequences on which a consist of type r can be used. For notational convenience, set S^r also contains sequences composed of a single train leg that can be covered by a consist of type r. The purpose of defining train sequences becomes more apparent when considering the network representation.

4.2.1 Network Representation

For each equipment type $k \in K$, we define a space-time network $G^k = (N^k, A^k)$ where N^k is the node set and A^k is the arc set. A portion of such a network is illustrated in Figure 4.1. Each station is represented by two lines corresponding to eastbound and westbound trains.

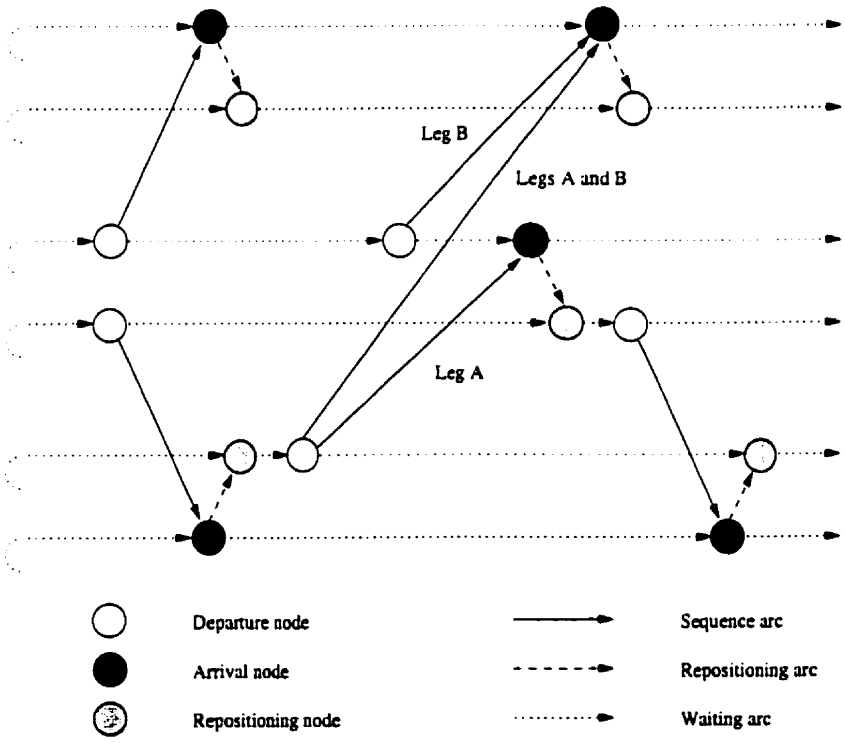

Figure 4.1: Portion of network G^k for equipment type k

Set N^k ($k \in K$) contains three types of nodes: for each consist type and each train leg on which equipment of type k can be used, departure, arrival and repositioning nodes are defined. The time associated with a departure node corresponds to the exact departure time of the corresponding train leg. However, the arrival node represents the moment defined by the arrival time plus an additional duration needed for train inspection and passenger exchange, called the *run-thru time*. Additional repositioning nodes are also used to represent the movement of a unit within the same station after its arrival. For example, if a station is located along an east-west track, then a train consist arriving on a westbound leg will need an extra amount of time, called the *turn-around time*, to reposition itself for an eastbound leg. When the train consist can be modified at the end of a leg, the run-thru and turn-around times include an additional duration, called the *switching time*, which is necessary to perform the modifications. Switching is said to occur whenever a car is added to or separated from a train consist during a connection in a station. If switching is always forbidden in a certain station, that station needs not be represented in the space-time network since no consist modifications will occur there: legs that have the corresponding station as an origin or a destination will necessarily be covered as part of a sequence containing two legs or more.

The arc set A^k ($k \in K$) contains a train sequence arc for every sequence on which equipment of type k may be used. Define $A^k_S \subseteq A^k$ as the subset of arcs in the graph G^k that are associated with train sequences. Each sequence arc links the origin node of the first leg to the arrival node of the last leg in the sequence. The purpose of sequences can now be made more explicit. Since car switching takes time, the true arrival time of a train depends on whether switching must be performed after the arrival. When switching does take place, then all units of equipment used on the arriving train are delayed. But because each type of equipment has its individual network, the model must ensure that whenever switching is performed, then all units of equipment become available at the same moment. This is accomplished by using train sequences as illustrated in the following example.

Consider train legs A and B in Figure 4.1. Because the switching time is larger than the run-thru time, these legs can be covered by the same consist only if it is not modified at the intermediate station. This is represented in the figure by the fact that the arrival node for leg A is located on the right of the departure node for leg B. By imposing a constraint stating that each leg must be covered within exactly one sequence, either legs A and B will both be covered using exactly the same equipment in a sequence containing the two legs, or else they will be covered using completely different units. Thus, the case where the same locomotive would cover both legs while a car would cover leg A but not leg B is not allowed. This is precisely what we wish to model since car switching implies that not only the switched car but also the rest of the consist used on leg A will be delayed after its arrival.

Set A^k also contains a repositioning arc for every possible movement within a station. For example, if a station is divided between eastbound and westbound trains, such arcs would be used to represent the change of orientation of a physical train consist. Generally, one repositioning arc is needed for each train leg which can occur last in a sequence. Finally, a waiting arc is defined for every pair of nodes that represent consecutive events (departure, arrival or repositioning) involving trains with the same orientation. Again, if a station is divided between eastbound and westbound trains, then waiting arcs exist between departure, arrival or repositioning nodes that involve the trains oriented accordingly. Since a periodic solution over a given horizon is sought, waiting arcs are also defined between the nodes that represent the last event and the first event of the period in each station.

4.2.2 A Multi-Commodity Network Flow Based Formulation

For every consist type $r \in R$ and for every sequence $s \in S^r$, let y_{rs} be a binary variable equal to 1 if and only if train sequence s is covered using a consist of type r. For every equipment type $k \in K$ and every arc $a \in A^k$, let x_a be a non-negative integer

variable representing the number of units of equipment k used on arc a, and let f_a represent the operational cost of using one unit of flow on that arc. For sequence arcs, this cost usually depends on the distance traveled in the sequence and on the type of equipment. For repositioning arcs, this cost can include a penalty to minimize unnecessary movements within a station. Finally, waiting arcs normally have a cost of zero.

For every equipment type $k \in K$ and every sequence arc $a \in A_S^k$, define $r_a \in R$ and $s_a \in S^{r_a}$ as the consist type and the sequence associated with the arc a, respectively. Since a given sequence and a given consist type usually have several arcs associated with them (one for each type of equipment used in the consist), it is convenient to be able to refer to the collection of all arcs associated with this sequence and this consist type: define $A_{rs} = \bigcup_{k \in K} \{a \in A_S^k | r_a = r, s_a = s\}$ as the set of all arcs associated with consist type $r \in R$ and sequence $s \in S^r$. For any arc a, let also $k_a = \{k \in K | a \in A^k\}$ represent the equipment type associated with this arc.

For every train leg $l \in L$ and every train sequence s ($r \in R; s \in S^r$), define the binary constant d_{ls} equal to 1 if and only if train leg l is part of sequence s. For every equipment type $k \in K$ and every arc $a \in A_S^k$, define ℓ_a as the minimum number of units of equipment k needed to cover train sequence s_a, and u_a as the maximum number of units of equipment k allowed on train sequence s_a. These numbers are used to impose demand constraints as well as locomotive pulling capacities. They have a meaning only if the corresponding sequence is covered with a consist of type r_a. Otherwise, no unit will be allowed on arc a.

Let $A_E^k \subseteq A^k$ be a set of pairwise incompatible arcs in G^k such that the removal of these arcs makes the network acyclic. For example, this cut can contain all arcs that traverse a given moment in time. It is easy to verify that when flow conservation equations are satisfied throughout the network, it suffices to impose an upper bound on the sum of the flows on all arcs of the cut A_E^k to ensure that equipment availability

will be satisfied at any time. The number of available units of equipment k is denoted by e_k.

Finally, for every node $n \in N^k$ $(k \in K)$, the sets $I_n \subseteq A^k$ and $O_n \subseteq A^k$ contain all arcs that are directed in and out of node n, respectively. The basic model for the *periodic equipment assignment problem* can be stated as follows:

$$\text{Minimize} \quad \sum_{k \in K} \sum_{a \in A^k} f_a x_a \tag{4.1}$$

subject to

$$\sum_{r \in R} \sum_{s \in S^r} d_{ls} y_{rs} = 1 \qquad (l \in L) \tag{4.2}$$

$$x_a - \ell_a y_{rs} \geq 0 \qquad (r \in R; s \in S^r; a \in A_{rs}) \tag{4.3}$$

$$x_a - u_a y_{rs} \leq 0 \qquad (r \in R; s \in S^r; a \in A_{rs}) \tag{4.4}$$

$$\sum_{a \in A_E^k} x_a \leq e_k \qquad (k \in K) \tag{4.5}$$

$$\sum_{a \in I_n} x_a - \sum_{a \in O_n} x_a = 0 \qquad (k \in K; n \in N^k) \tag{4.6}$$

$$x_a \geq 0, \text{ integer} \qquad (k \in K; a \in A^k) \tag{4.7}$$

$$y_{rs} \in \{0, 1\} \qquad (r \in R; s \in S^r). \tag{4.8}$$

The objective function (4.1) minimizes the sum of operational costs. Constraints (4.2) require that each train leg be part of exactly one sequence covered by an appropriate consist. Constraints (4.3) and (4.4) impose lower and upper bounds on sequence arcs of all networks depending on the choice of sequences and consist types. Equipment availability is respected at any time via constraints (4.5). Flow conservation at every node for each equipment type is imposed by constraints (4.6). Finally, all x_a variables must assume non-negative integers values, while y_{rs} variables are restricted to be binary.

4.3 Extensions to the Basic Model

4.3.1 Maintenance Constraints

When performing the assignment of locomotives and cars to trains, planners usually must take into consideration some form of maintenance requirements. These requirements may be expressed in different ways but the most popular approach is to specify, for each type of equipment, a maximum number of days between two successive stops at a maintenance center for any unit of that type. Maintenance centers are usually attached to stations of the physical network and are thus accessible directly at the end of some legs. These regular stops at maintenance centers are necessary to allow for minor repairs and to comply with safety regulations. Although the solution to the basic model can sometimes satisfy these requirements, maintenance constraints must normally be imposed explicitly if maintenance frequency is high or if the set of stations where maintenance can be performed is limited.

Suppose that every unit of equipment must make a stop in one of a specified set of stations at least once every m units of time. Our approach to impose these constraints is to replace the single-commodity network G^k associated with each type of equipment k with a multi-commodity network. Let D represent the set of commodities. Commodity $d \in D$ will correspond to paths in the graph G^k starting at time d and finishing at most m units of time later. To reduce the cardinality of D, one may discretize the planning period as follows. Let t denote the length of the planning period. Divide the planning period $[0, t-1]$ into p disjoint but consecutive subperiods of equal length. For each subperiod $j = 1, \ldots, p$, let a_j and b_j denote the start and the end of this subperiod, respectively. One then obtains $a_1 = 0, b_1 = (t/p) - 1, a_2 = t/p, b_2 = 2(t/p) - 1, \ldots$ Commodity $d \in D$ will then correspond to paths starting from a maintenance center between a_d and b_d and

returning to a maintenance center before $a_d + m \mod t$. Then, by making sure that a unit of flow in graph G^k can only switch commodities at special nodes representing maintenance activities, it will not be possible to find a path in G^k that lasts more than m units of time without visiting a maintenance center. On the other hand, if more frequent visits to a maintenance center can yield a solution with a smaller cost, this will be allowed by the model.

Using a discretization of the planning period introduces a small error in enforcing maintenance constraints since two units leaving the maintenance center at times a_j and b_j must both return to a maintenance center before time $a_j + m \mod t$ even though the second unit has left the maintenance center $b_j - a_j$ units of time later than the first. This error decreases as the cardinality of D increases. For example, using 7 commodities in a 7-day planning horizon gives a maximum error of 24 hours while using 28 commodities decreases this maximum error to 6 hours. For most practical applications, this approximation error can be tolerated since maintenance operations are not synchronized with great precision. This approach is also conservative: no path can exceed m units of time between two maintenance stops.

The network for each type of equipment is augmented as follows. For each train leg whose destination station has an associated maintenance center and after which maintenance is allowed, one introduces an additional maintenance node representing a maintenance activity taking place after the arrival of the train. One also introduces an arc from the arrival node of the leg to the maintenance node. Then, for each part of the station, there is an arc from the maintenance node to the first node in that part of the station corresponding to a time larger than or equal to the arrival time plus the maintenance duration. Hence, if the station is divided between eastbound and westbound trains, two such arcs are required. This is illustrated in Figure 4.2. Finally, additional flow conservation constraints at maintenance nodes will link commodities

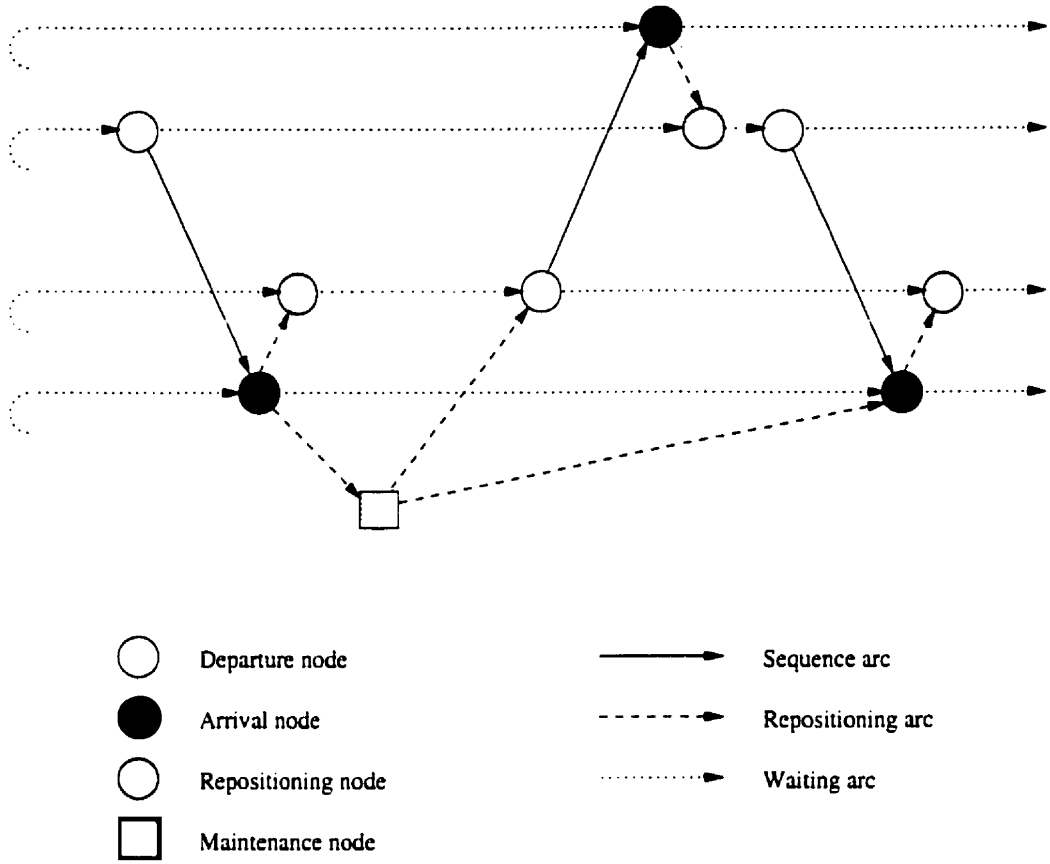

Figure 4.2: Modified network to incorporate maintenance constraints

and allow units of flow to switch commodities when passing by these nodes which act as sources and sinks for the different commodities.

For each equipment type k, variable x_a associated with arc $a \in A^k$ is now replaced by a set of variables x_a^d ($d \in D$). However, x_a^d is defined if and only if arc a could belong to a path associated with commodity d. More precisely, variable x_a^d is defined if and only if both the tail and head nodes of arc a are associated with events that occur between a_d and $a_d + m \mod t$. For each equipment type k, let $M^k \subset N^k$ be the subset of maintenance nodes in the graph G^k. The time associated with a maintenance node $n \in M^k$ is taken as the arrival time of the train leg with which it is associated. Hence, any unit of flow of commodity $d \in D$ has to enter a maintenance

node at or before time $a_d + m \mod t$. Arcs linking the arrival node of a train leg to the maintenance node can be given a positive cost so as to minimize maintenance frequency while still satisfying the minimum requirements.

Given these definitions, the following model may then be used to appropriately enforce maintenance constraints:

$$\text{Minimize} \quad \sum_{k \in K} \sum_{a \in A^k} \sum_{d \in D} f_a x_a^d \tag{4.9}$$

subject to

$$\sum_{r \in R} \sum_{s \in S^r} d_{ls} y_{rs} = 1 \qquad (l \in L) \tag{4.10}$$

$$\sum_{d \in D} x_a^d - \ell_a y_{rs} \geq 0 \qquad (r \in R; s \in S^r; a \in A_{rs}) \tag{4.11}$$

$$\sum_{d \in D} x_a^d - u_a y_{rs} \leq 0 \qquad (r \in R; s \in S^r; a \in A_{rs}) \tag{4.12}$$

$$\sum_{a \in A_E^k} \sum_{d \in D} x_a^d \leq e_k \qquad (k \in K) \tag{4.13}$$

$$\sum_{d \in D} \sum_{a \in I_n} x_a^d - \sum_{d \in D} \sum_{a \in O_n} x_a^d = 0 \qquad (k \in K; n \in M^k) \tag{4.14}$$

$$\sum_{a \in I_n} x_a^d - \sum_{a \in O_n} x_a^d = 0 \qquad (k \in K; n \in N^k \setminus M^k; d \in D) \tag{4.15}$$

$$x_a^d \geq 0, \text{ integer} \qquad (k \in K; a \in A^k; d \in D) \tag{4.16}$$

$$y_{rs} \in \{0, 1\} \qquad (r \in R; s \in S^r). \tag{4.17}$$

Whereas constraints (4.10) are identical to their counterparts (4.2) of the basic model, the objective function (4.1) as well as constraint sets (4.3), (4.4) and (4.5) are modified by replacing each variable x_a by the sum of x_a^d variables for $d \in D$. Hence, each train sequence can now be covered by units of equipment that were last inspected

at different moments in time. In addition, flow conservation constraints are now divided into two groups. For each equipment type, linking constraints (4.14) enforce flow conservation between commodities at all maintenance nodes while constraints (4.15) ensure that flow conservation is satisfied for each commodity at all departure, arrival and repositioning nodes. Thus, decomposing the solution of the problem will yield cycles that change commodities at least once every m units of time in one of the available maintenance centers.

4.3.2 Equipment Switching Penalties

Equipment switching is said to occur after a given train leg l_i if there are at least two units of equipment used on leg l_i such that one of them is next used on leg l_j while the other one is next used on leg $l_k \neq l_j$. While equipment switching enables the railway to decrease its fuel and maintenance expenses by reducing the total number of miles traveled by inactive units, it can also be a source of operating delays since separating or assembling cars and locomotives requires a certain time that may vary according to station congestion and resources availability. Hence, although switching must be performed at least to some extent, it is sometimes desirable to limit such consist modifications when they may have a negative impact on schedule compliance.

In the basic model, the time associated with an arrival node includes the minimum time needed for switching under ideal operating conditions. Thus, switching is forbidden between two legs if the connection time between these legs is less than the minimum switching time. We now explain how a penalty can be imposed to switchings that are feasible but occur shortly after the arrival of a train. Because they may cause some trains to be delayed in situations of high station congestion, such switchings should be allowed but minimized.

Consider a train leg $l \in L$ covered in a sequence that terminates with leg l. After the arrival of the train, the consist used on that leg can either be used unmodified on a different leg or its equipment units can be separated and recombined with other units to form new outbound trains. In the former case, the consist will perform a *direct connection* whereas in the latter, it will perform a *switching connection*. Because a penalty should be imposed when switching is performed shortly after the arrival, one must also distinguish between *short* and *long* switching connections. These different possibilities are illustrated in Figure 4.3. Each arrival node is the tail of a short switching arc, a long switching arc and a certain number of direct connection arcs that link the arrival node to departure nodes of other legs. By imposing the constraint that exactly one possibility be chosen, either all units will perform a direct connection to the same next train leg or they will all perform a short or long switching connection. Similarly, a train consist leaving the station can either perform a direct connection from a previous train or it can be formed by assembling units that were previously switched and reassembled after the arrival of preceding trains.

For each consist type $r \in R$, let C^r denote the set of all possible direct and switching connections for equipment units used in a consist of type r. The set of connections may differ for each consist type since its operating speed affects the arrival times of the trains. For each equipment type k, additional nodes and arcs must be introduced in the sets N^k and A^k to represent these various possibilities. Indeed, for each consist type $r \in R$ and each connection $c \in C^r$, one must introduce a new arc if equipment k is required in a consist of type r. Let $A^k_C \subset A^k$ denote the set of all arcs associated with connections. Then, for every arc $a \in A^k_C$, let c_a and r_a denote, respectively, the associated connection and consist type.

Additional variables and constraints are also needed to ensure that all units of equipment used in the same train consist will perform the same connection. For each consist type $r \in R$ and each connection $c \in C^r$, define a binary variable w_{rc} equal to

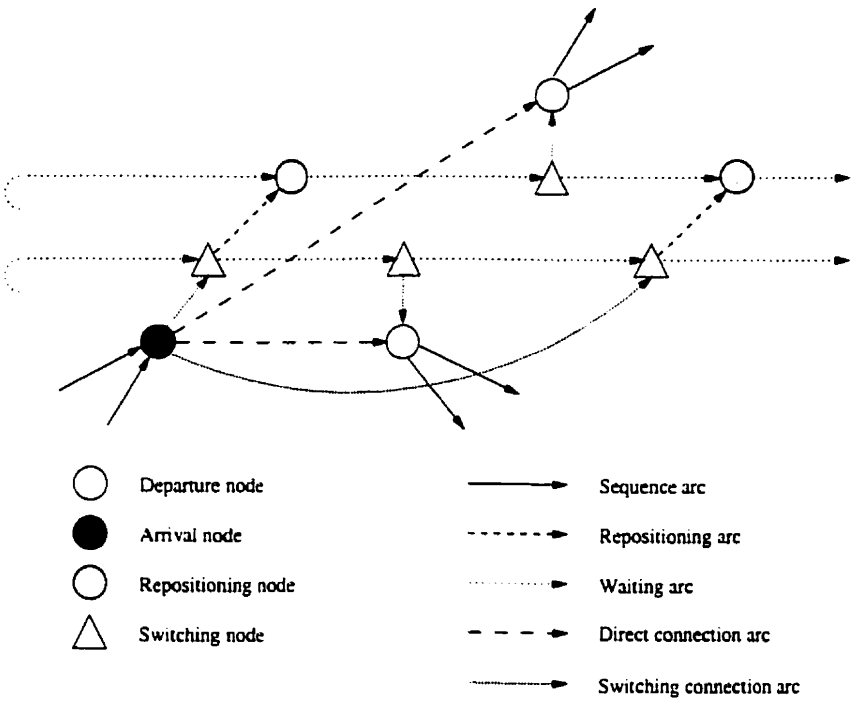

Figure 4.3: Modified network to incorporate car switching penalties

1 if and only if the given connection is performed. For each consist type $r \in R$, let L^r denote the set of train legs on which a consist of type r can be used. Then, for each consist type $r \in R$ and each train leg $l \in L^r$, let $S^r_+(l) \subseteq S^r$ and $S^r_-(l) \subseteq S^r$ designate, respectively, the subsets of train sequences that begin and terminate with train leg l. Similarly, let $C^r_-(l) \subseteq C^r$ and $C^r_+(l) \subseteq C^r$ represent the sets of feasible connections before and after leg l. Then, the constraints

$$\sum_{s \in S^r_+(l)} y_{rs} - \sum_{c \in C^r_-(l)} w_{rc} = 0 \qquad (r \in R; l \in L^r) \tag{4.18}$$

$$\sum_{s \in S^r_-(l)} y_{rs} - \sum_{c \in C^r_+(l)} w_{rc} = 0 \qquad (r \in R; l \in L^r) \tag{4.19}$$

$$\sum_{d \in D} x^d_a - u_a w_{r_a c_a} \leq 0 \qquad (k \in K; a \in A^k_C) \tag{4.20}$$

$$w_{rc} \in \{0, 1\} \qquad (r \in R; c \in C^r) \tag{4.21}$$

must be added to model (4.9)-(4.17) to impose switching penalties. Given (4.10), constraints (4.18) ensure that exactly one connection is chosen before the departure of leg l if this leg is covered by a consist of type r in a sequence that begins with leg l. Constraints (4.19) serve the same purpose for sequences that terminate with leg l. For every $k \in K$ and every $a \in A_C^k$, define u_a as an upper bound on the number of units of equipment k on arc a. Then, constraints (4.20) guarantee that a connection arc is not used unless the corresponding connection is chosen. Finally, penalties are imposed by adding the term $\sum_{r \in R} \sum_{c \in C^r} f_{rc} w_{rc}$ to the objective function, where f_{rc} is the non-negative cost of connection c.

4.3.3 Equipment Substitutions

The last extension concerns the possibility of using a unit of equipment of a given type i where a unit of type j is required. Consider the set J of substitution possibilities: $J = \{(i,j) \mid i,j \in K$ and type i can be substituted for type $j\}$. For each consist type $r \in R$ and each sequence $s \in S^r$, let $J_{rs} \subseteq J$ be the set of allowed substitutions in a consist of type r used on sequence s. Then, for each $r \in R$, each $s \in S^r$ and each $(i,j) \in J_{rs}$, define a non-negative integer variable v_{rs}^{ij} indicating the number of units of type i substituted for units of type j in a consist of type r on sequence s, and let f_{rs}^{ij} denote the cost (or penalty) associated with the substitution of one such unit. Then, the term

$$\sum_{r \in R} \sum_{s \in S^r} \sum_{(i,j) \in J_{rs}} f_{rs}^{ij} v_{rs}^{ij}$$

must be added to the objective function.

For every $k \in K$ and every $a \in A^k$, recall that k_a is the equipment type associated with arc a. Lower bound constraints (4.11) are then replaced with the following:

$$\sum_{d \in D} x_a^d - \ell_a y_{rs} - \sum_{j \in K} v_{rs}^{k_a j} + \sum_{i \in K} v_{rs}^{i k_a} \geq 0 \qquad (r \in R; s \in S^r; a \in A_{rs}) \quad (4.22)$$

$$v_{rs}^{ij} \geq 0, \text{ integer} \qquad (r \in R; s \in S^r; (i,j) \in J_{rs}) \quad (4.23)$$

These ensure that the flow on arc $a \in A_S^k$ satisfies the minimum requirement l_a, plus substitutions of equipment k_a for other types j, minus substitutions of other types i for equipment k_a.

4.4 Solution Methodology

Even for small instances of the problem, model (4.9)-(4.17) contains a very large number of variables and constraints. The approach that we propose to solve this model consists of first relaxing the integrality requirements on the x_a flow variables and gradually imposing these constraints by a branch-and-bound method. At each node of the tree, one thus obtains a mixed-integer problem (integrality is still required on the y_{rs} variables) that is solved by a Benders decomposition (BENDERS, 1962). The subproblem in this decomposition is the LP relaxation of a set of multi-commodity network flow problems that can be solved either by the simplex algorithm or by a Dantzig-Wolfe decomposition. When the mixed-integer problem is feasible at a given node of the tree, a feasible integer solution to (4.9)-(4.17) can also be computed by solving the multi-commodity network flow subproblems with the added integrality requirements. This feasible integer solution provides an upper bound that can be used to prune branches of the search tree. We first present this approach on model (4.9)-(4.17) and then describe the adaptations that are required to deal with switching penalties and substitution possibilities.

4.4.1 Benders Decomposition

Let $\boldsymbol{Y}$ be the set of binary vectors satisfying constraints (4.10) and (4.17). For a given $\bar{\boldsymbol{y}} \in \boldsymbol{Y}$, model (4.9)-(4.17) reduces to the following problem:

$$v(\bar{\boldsymbol{y}}) = \text{Minimize} \quad \sum_{k \in K} \sum_{a \in A^k} \sum_{d \in D} f_a x_a^d \tag{4.24}$$

subject to

$$\sum_{d \in D} x_a^d \geq l_a \bar{y}_{rs} \qquad (r \in R; s \in S^r; a \in A_{rs}) \tag{4.25}$$

$$\sum_{d \in D} x_a^d \leq u_a \bar{y}_{rs} \qquad (r \in R; s \in S^r; a \in A_{rs}) \tag{4.26}$$

$$\sum_{a \in A_E^k} \sum_{d \in D} x_a^d \leq e_k \qquad (k \in K) \tag{4.27}$$

$$\sum_{d \in D} \sum_{a \in I_n} x_a^d - \sum_{d \in D} \sum_{a \in O_n} x_a^d = 0 \qquad (k \in K; n \in M^k) \tag{4.28}$$

$$\sum_{a \in I_n} x_a^d - \sum_{a \in O_n} x_a^d = 0 \qquad (k \in K; n \in N^k \setminus M^k; d \in D) \tag{4.29}$$

$$x_a^d \geq 0, \text{ integer} \qquad (k \in K; a \in A^k; d \in D). \tag{4.30}$$

Model (4.24)-(4.30) decomposes into a multi-commodity network flow problem for each type of equipment $k \in K$, where the commodities are the elements of the set D. Hence, model (4.24)-(4.30) does not have the integrality property. It may also fail to be feasible even when its linear relaxation is feasible. In any case, the *primal subproblem* designates the linear relaxation obtained by dropping integrality requirements in model (4.24)-(4.30).

Let $\boldsymbol{\beta} = (\beta_a \geq 0 | r \in R; s \in S^r; a \in A_{rs})$, $\boldsymbol{\gamma} = (\gamma_a \leq 0 | r \in R; s \in S^r; a \in A_{rs})$, $\boldsymbol{\delta} = (\delta_k \leq 0 | k \in K)$, $\boldsymbol{\eta} = (\eta_n | k \in K; n \in M^k)$ and $\boldsymbol{\pi} = (\pi_n | k \in K; n \in N^k \setminus M^k)$ be the dual variables associated with constraints (4.25)-(4.29), respectively. The dual of the primal subproblem, called the *dual subproblem*, can be expresses as

$$\text{Maximize} \quad \sum_{r \in R} \sum_{s \in S^r} \sum_{a \in A_{rs}} (\ell_a \bar{y}_{rs} \beta_a + u_a \bar{y}_{rs} \gamma_a) + \sum_{k \in K} e_k \delta_k \tag{4.31}$$

subject to

$$(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\pi}) \quad \in \quad \boldsymbol{\Delta}, \tag{4.32}$$

where $\boldsymbol{\Delta}$ denotes the polyhedron defined by the constraints of the problem.

Observe that the set $\boldsymbol{\Delta}$ does not depend on $\bar{\boldsymbol{y}}$ since this vector appears only in the objective function of the dual subproblem. The dual subproblem has one constraint for each x_a^d variable in the primal subproblem. But because each of these variables is non-negative, all constraints of the dual are of the form $\leq f_a$. Hence, $\boldsymbol{\Delta} \neq \varnothing$ whenever $f_a \geq 0$ $(k \in K; a \in A^k)$ since the null vector $\mathbf{0}$ is then a feasible solution to the dual subproblem.

In these conditions, either the primal subproblem is infeasible or it is feasible and bounded. Let $P_{\boldsymbol{\Delta}}$ and $Q_{\boldsymbol{\Delta}}$ represent the sets of extreme points and extreme rays of $\boldsymbol{\Delta}$, respectively.

If, for a given $\bar{\boldsymbol{y}} \in \boldsymbol{Y}$, one has

$$\sum_{r \in R} \sum_{s \in S^r} \sum_{a \in A_{rs}} (\ell_a \bar{y}_{rs} \beta_a + u_a \bar{y}_{rs} \gamma_a) + \sum_{k \in K} e_k \delta_k \leq 0$$

for all extreme rays $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\pi}) \in Q_{\Delta}$, then the dual subproblem (a maximization problem) is bounded and the primal subproblem is feasible. The optimal value of both problems is then equal to

$$\max_{(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\eta},\boldsymbol{\pi}) \in P_{\Delta}} \sum_{r \in R} \sum_{s \in S^r} \sum_{a \in A_{rs}} (\ell_a \bar{y}_{rs} \beta_a + u_a \bar{y}_{rs} \gamma_a) + \sum_{k \in K} e_k \delta_k.$$

If, however, there exists an extreme ray $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\pi}) \in Q_{\Delta}$ for which

$$\sum_{r \in R} \sum_{s \in S^r} \sum_{a \in A_{rs}} (\ell_a \bar{y}_{rs} \beta_a + u_a \bar{y}_{rs} \gamma_a) + \sum_{k \in K} e_k \delta_k > 0$$

then the dual subproblem is unbounded and the primal subproblem must be infeasible.

Model (4.9)-(4.17) can thus be restated as the following *Benders master problem*:

$$\text{Minimize} \quad z \tag{4.33}$$

subject to

$$z - \left(\sum_{r \in R} \sum_{s \in S^r} \sum_{a \in A_{rs}} (\ell_a \beta_a + u_a \gamma_a) y_{rs} + \sum_{k \in K} e_k \delta_k \right) \geq 0 \qquad ((\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\pi}) \in P_{\Delta}) \quad (4.34)$$

$$\sum_{r \in R} \sum_{s \in S^r} \sum_{a \in A_{rs}} (\ell_a \beta_a + u_a \gamma_a) y_{rs} + \sum_{k \in K} e_k \delta_k \leq 0 \qquad ((\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\pi}) \in Q_{\Delta}) \quad (4.35)$$

$$\sum_{r \in R} \sum_{s \in S^r} d_{ls} y_{rs} = 1 \qquad (l \in L) \tag{4.36}$$

$$y_{rs} \in \{0, 1\} \quad (r \in R; s \in S^r). \tag{4.37}$$

Formulation (4.33)-(4.37) contains an enormous number of constraints. However, most of these constraints are inactive in any optimal solution. Hence, instead of

enumerating all extreme points and extreme rays *a priori*, an iterative algorithm can be used to generate only small subsets of *optimality cuts* (4.34) and *feasibility cuts* (4.35). In the worst case, all extreme points and extreme rays of $\boldsymbol{\Delta}$ will be enumerated. See BENDERS (1962) and CORDEAU *et al.* (1998b) for more details on this algorithm.

4.4.2 Computing Upper Bounds

Whenever the relaxation (mixed-integer problem) is feasible at a given node of the branch-and-bound tree, a heuristic can be used to generate a feasible integer solution to model (4.9)-(4.17). If the mixed-integer problem is feasible, one obtains a vector $\bar{\boldsymbol{y}} \in \boldsymbol{Y}$ that yields a feasible primal subproblem. Integrality constraints can then be imposed on all x_a variables, and the resulting integer programming problem can be solved. If this problem is feasible, any feasible integer solution $\bar{\boldsymbol{x}}$ together with the vector $\bar{\boldsymbol{y}}$ constitute a feasible solution to model (4.9)-(4.17). The cost of this solution provides an upper bound on the optimal value of the problem and it can be used to prune branches of the enumeration tree. In particular, if model (4.9)-(4.17) is feasible and there is no integrality gap in the subproblem, then an optimal solution to the problem can be computed at the first node of the branch-and-bound tree. Otherwise, branching must be performed on the x_a flow variables.

4.4.3 Reintroducing Switching Penalties and Substitutions

We now discuss the modifications that must be made to the solution approach to deal with switching penalties and substitution possibilities.

Let $(\boldsymbol{W}, \boldsymbol{Y})$ be the set of binary vectors satisfying constraints (4.10), (4.17)-(4.19) and (4.21). For a given vector $(\bar{\boldsymbol{w}}, \bar{\boldsymbol{y}}) \in (\boldsymbol{W}, \boldsymbol{Y})$, constraints (4.20) become

$$\sum_{d \in D} x_a^d \leq u_a \bar{w}_{r_a c_a} \qquad (k \in K; a \in A_C^k). \tag{4.38}$$

These constraints are added to the primal subproblem and also affect the dual subproblem. Let $\boldsymbol{\phi} = \{\phi_a \leq 0 | k \in K; a \in A_C^k\}$ be the dual variables associated with constraints (4.38).

If, in addition, constraints (4.11) are replaced with constraints (4.22) to permit substitutions, then a solution to the master problem becomes a triplet $(\bar{\boldsymbol{v}}, \bar{\boldsymbol{w}}, \bar{\boldsymbol{y}})$. Associating the dual variables $\boldsymbol{\beta}$ with (4.22), the objective function of the dual subproblem becomes

$$\sum_{r \in R} \sum_{s \in S^r} \sum_{a \in T^{rs}} \left[(l_a \bar{y}_{rs} + \sum_{j \in K} \bar{v}_{rs}^{k_a j} - \sum_{i \in K} \bar{v}_{rs}^{i k_a}) \beta_a + u_a \bar{y}_{rs} \gamma_a \right] + \sum_{k \in K} \sum_{a \in A_C^k} \phi_a u_a \bar{w}_{r_a c_a} + \sum_{k \in K} e_k \delta_k.$$

For given $r \in R$, $s \in S^r$ and $(i, j) \in J_{rs}$, let a_i and a_j denote the arcs associated with the corresponding substitution. Temporarily redefining P_{Δ} and Q_{Δ} according to the new set of dual variables, one then replaces (4.34) and (4.35) with

$$z - \left[\sum_{r \in R} \sum_{s \in S^r} \left(\sum_{a \in T^{rs}} (l_a \beta_a + u_a \gamma_a) y_{rs} + \sum_{(i,j) \in J_{rs}} (\beta_{a_i} - \beta_{a_j}) v_{rs}^{ij} \right) + \sum_{k \in K} \sum_{a \in A_C^k} \phi_a u_a w_{r_a c_a} + \sum_{k \in K} e_k \delta_k \right] \geq 0 \qquad ((\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{\phi}) \in P_{\Delta})$$

and

$$\sum_{r\in R}\sum_{s\in S^r}\left(\sum_{a\in T^{rs}}(l_a\beta_a+u_a\gamma_a)y_{rs}+\sum_{(i,j)\in J_{rs}}(\beta_{a_i}-\beta_{a_j})v_{rs}^{ij}\right)+$$
$$\sum_{k\in K}\sum_{a\in A_C^k}\phi_a u_a w_{r_a c_a}+\sum_{k\in K}e_k\delta_k\leq 0 \qquad ((\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\eta},\boldsymbol{\pi},\boldsymbol{\phi})\in Q_\Delta).$$

The rest of the solution method is unaffected by these modifications. In addition, upper bounds can still be computed as explained in Section 4.4.2.

4.5 Computational Considerations

In our previous article (CORDEAU *et al.*, 1998b), we proposed several ways to improve the performance of the Benders decomposition algorithm when solving the mixed-integer problem in the context of the basic model. First, we suggested that individual cuts should be generated from the subproblems associated with the different equipment types instead of generating a single cut from the global subproblem. To this purpose, z is replaced with $|K|$ variables z_k $(k \in K)$ in (4.33)-(4.37). Next, we observed that, at each node of the branch-and-bound tree, a significant speed improvement can be obtained by first solving the LP relaxation of the master problem before reintroducing the integrality constraints on the y_{rs} variables (MCDANIEL and DEVINE, 1977). Finally, we presented two approaches to generate initial valid cuts for the master problem so as to reduce the number of iterations of the Benders decomposition algorithm.

All these ideas apply here with very simple modifications. We now discuss computational considerations which are more specific to the model obtained when

considering the extensions discussed in Section 4.3. These refinements are presented independently with regard to model (4.9)-(4.17). However, they can be combined with little effort and are also valid when switching penalties and substitution possibilities are considered.

4.5.1 Generating Cuts from a Relaxation of Maintenance Constraints

Recall that in model (4.9)-(4.17), variable x_a^d is defined only if arc a can belong to a path for commodity d. A simple idea which has proven to be quite effective in accelerating the solution of the problem consists of first solving the relaxation obtained by defining x_a^d for all $d \in D$. This is clearly a relaxation of the problem because maintenance constraints are no longer imposed. However, all extreme points and extreme rays generated when solving this relaxation can be used to initialize the corresponding sets of cuts for solving (4.9)-(4.17).

Indeed, model (4.9)-(4.17) is obtained from this relaxation by setting $x_a^d = 0$ if arc a cannot belong to a path for commodity d. Since restricting a problem corresponds to relaxing the dual of this problem, the polyhedron of the dual subproblem associated with the relaxation is thus contained in that of the dual subproblem of model (4.9)-(4.17). Hence, every feasible point for the dual subproblem of the relaxation is a feasible point for the dual subproblem of (4.9)-(4.17). Therefore, all cuts generated from extreme points and extreme rays when solving the relaxation are still valid for solving the model with the maintenance constraints imposed. These points and rays may lie in the interior of the dual subproblem polyhedron for model (4.9)-(4.17) but they nevertheless yield valid cuts.

Furthermore, one can check that the relaxation described above is equivalent to model (4.1)-(4.8). In fact, any solution to this relaxation can be transformed into a solution to the latter model by setting $x_a = \sum_{d \in D} x_a^d$. The advantage of first solving model (4.1)-(4.8) is that a large number of cuts are then generated from subproblems that are considerably smaller and easier to solve.

4.5.2 Identifying Pareto-optimal Cuts

Whenever the primal subproblem (4.24)-(4.30) is degenerate, there may exist more than one optimal solution to the dual subproblem. Although any of these points leads to a valid optimality cut, some can yield stronger cuts than others. The cut generated from the extreme point $(\boldsymbol{\beta}^1, \boldsymbol{\gamma}^1, \boldsymbol{\delta}^1, \boldsymbol{\eta}^1, \boldsymbol{\pi}^1)$ dominates the cut generated from the extreme point $(\boldsymbol{\beta}^2, \boldsymbol{\gamma}^2, \boldsymbol{\delta}^2, \boldsymbol{\eta}^2, \boldsymbol{\pi}^2)$ if and only if

$$\sum_{r \in R} \sum_{s \in S^r} \sum_{a \in A_{rs}} (\ell_a \beta_a^1 + u_a \gamma_a^1) y_{rs} + \sum_{k \in K} e_k \delta_k^1 \geq \sum_{r \in R} \sum_{s \in S^r} \sum_{a \in A_{rs}} (\ell_a \beta_a^2 + u_a \gamma_a^2) y_{rs} + \sum_{k \in K} e_k \delta_k^2$$

for all $\boldsymbol{y} \in \boldsymbol{Y}$ with strict inequality for at least one point. A cut is Pareto-optimal if no other cut dominates it (MAGNANTI and WONG, 1981).

Let $\boldsymbol{Y}^{LP}$ be the polyhedron defined by (4.10) and the constraints $0 \leq y_{rs} \leq 1$ $(r \in R; s \in R^s)$, and let $ri(\boldsymbol{Y}^{LP})$ denote the relative interior of $\boldsymbol{Y}^{LP}$. For a given vector $\bar{\boldsymbol{y}} \in \boldsymbol{Y}^{LP}$ for which the primal subproblem is feasible, let $v(\bar{\boldsymbol{y}})$ denote the optimal value of the subproblem. To identify an optimal solution to the dual subproblem that yields a Pareto-optimal cut, one can solve the following problem, where $\boldsymbol{y}^0 \in ri(\boldsymbol{Y}^{LP})$:

$$\text{Maximize} \quad \sum_{r\in R}\sum_{s\in S^r}\sum_{a\in A_{rs}}(\ell_a y^0_{rs}\beta_a + u_a y^0_{rs}\gamma_a) + \sum_{k\in K} e_k\delta_k \tag{4.39}$$

subject to

$$\sum_{r\in R}\sum_{s\in S^r}\sum_{a\in A_{rs}}(\ell_a \bar{y}_{rs}\beta_a + u_a \bar{y}_{rs}\gamma_a) + \sum_{k\in K} e_k\delta_k = v(\bar{\boldsymbol{y}}) \tag{4.40}$$

$$(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\pi}) \in \boldsymbol{\Delta}. \tag{4.41}$$

The additional constraint (4.40) ensures that one will choose an extreme point from the set of optimal solutions to the original dual subproblem. Let q be the dual variable associated with constraint (4.40). Instead of solving model (4.39)-(4.41), one can solve the dual auxiliary problem:

$$\text{Minimize} \quad \sum_{k\in K}\sum_{a\in A^k}\sum_{d\in D} f_a x^d_a + v(\bar{\mathbf{y}})q \tag{4.42}$$

subject to

$$\sum_{d\in D} x^d_a + l_a \bar{y}_{rs} q \geq l_a y^0_{rs} \qquad (r\in R; s\in S^r; a\in A_{rs}) \tag{4.43}$$

$$\sum_{d\in D} x^d_a + u_a \bar{y}_{rs} q \leq u_a y^0_{rs} \qquad (r\in R; s\in S^r; a\in A_{rs}) \tag{4.44}$$

$$\sum_{a\in A^k_E}\sum_{d\in D} x^d_a + e_k q \leq e_k \qquad (k\in K) \tag{4.45}$$

$$\sum_{d\in D}\sum_{a\in I_n} x^d_a - \sum_{d\in D}\sum_{a\in O_n} x^d_a = 0 \qquad (k\in K; n\in M^k) \tag{4.46}$$

$$\sum_{a\in I_n} x^d_a - \sum_{a\in O_n} x^d_a = 0 \qquad (k\in K; n\in N^k \setminus M^k) \tag{4.47}$$

$$x^d_a \geq 0 \qquad (k\in K; a\in A^k; d\in D). \tag{4.48}$$

This model is also obtained by introducing the additional variable q in the LP relaxation of (4.24)-(4.30). Hence, solving the problem in this form is very convenient in terms of ease of implementation and computational efficiency since the same basic representation can be used to solve both the subproblem (4.24)-(4.30) and the auxiliary problem (4.42)-(4.48). When the problem is large, significant memory savings can be obtained by using this implementation.

For every consist type $r \in R$ and every sequence $s \in S^r$, let ζ_{rs} be a binary variable equal to 1 if and only if $0 < y_{rs} < 1$. Let also $\epsilon > 0$ be a small positive value such that $\epsilon < 1/\sum_{r \in R} |S^r|$. A point of $ri(\boldsymbol{Y}^{LP})$ can be identified by solving the problem

$$\text{Maximize} \quad \sum_{r \in R} \sum_{s \in S^r} \zeta_{rs} \tag{4.49}$$

subject to

$$\sum_{r \in R} \sum_{s \in S^r} d_{ls} y_{rs} = 1 \qquad (l \in L) \tag{4.50}$$

$$y_{rs} - \epsilon \zeta_{rs} \geq 0 \qquad (r \in R; s \in S^r) \tag{4.51}$$

$$y_{rs} + \epsilon \zeta_{rs} \leq 1 \qquad (r \in R; s \in S^r) \tag{4.52}$$

$$\zeta_{rs} \in \{0, 1\} \qquad (r \in R; s \in S^r). \tag{4.53}$$

Choosing an interior point in this way can possibly lead to the infeasibility of model (4.42)-(4.48) since not all vectors $\boldsymbol{y}^0 \in ri(\boldsymbol{Y}^{LP})$ yield feasible primal subproblems. This can be avoided by iteratively adding feasibility cuts to problem (4.49)-(4.53) until its optimal solution yields a feasible primal subproblem. Since the structure of model (4.49)-(4.53) is similar to that of the master problem for Benders decomposition, the same methodology can be used to generate feasibility cuts.

When solving the integer master problem, one should ideally generate Pareto-optimal cuts from a point $y^0 \in ri(Y^c)$ where Y^c denotes the convex hull of Y. However, identifying such a point is difficult since a description of the convex hull is not available. Instead, one can use a point $y^0 \in ri(Y^{LP})$ but the generated cuts may then be dominated on Y^c although they are not dominated on Y^{LP}.

4.5.3 Solving the Primal Subproblem with a Dantzig-Wolfe Decomposition

As explained in Section 4.4.1, the primal subproblem (4.24)-(4.30) decomposes into one multi-commodity network flow problem for each equipment type $k \in K$. Although these problems can be solved directly by the simplex algorithm, a decomposition approach may be more appropriate when the number of commodities $|D|$ is large.

Consider an arbitrary equipment type $k \in K$. If constraints (4.25)-(4.28) are relaxed, the multi-commodity network flow problem for equipment k decomposes into a set of $|D|$ pure network flow problems. If the number of relaxed constraints is not too large, this problem may be solved by a decomposition approach such as Lagrangian relaxation (GEOFFRION, 1974) or Dantzig-Wolfe decomposition (DANTZIG and WOLFE, 1960).

Let Ω_d^k be the set of feasible paths for commodity $d \in D$. These paths must start in the interval $[a_d, b_d]$ and finish before $a_d + m \mod t$. The elements of Ω_d^k are in one-to-one correspondence with the extreme rays of the polyhedron defined by (4.29) and non-negativity constraints. For every $\omega \in \Omega_d^k$, let θ_ω be the flow on path ω and let f_ω be the cost of sending one unit on this path. Define a binary constant $b_{a\omega}$ equal to 1 if and only if arc $a \in A^k$ belongs to path $\omega \in \Omega^k$. The primal subproblem

(4.24)-(4.30) can be restated as the following master problem:

$$\text{Minimize} \quad \sum_{k\in K}\sum_{d\in D}\sum_{\omega\in\Omega_d^k} f_\omega\theta_\omega \tag{4.54}$$

subject to

$$\sum_{d\in D}\sum_{\omega\in\Omega_d^{k_a}} b_{a\omega}\theta_\omega \geq l_a\bar{y}_{rs} \qquad (r\in R; s\in S^r; a\in A_{rs}) \tag{4.55}$$

$$\sum_{d\in D}\sum_{\omega\in\Omega_d^{k_a}} b_{a\omega}\theta_\omega \leq u_a\bar{y}_{rs} \qquad (r\in R; s\in S^r; a\in A_{rs}) \tag{4.56}$$

$$\sum_{a\in A_E^k}\sum_{d\in D}\sum_{\omega\in\Omega_d^k} b_{a\omega}\theta_\omega \leq e_k \qquad (k\in K) \tag{4.57}$$

$$\sum_{d\in D}\sum_{a\in I_n}\sum_{\omega\in\Omega_d^k} b_{a\omega}\theta_\omega - \sum_{d\in D}\sum_{a\in O_n}\sum_{\omega\in\Omega_d^k} b_{a\omega}\theta_\omega = 0 \qquad (k\in K; n\in M^k) \tag{4.58}$$

$$\sum_{\omega\in\Omega_d^k} b_{a\omega}\theta_w \geq 0, \text{ integer} \qquad (k\in K; a\in A^k; d\in D) \tag{4.59}$$

$$\theta_\omega \geq 0 \qquad (k\in K; d\in D; \omega\in\Omega_d^k). \tag{4.60}$$

Columns for the master problem are generated by solving the subproblem (4.29)-(4.30) with an objective that is updated at every iteration to reflect the new values of the dual variables. For a given equipment type $k\in K$ and a given arc $a\in A^k$, the reduced-cost of arc a is $\tilde{f}_a = f_a-\beta_a-\gamma_a-\delta_k I(a\in A_E^k)-\pi_{j_a}I(j_a\in M^k)+\pi_{i_a}I(i_a\in M^k)$ where i_a and j_a represent, respectively, the tail and head nodes of arc a, and $I(\cdot)$ is the indicator function.

If model (4.54)-(4.60) is feasible, then the optimal values of the dual variables associated with the constraints of the Dantzig-Wolfe master problem are an extreme point of the dual subproblem polyhedron $\mathbf{\Delta}$. Even though only a subset of all columns has been generated, this point is an extreme point of the dual subproblem since

all other constraints of the dual (which correspond to columns that have not been generated) are automatically satisfied. If model (4.54)-(4.60) is infeasible, an extreme ray can be computed by using the big M method. In this case, artificial variables are present in the basis at optimality. However, one has generated all columns (all constraints of the dual) necessary to identify an extreme ray. Generating additional columns would only add already satisfied constraints to the dual problem. The direction of the extreme ray can then be determined by identifying the constraints for which the artificial variable is still basic.

4.6 Computational Experimentation

The development of the model and solution approach proposed in the present paper was motivated by two real-life applications of the locomotive and car assignment problem. The first of these applications concerns the passenger trains operated by VIA Rail Canada in the Québec–Windsor corridor. The second application is a study realized by VIA to evaluate the costs and benefits of a project to increase service frequency and replace its current fleet of locomotives and cars with self-powered car modules. We now describe the data used in the computational experiments and give a summary of the results obtained for each application. All experiments were performed on a Sun Ultra2 workstation (300 MHz). The algorithm is coded in C and uses the CPLEX Callable Library (CPLEX, 1997) to solve linear and integer subproblems.

4.6.1 First Group of Experiments

Description of data sets. The context of the first group of computational experiments is described in detail in our previous article (CORDEAU *et al.*, 1998b)

and we only briefly recall it here. VIA currently uses two types of locomotives (F40 and LRC) and two types of first-class and second-class cars (LRC and HEP) that yield three consist types with different operating speeds: F40 locomotives combine with both LRC and HEP cars but LRC locomotives combine only with LRC cars. Equipment availability is limited and the objective is to minimize the sum of operational costs related to mileage. All train legs begin and terminate in one of the nine major stations of the physical rail network, but switching is allowed only in two of these stations (Montréal and Toronto). For each train leg, demand is expressed as the number of first-class and second-class cars required. Most train legs require a single locomotive but a few exceptions require two.

Three instances corresponding to the schedules of different seasons were used in the experiments. In addition, two variants were considered for each instance. In the first variant (instances 1a to 3a), the type of consist used on each leg is fixed and matches the assignment used by VIA. In the second variant (instances 1b to 3b), more than 50% of all train legs can be covered by either two or three consist types. All these instances correspond to weekly problems.

For each variant of each instance, three scenarios were compared. In the first scenario, maintenance constraints are imposed but switching is not penalized and substitutions are forbidden. Every unit of equipment must be inspected at least once every seven days at the unique maintenance center located in Montréal. Maintenance can be performed after the arrival of any train in that station and the minimum time required for maintenance is five hours. In the second scenario, maintenance constraints are still imposed but switching is now penalized. For example, at the station associated with the maintenance center, the minimum time required for switching is two hours but switching is however penalized if less than five hours are available for connection. Finally, the third scenario incorporates all three extensions to the basic model and adds the possibility to substitute a first-class car for a second-

class car on any train sequence. In all scenarios, a 7-subperiod discretization is used to impose maintenance constraints; this approach is used by VIA for planning purposes.

Table 4.1 reports the size of the Benders master problem and subproblem for each of the six instances when considering the different scenarios. For example, the schedule for instance VIA1a has 330 train legs, leading to a total of 16,368 sequences. Under the first scenario, the master problem (4.33)-(4.37) contains one constraint of the form (4.36) for each train leg, one y_{rs} variable for each sequence, and one z_k cost variable for each of the six equipment types. The primal subproblem (4.24)-(4.30) contains 109,056 constraints and 366,414 flow variables x_a^d. Since the primal subproblem decomposes into a set of six multi-commodity network flow problems, each of them has on average more than 18,000 constraints and 60,000 variables. Under the second scenario, the number of variables and constraints increases slightly following the introduction of connection variables w_{rc} and the associated constraints (4.18)-(4.21). Finally, the number of variables in the master problem nearly doubles in the third scenario since one substitution variable w_{ij}^{rs} is added for every sequence. The size of the subproblem is however not affected by the introduction of substitution possibilities.

Summary of results. To solve the mixed-integer problem at the first node of the branch-and-bound tree, the algorithm actually proceeds in three phases. In phase I, initial cuts are generated by solving the LP relaxation of (4.1)-(4.8) as explained in Section 4.5.1. Then, phase II solves the LP relaxation of (4.9)-(4.17) to optimality. In phase III, integrality is finally imposed on the variables of the master problem, and the algorithm iteratively solves the integer master problem and generates additional cuts until an optimal solution is found for the mixed-integer problem.

The first step in our experiments was to analyze the effects on computing time and convergence of generating initial cuts by solving the LP relaxation of (4.1)-(4.8).

Table 4.1: Model size for Benders decomposition of the first set of instances

			Master problem		Subproblem	
Instance	**Legs**	**Sequences**	**Constraints**	**Variables**	**Constraints**	**Variables**
VIA1a-1	330	16 368	330	16 374	109 056	366 414
VIA1a-2			740	17 101	122 199	388 338
VIA1a-3			740	33 469	122 199	388 338
VIA2a-1	352	17 478	352	17 484	116 178	390 417
VIA2a-2			784	18 270	130 191	414 042
VIA2a-3			784	35 748	130 191	414 042
VIA3a-1	348	14 296	348	14 302	96 822	322 923
VIA3a-2			772	15 073	110 622	346 149
VIA3a-3			772	29 369	110 622	346 149
VIA1b-1	330	26 691	330	26 697	174 250	590 470
VIA1b-2			922	27 763	193 597	622 789
VIA1b-3			922	54 454	193 597	622 789
VIA2b-1	352	25 139	352	25 145	165 491	558 767
VIA2b-2			976	26 301	186 116	593 648
VIA2b-3			976	51 440	186 116	593 648
VIA3b-1	348	22 123	348	22 129	147 131	494 759
VIA3b-2			964	23 287	167 594	529 598
VIA3b-3			964	45 410	167 594	529 598

We have determined that when these initial cuts are not generated, CPU times are clearly excessive because of the large size of the subproblem. On the other hand, if model (4.1)-(4.8) is solved first, then a few additional iterations of the algorithm with subproblem (4.24)-(4.30) are sufficient to find an optimal solution to (4.9)-(4.17). In this application, maintenance constraints are easily satisfied and the optimal solution to (4.9)-(4.17) often differs only slightly from the optimal solution to the maintenance relaxation.

To illustrate the benefits of generating these initial cuts, three smaller instances were obtained by considering an hypothetical scenario in which switching would be permitted in all nine stations of the network. This considerably reduces the number of sequences and the size of the model without affecting the structure of the problem: the number of constraints in the master problem remains the same but the number of variables and the size of the subproblem are divided by a factor of ten. Each of these instances was then solved with and without the initial cuts. Table 4.2 indicates the number of iterations, number of cuts generated and the CPU time (in minutes)

needed to find an optimal solution to the LP relaxation of model (4.9)-(4.17) by the two methods.

Table 4.2: Effect of generating initial cuts from relaxation

	Basic algorithm			Two-phase algorithm					
				Phase I			Phase II		
Instance	Iter.	Cuts	CPU	Iter.	Cuts	CPU	Iter.	Cuts	CPU
VIA1a-0	67	227	66.9	12	40	0.02	1	0	0.72
VIA2a-0	302	943	440.2	37	76	0.05	1	0	0.93
VIA3a-0	342	1025	413.3	21	63	0.04	1	0	0.90

With the two-phase algorithm, computing times are divided by more than one hundred and the number of iterations performed also decreases very significantly. This is explained by the fact that the feasibility cuts generated from the multi-commodity subproblem are weaker than those generated from the single-commodity subproblem of the relaxation. When first solving the maintenance relaxation, a single extra iteration with subproblem (4.24)-(4.30) was sufficient to find an optimal solution satisfying the maintenance constraints. On the large instances of Table 4.1, CPU times exceeded 24 hours when the relaxation of maintenance constraints was not solved first. Thus, initial cuts were generated in all further experiments.

Table 4.3 presents the results obtained when solving each of the six instances under the three scenarios. The numbers indicate the total work for the three phases just described. The CPU time also includes the time needed to compute an integer solution with the upper bounding procedure explained in Section 4.4.2. In all these experiments, this procedure found an optimal solution at the first node of the tree since there was no integrality gap in any of the subproblems. Hence, branching was not required.

When considering only maintenance constraints (scenario 1), all instances are solved in less that 20 minutes and only a few iterations are required to determine an optimal integer solution. When switching penalties (scenario 2) and substitution

Table 4.3: Computational results for the first set of instances

	Scenario 1			Scenario 2			Scenario 3			Maximum
Instance	Iter.	Cuts	CPU	Iter.	Cuts	CPU	Iter.	Cuts	CPU	IP Gap (%)
VIA1a	5	20	9.9	25	104	6.8	47	138	15.5	0.2348
VIA2a	7	24	16.4	29	110	13.5	38	123	60.3	0.1506
VIA3a	6	22	9.3	33	106	16.8	28	87	63.8	0.2119
VIA1b	7	29	14.7	79	323	38.6	83	331	90.0	0.0625
VIA2b	9	41	18.9	74	294	43.1	85	327	87.7	0.0461
VIA3b	7	34	17.5	90	366	29.9	84	341	62.2	0.0376

possibilities (scenarios 3) are introduced, the total effort needed to solve a given instance grows moderately. The CPU time remains reasonable considering that an optimal solution is computed. This good performance is in part explained by the fact that the integrality gap is very small in these instances. For every instance, the largest integrality gap is observed for the third scenario and is always below 0.25%. This gap is explored when solving the integer master problem which has relatively few rows and is solved rather quickly despite its large number of variables.

In our previous article (CORDEAU *et al.*, 1998b), we compared the performance of the Benders decomposition algorithm to those of Lagrangian relaxation and Dantzig-Wolfe decomposition for solving the LP relaxation of the basic model. We also compared our complete algorithm to a simplex-based branch-and-bound method. According to these results, the approach presented here can solve the extended model in less time than what is required by the other approaches for simply solving the basic model. Further comparisons with these methods would thus be pointless.

This performance is also superior to that of the Dantzig-Wolfe decomposition method of CORDEAU *et al.* (1998a). On similar instances, the new method can find an optimal solution in less CPU time than what is needed by the former one to identify an approximate solution. A direct comparison of the two methods is however difficult because they use different modeling approaches. The first one places the emphasis on the minimization of train modifications by penalizing all switchings and using

compound modules containing several units of equipment. The second one places the emphasis on the minimization of mileage costs by penalizing only short switchings and using disaggregated equipment units. In addition, some small features that are necessary in a commercial implementation have been omitted here. These could nevertheless be added to the model with little effect on algorithmic performance.

4.6.2 Second Group of Experiments

Description of data sets. In an alternative studied by VIA, train frequencies would be increased significantly and the current fleet of locomotives and cars would be replaced by a set of self-powered car modules containing two, three or four cars each. Our mandate was to determine the number of modules of each type that should be acquired so as to minimize a weighted combination of capital costs and future operating cots.

The input for these experiments is the expected demand in passengers on a set of 548 train legs from a weekly schedule. The demand on each train leg can be satisfied with at most eight cars and a train consist contains at most two active modules. A consist type is defined for each of the nine possible ways of choosing one or two module types among the three types available. For example, a single three-car module and two three-car modules represent distinct consist types. Then, for each train leg, the set of possible consist types is determined according to the demand: all consist types that provide enough seating capacity can be used to cover the given leg. As in the first application, maintenance must be performed weekly on every module. However, because all modules are self-powered, switching can now be performed in very little time in all nine stations of the network and should not be penalized. Finally, substitutions are not necessary since all cars provide the same type of service.

Three instances were obtained from different evaluations of the demand ranging from light to heavy. In addition, two variants were considered for each instance. In the first variant (instances 4a to 6a), a unique maintenance center is used with a 7-subperiod discretization. In the second variant (instances 4b to 6b), two maintenance centers are available (Montréal and Toronto) and a 28-subperiod discretization is used. Table 4.4 reports the size of each instance. Since switching is allowed in every station, the number of sequences is considerably smaller in this application. On the other hand, the subproblem becomes very large when a finer discretization is considered. Given that there are three equipment types, each multi-commodity network flow subproblem contains approximately 40,000 constraints and 120,000 variables in the second variant.

Table 4.4: Model size for Benders decomposition of the second set of instances

			Master problem		Subproblem	
Instance	**Legs**	**Sequences**	**Constraints**	**Variables**	**Constraints**	**Variables**
VIA4a	548	1 734	548	1 737	29 801	65 894
VIA5a	548	1 743	548	1 746	29 726	66 881
VIA6a	548	1 757	548	1 760	29 642	68 372
VIA4b	548	1 734	548	1 737	121 438	340 782
VIA5b	548	1 743	548	1 746	120 675	348 077
VIA6b	548	1 757	548	1 760	119 621	355 543

Summary of results. The instances in this group of experiments are more difficult to solve that those of the previous group for several reasons. First, the average number of possible consist types for each leg is higher in these problems and it seems that the performance of the algorithm is more affected by the number of consist types than by the number of sequences. Second, maintenance constraints are more difficult to satisfy here and several iterations with the multi-commodity network flow subproblem (4.24)-(4.30) are sometimes needed to find an optimal solution. Finally, these instances include fixed costs that take equipment ownership into consideration. The objective then contains two terms that are in contradiction since reducing fixed costs leads to an increase in equipment utilization and operating costs.

Table 4.5 summarizes the computational statistics obtained when solving each of the first three instances. We report the number of iterations, number of cuts, and CPU time (in minutes) for each of the three phases: solving the LP relaxation of (4.1)-(4.8), solving the LP relaxation of (4.9)-(4.17), and solving the mixed-integer problem. Here again, computing an integer solution at the first node of the search tree provided an optimal solution to the problem. The time needed to compute this solution is reported separately in Table 4.7.

Table 4.5: Computational results for second set of instances (first variant)

	Phase I			**Phase II**			**Phase III**			
Instance	Iter.	Cuts	CPU	Iter.	Cuts	CPU	Iter.	Cuts	CPU	IP Gap (%)
VIA4a	179	463	3.5	85	223	509.3	3	4	82.7	0.0464
VIA5a	216	569	5.2	30	77	199.7	6	10	488.3	0.0364
VIA6a	147	414	2.9	47	120	302.1	1	3	31.9	0.1042

The first phase is completed in a few minutes on all instances although most of the cuts are generated during that phase. In phase II, each iteration of the algorithm takes much longer because three large problems must be solved twice with the simplex algorithm. The primal subproblem (4.24)-(4.30) (which decomposes into three multi-commodity network flow problems) is solved first, followed by the auxiliary problem (4.42)-(4.48) to identify Pareto-optimal cuts. At least 90% of the total CPU time is spent in solving these problems with the largest portion used for the primal subproblem. The third phase requires only a few iterations but each of them takes even longer because the integer master problem must be solved by a branch-and-bound algorithm. This is particularly time-consuming for instance VIA5a since each iteration requires on average more than 45 minutes. The total computing times are large but they are acceptable considering that optimal solutions are computed for a strategic problem of resource acquisition. Again, integrality gaps are very small. These small gaps are a result of the problem formulation: enumerating the set of possible consist types and imposing constraints (4.10) requires that one consist be supplied although a fraction of a consist could sometimes be sufficient if demand

constraints were expressed as a seating capacity to be provided. The LP relaxation of the problem is thus very strong.

Because several iterations are performed in phase II of the algorithm with the multi-commodity subproblem, generating Pareto-optimal cuts as explained in Section 4.5.2 is often necessary to obtain convergence in reasonable time. Indeed, the primal subproblem (4.24)-(4.30) is normally highly degenerate since the bulk of its constraints are flow conservation equations (4.28) and (4.29). As a result, the optimality cuts may be extremely weak if they are generated from an arbitrary dual optimal solution. Figure 4.4 plots the value of the lower bound provided by the master problem and the value of the upper bound provided by the subproblem as a function of CPU time when solving the LP relaxation in Phase II for instance VIA4a.

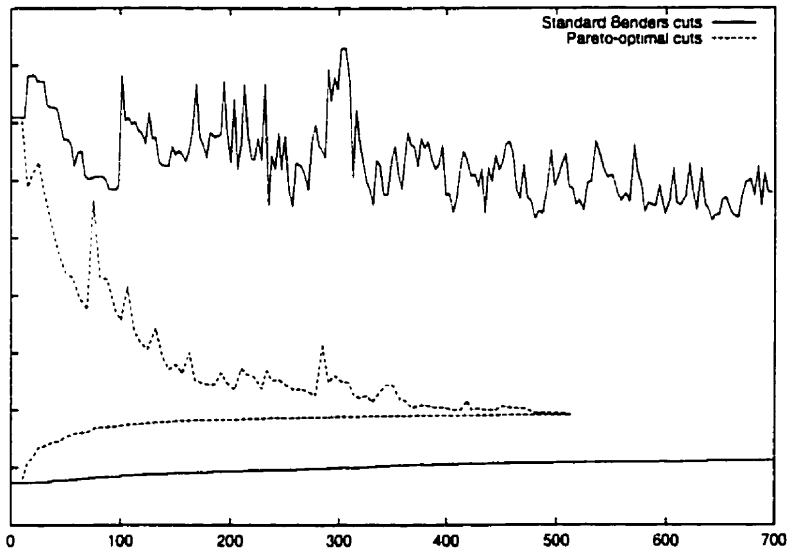

Figure 4.4: Values of lower and upper bounds as a function of CPU time

As the figure indicates, generating Pareto-optimal cuts improves the performance of the algorithm considerably. While the lower bound provided by the value of the master problem increases very slowly when cuts are generated from an arbitrary

optimal solution, this bound grows quickly in the first few iterations when Pareto-optimal cuts are used. Also, while the upper bound provided by the value of the subproblem is obviously not monotonically decreasing in any case, it exhibits a more stable behavior when non-dominated cuts are generated. Using these cuts, the algorithm converged to an optimal solution after 509.3 minutes, as indicated in Table 4.5. Computing Pareto-optimal cuts requires a bit of extra work but allows a very significant reduction of the total CPU time on all instances.

Because two maintenance centers are available in the last three instances, these were slightly easier to solve than the first three. For every instance, a single iteration with subproblem (4.24)-(4.30) was necessary to obtain an optimal solution to the LP relaxation of (4.9)-(4.17). Then, solving the integer master problem only once yielded an optimal solution to the mixed-integer problem at the first node of the branch-and-bound tree. Again, an optimal integer solution was found by solving the subproblem once with the integrality requirements. The corresponding statistics are summarized in Table 4.6 below.

Table 4.6: Computational results for second set of instances (second variant)

	Phase I			**Phase II**			**Phase III**			
Instance	Iter.	Cuts	CPU	Iter.	Cuts	CPU	Iter.	Cuts	CPU	IP Gap (%)
VIA4b	173	468	3.5	1	0	21.4	1	0	38.7	0.1410
VIA5b	194	504	4.5	1	0	25.5	1	0	38.3	0.0612
VIA6b	153	426	3.3	1	0	25.2	1	0	51.1	0.3473

In this case, a different approach was necessary to solve the subproblem (4.24)-(4.30) because of the 28-subperiod discretization. For these instances, the subproblem was optimized with a Dantzig-Wolfe decomposition (see Section 4.5.3). The decomposition approach becomes attractive here because the larger number of subproblems used for column generation does not have a great impact on computing

times. On the other hand, solving the subproblem with the simplex algorithm is very time-consuming because of the large number of constraints.

Table 4.7 reports the CPU time needed to solve the primal subproblem (4.24)-(4.30) once with the simplex algorithm and with Dantzig-Wolfe decomposition. These timings were collected the first time the subproblem was solved in phase II of the algorithm. For the 7-subperiod discretization used in the first three instances, the simplex algorithm is approximately two times faster than Dantzig-Wolfe decomposition. However, this conclusion reverses in the case of the 28-subperiod discretization used in the last three instances. In both cases, the master problem (4.54)-(4.60) of the Dantzig-Wolfe decomposition for each type of equipment contains approximately 1000 constraints. Several independent paths are generated at each iteration of the column generation process by solving a network flow problem with upper bounds set to 1 on all arcs of the network.

Table 4.7: CPU time (in minutes) needed to solve subproblem (4.24)-(4.30)

	LP relaxation solution		**Optimal integer solution**	
Instance	Simplex	D-W decomposition	BB with simplex	BB with D-W decomposition
VIA4a	3.1	6.3	124.2	-
VIA5a	2.9	7.3	97.8	-
VIA6a	3.1	6.1	85.2	-
VIA4b	73.7	21.4	-	320.6
VIA5b	80.5	25.5	-	475.7
VIA6b	75.6	25.2	-	221.3

This table also indicates the CPU time required to solve the integer subproblem by branch-and-bound. These were computed only for the faster of the two methods in each case. When solving the subproblem with a simplex-based branch-and-bound method, strong branching is used with a best-bound search. When the linear relaxations are solved with Dantzig-Wolfe decomposition, a depth-first search is used and branching is performed on the path variable whose value is closest to the next integer. To obtain the total time required for solving a given instance, one must add

the time from the last two columns of Table 4.7 to those of the three phases. For example, instance VIA4b required a total of 384.2 minutes of which 320.6 were spent solving the integer subproblem.

4.7 Conclusions

In this paper, we have presented a basic formulation and three extensions for the simultaneous assignment of locomotives and cars to trains in the context of passenger transportation. The resulting model is a robust and flexible starting point for the development of optimization systems capable of handling large and complex problems that occur in actual operations. The model is also very tractable and can be solved to optimality in reasonable time for instances of realistic size. Our solution method combines relaxation and decomposition principles in an efficient manner that takes advantage of several problem characteristics. The approach was used in practice to determine the best mix of equipment that a railway should acquire so as to minimize a combination of capital and operating costs.

The model is particularly useful in tactical and strategic planning but does not address the operational planning problem which deals with the daily operations of a railway. Short-term planning requires that several factors such as train delays and equipment position and orientation on the train be taken into account. In addition, fast solution methods are necessary so that the model can be used in real-time to analyze different scenarios or determine the changes to be made following a mechanical failure or train delay. Given the separability of our formulation and the fact that the subproblem for each equipment type can itself be decomposed, the approach

introduced here can be adapted to deal with the operational problem. These further extensions will be addressed in subsequent research.

Acknowledgments

We thank Mr. Steve Del Bosco and Mr. Alain Vigeant of VIA Rail Canada who provided the data used in the computational experiments. This work was supported by the Québec Government (Fonds pour la Formation de Chercheurs et l'Aide à la Recherche), the Natural Sciences and Engineering Research Council of Canada, and by Ad Opt Technologies Inc.

Conclusion

La première contribution de cette thèse est de présenter un cadre de modélisation à la fois général et détaillé pour l'affectation des locomotives et des wagons aux trains de passagers. Ce cadre original s'inspire en partie des approches utilisées en transport ferroviaire de marchandises et en transport aérien mais introduit également de nombreux éléments de modélisation qui sont propres au transport ferroviaire de passagers.

Tout d'abord, la prise en compte des incompatiblités et des interdépendances de nature temporelle entre les différents types d'équipement requiert une approche spécifique au problème étudié. En fait, les modèles proposés jusqu'à maintenant pour l'affectation des locomotives aux trains de marchandises ne permettent de traiter que les incompatibilités entre les trains et les types de locomotives et ne considèrent pas les incompatiblités entre les types de locomotives eux-mêmes. De plus, aucun de ces modèles ne considère l'effet des combinaisons d'équipement sur la vitesse d'opération des trains. Finalement, ces modèles ne considèrent pas les temps de connexion variables en fonction du type de connexion ou les effets du couplage et du découplage. Ce type de difficultés n'est par ailleurs pas présent dans les problèmes d'affectation d'équipement en transport aérien puisqu'un seul appareil est utilisé sur chaque vol.

Comme en témoignent les différents modèles présentés dans la thèse, le cadre de modélisation permet aussi de traduire un très large éventail de possibilités et de contraintes qui peuvent être communes ou spécifiques à différentes applications pratiques. En particulier, les contraintes d'entretien, les possibilités de substitution

et les pénalités pour le couplage et le découplage des wagons peuvent être prises en compte sans trop affecter la structure de base des modèles.

Une autre contribution importante de cette thèse est le développement et la comparaison de diverses approches de résolution pour les modèles proposés. Au chapitre 2, nous avons d'abord utilisé une approche basée sur la génération de colonnes pour résoudre un problème pratique avec un objectif et des contraintes complexes. Cette approche permet de résoudre de manière approximative, en quelques heures de temps de calcul, des instances comportant six types d'équipement et plus de 300 trains par semaine. Les comparaisons avec les solutions produites manuellement par les employés de VIA Rail montrent que cette approche permet habituellement de réduire à la fois les coûts et le nombre de couplages et de découplages de wagons. De plus, ces résultats indiquent que des économies considérables peuvent être réalisées au niveau des coûts variables d'opération en augmentant très légèrement le nombre de fois qu'un wagon change de locomotive à l'extérieur du centre d'entretien. Des économies de l'ordre de 10% peuvent souvent être réalisées en augmentant de quelques unités le nombre de couplages ou de découplages effectuées durant une semaine d'opération.

La principale faiblesse de cette approche est que le temps de calcul augmente très rapidement lorsqu'on considère la possibilité de choisir de manière endogène la combinaison d'équipement à utiliser sur chaque train. Par exemple, en considérant deux possiblités pour 30% des trains et une seule possibilité pour les autres, le temps de calcul pour une des instances est passé d'environ 3 heures à plus de 14 heures. Au chapitre 3, nous avons donc présenté un second modèle, plus simple, mais mieux adapté à la possibilité de pouvoir choisir parmi plusieurs combinaisons celle utilisée sur chaque train. En ayant recours à une formulation différente du problème, nous avons donc obtenu un modèle pour lequel la méthode de décomposition de Benders fournit une approche de résolution très efficace. En combinant certaines

améliorations à l'algorithme de base, cette approche permet par exemple de résoudre à l'optimalité en moins de 15 minutes des instances dans lesquelles deux ou trois combinaisons d'équipement sont possibles pour plus de la moitié des trains. De plus, les comparaisons avec une méthode de séparation et d'évaluation progressive basée sur la résolution de programmes linéaires par l'algorithme du simplexe indiquent que la méthode de décomposition est au moins dix fois plus rapide. Ce gain s'explique en bonne partie par le fait que les contraintes de demande et de capacité deviennent de simples bornes sur les arcs d'un réseau lorsque la décomposition de Benders est utilisée.

Le modèle simplifié incorpore cependant trop peu d'éléments pour être utilisé dans des applications pratiques. Dans le dernier chapitre, nous avons donc présenté une généralisation du modèle simplifié qui incorpore les contraintes d'entretien, des pénalités de couplage et découplage des wagons ainsi que les possiblités de substitution. Le modèle résultant possède donc un niveau de détail semblable au premier modèle du chapitre 2 mais conserve une structure propice à l'utilisation de la décomposition de Benders. Ce nouveau modèle vise par ailleurs à être plus général que le premier en permettant facilement la présence de plusieurs centres d'entretien et de trains fonctionnant durant la nuit. En résolvant d'abord la relaxation correspondant au modèle du chapitre 3, cette nouvelle approche permet de résoudre, à l'optimalité et avec toute la gamme des contraintes considérés au chapitre 2, des instances de même taille que précédemment en moins de 90 minutes de temps de calcul.

Une comparaison directe des deux approches est cependant difficile car elles traitent des variantes légèrement différentes du problème. Par exemple, alors que le premier modèle considère des modules composés de plusieurs unités d'équipement et pénalise tous les couplages et découplages de wagons, le second considère des équipements désaggrégés et n'impose des pénalités qu'aux modifications apportées

aux trains lors de connexions courtes. Les contraintes de capacité des locomotives sont également traitées différemment: afin de conserver la séparabilité du sous-problème, le second modèle répartit *a priori* la capacité des locomotives entre les différents types de wagons. Finalement, la dernière approche néglige certaines fonctionalités nécessaires dans un logiciel commercial telles que la possibilité de violer certaines contraintes moyennant une pénalité de façon à assurer la réalisabilité du problème.

L'utilisation d'une approche basée sur la génération de colonnes dans la première application s'explique entre autres par le fait que le projet a débuté avant que ne s'effectue le développement du modèle présenté au chapitre 3. En utilisant le concept d'équipement de base, le premier modèle permet par ailleurs de mieux contrôler la fréquence des couplages et découplages de wagons, ce qui constitue un objectif important pour VIA Rail. Le modèle du chapitre 4 permet aussi l'utilisation de modules mais ceux-ci doivent être formés *a priori* et les unités d'équipement qu'ils contiennent ne peuvent être recombinées au cours de la période. Dans le cas où la disponibilité de l'équipement est très contraignante, ceci constitue une restriction importante. Finalement, le besoin de disposer d'un algorithme robuste pouvant être incorporé à un logiciel commercial a motivé le choix de la génération de colonnes pour le développement du logiciel implanté chez VIA Rail.

Bien que cette thèse contienne très peu de développements théoriques concernant les méthodes de décomposition utilisées, elle fournit néanmoins beaucoup d'informations utiles sur leur utilisation pratique. En particulier, les nombreuses idées proposées pour accélérer l'algorithme de décomposition de Benders peuvent être appliquées à plusieurs autres problèmes. L'expérimentation a d'abord fait ressortir l'importance de générer un bon ensemble de coupes initiales afin d'assurer une convergence rapide de l'algorithme. Elle a également témoigné du fait qu'un gain important de rapidité peut être obtenu en résolvant d'abord la relaxation linéaire du problème. Dans le cas

du modèle étendu incluant les contraintes d'entretien, la méthode de décomposition de Benders est accélérée encore davantage en résolvant d'abord une relaxation du problème. Finalement, l'importance de générer des coupes Pareto-optimales a été clairement illustrée lors de la résolution de certains problèmes.

Les résultats présentés dans les deux derniers chapitres de la thèse confirment que la décomposition de Benders peut être une méthode de résolution très efficace lorsque le problème possède une structure appropriée. Dans le modèle du chapitre 3, les contraintes liantes deviennent de simples bornes sur les arcs d'un réseau lorsque sont fixées les valeurs des variables du problème maître. Dans le modèle du chapitre 4, une telle simplification n'est plus possible car ces contraintes lient les arcs des modèles multi-flots. En résolvant une relaxation du problème, on peut cependant générer rapidement un sous-ensemble des coupes nécessaires pour identifier une solution optimale. Un nombre réduit d'itérations avec le modèle complet est ensuite suffisant pour atteindre l'optimalité.

Les approches présentés dans cette thèse permettent de traiter le problème de planification tactique tout en incorporant un niveau de détail relativement élevé quant à l'opération des trains. Ces approches peuvent également être utilisées pour résoudre des problèmes de planification à plus long terme tels que celui de déterminer la composition optimale de la flotte. Des approches similaires pourraient par ailleurs être utilisées pour traiter d'autres problèmes dans lesquels doivent être affectées des unités de différentes natures. Par exemple, l'affectation simultanée d'appareils et d'équipages aux vols d'un transporteur aérien pourrait se faire en utilisation une approche semblable à celles décrites dans cette thèse.

Dans leur forme actuelle, ces approches ne permettent cependant pas de résoudre le problème de gestion opérationnelle de l'équipement. Un développement subséquent

consisterait à adapter les approches proposées de manière à traiter ce problème qui requiert que l'on tienne compte de la position et de l'orientation de chaque locomotive et de chaque wagon dans le réseau. Au niveau opérationnel, des méthodes de résolution très rapides sont en général nécessaires afin d'obtenir rapidement une bonne solution suite à un changement ou un retard se produisant dans le réseau.

Bibliographie

ADAMIDOU, E. A., KORNHAUSER, A. L. et KOSKOSIDIS, Y. A. (1993). A game theoretic/network equilibrium solution approach for the railroad freight car management problem. *Transportation Research*, 27B:237–252.

ALLMAN, W. P. (1972). An optimization approach to freight car allocation under time-mileage per diem rental rates. *Management Science*, 18B:567–574.

ASSAD, A. A. (1980a). Modelling of rail networks: Toward a routing/makeup model. *Transportation Research*, 14B:101–114.

ASSAD, A. A. (1980b). Models for rail transportation. *Transportation Research*, 14A:205–220.

ASSAD, A. A. (1981). Analytical models in rail transportation: An annotated bibliography. *INFOR*, 19:59–80.

ASSAD, A. A. (1982). A class of train-scheduling problems. *Transportation Science*, 16:281–310.

ASSAD, A. A. (1983). Analysis of rail classification policies. *INFOR*, 21:293–314.

AVI-ITZHAK, B., BENN, B. A. et POWELL, B. A. (1967). Car pool systems in railroad transportation: Mathematical models. *Management Science*, 13:694–711.

AVRAMOVIĆ, Z. Ž. (1995). Method for evaluating the strength of retarding steps on a marshalling yard hump. *European Journal of Operational Research*, 85:504–514.

BARTLETT, T. E. (1957). An algorithm for the minimum number of transport units to maintain a fixed schedule. *Naval Research Logistics Quarterly*, 4:139–149.

BARTLETT, T. E. et CHARNES, A. (1957). Cyclic scheduling and combinatorial topology: Assignment and routing of motive power to meet scheduling and maintenance requirements; Part II generalization and analysis. *Naval Research Logistics Quarterly*, 4:207–220.

BEAUJON, G. J. et TURNQUIST, M. A. (1991). A model for fleet sizing and vehicle allocation. *Transportation Science*, 25:19–45.

BECKMANN, M., MCGUIRE, C. B. et WINSTON, C. B. (1956). *Studies in the Economics of Transportation*. Yale University Press, New Haven, CT.

BEN-KHEDER, N., KINTANAR, J., QUEILLE, C. et STRIPLING, W. K. (1997). Decision support scheduling systems for SNCF. Présenté au INFORMS Fall Meeting, Dallas.

BENDERS, J. F. (1962). Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4:238–252.

BERTOSSI, A. A., CARRARESI, P. et GALLO, G. (1987). On Some Matching Problems Arising in Vehicle Scheduling Models. *Networks*, 17:271–281.

BODIN, L. D., GOLDEN, B. L., SCHUSTER, A. D. et ROMIG, W. (1980). A model for the blocking of trains. *Transportation Research*, 14B:115–120.

BOOLER, J. M. P. (1980). The solution of a railway locomotive scheduling problem. *Journal of the Operational Research Society*, 31:943–948.

BOOLER, J. M. P. (1995). A note on the use of Lagrangean relaxation in railway scheduling. *Journal of the Operational Research Society*, 46:123–127.

BRÄNNLUND, U., LINDBERG, P. O., NÕU, A. et NILSSON, J. E. (1998). Railway timetabling using Lagrangian relaxation. *Transportation Science*, 32:358–369.

BUSSIECK, M. R., KREUZER, P. et ZIMMERMANN, U. T. (1996). Optimal lines for railway systems. *European Journal of Operational Research*, 96:54–63.

BUSSIECK, M. R., WINTER, T. et ZIMMERMANN, U. T. (1997). Discrete optimization in public rail transport. *Mathematical Programming*, 79:415–444.

CAMPBELL, K. C. (1996). *Booking and Revenue Management for Rail Intermodal Services*. Thèse de doctorat, Department of Systems Engineering, University of Pennsylvania, Philadelphia, PA.

CAPRARA, A., FISCHETTI, M., TOTH, P., VIGO, D. et GUIDA, P. L. (1997). Algorithms for railway crew management. *Mathematical Programming*, 79:125–141.

CAREY, M. (1994a). A model and strategy for train pathing with choice of lines, platforms and routes. *Transportation Research*, 28B:333–353.

CAREY, M. (1994b). Extending a train pathing model from one-way to two-way track. *Transportation Research*, 28B:395–400.

CAREY, M. et LOCKWOOD, D. (1995). A model, algorithms and strategy for train pathing. *Journal of the Operational Research Society*, 46:988–1005.

CEDER, A. (1991). A procedure to adjust transit trip departure times through minimizing the maximum headway. *Computers and Operations Research*, 18:417–431.

CHARNES, A. et MILLER, M. H. (1957). A model for the optimal programming of railway freight train movements. *Management Science*, 3:74–92.

CHEN, B. et HARKER, P. T. (1990). Two moments estimation of the delay on single-track rail lines with scheduled traffic. *Transportation Science*, 24:261–275.

CHIH, K. C., HORNUNG, M. A., ROTHENBERG, M. S. et KORNHAUSER, A. L. (1990). Implementation of a real time locomotive distribution system. Dans T. K. S. Murthy, R. E. Rivier, G. F. List et J. Mikolaj, rédacteurs, *Computer Applications*

in Railway Planning and Management, pages 39–49. Computational Mechanics Publications, Southampton, U.K.

CHURCHOD, A. et EMERY, D. (1987). Computer-aided planning for major railway stations. Dans T. K. S. Murthy, F. E. Young, S. Lehmann et W. R. Smith, rédacteurs, *Computers in Railway Installations, Track and Signalling*, pages 3–19. Computational Mechanics Publications, Berlin.

COOPER, L. et LEBLANC, L. J. (1977). Stochastic transportation problems and other network related convex problems. *Naval Research Logistics Quarterly*, 24:327–337.

CORDEAU, J.-F., DESAULNIERS, G., LINGAYA, N., SOUMIS, F. et DESROSIERS, J. (1998a). Simultaneous locomotive and car assignment at VIA Rail Canada. Rapport technique G-98-61, GERAD, École des Hautes Études Commerciales de Montréal, Canada.

CORDEAU, J.-F., SOUMIS, F. et DESROSIERS, J. (1998b). A Benders decomposition approach for the locomotive and car assignment problem. Rapport technique G-98-35, GERAD, École des Hautes Études Commerciales de Montréal, Canada.

CORDEAU, J.-F., TOTH, P. et VIGO, D. (1998c). A survey of optimization models for train routing and scheduling. *Transportation Science*, 32:380–404.

CPLEX (1997). *Using the CPLEX Callable Library 5.0*. ILOG Inc., Incline Village, NV.

CRAINIC, T. G. (1984). A comparison of two methods for tactical planning in rail freight transportation. Dans J. P. Brans, rédacteur, *Operational Research '84*, pages 707–720. Elsevier Science Publishers, Amsterdam.

CRAINIC, T. G., FERLAND, J.-A. et ROUSSEAU, J.-M. (1984). A tactical planning model for rail freight transportation. *Transportation Science*, 18:165–184.

CRAINIC, T. G., FLORIAN, M. et LÉAL, J.-E. (1990a). A model for the strategic planning of national freight transportation by rail. *Transportation Science*, 24:1–24.

CRAINIC, T. G., GENDREAU, M. et DEJAX, P. (1990b). Modelling the container fleet management problem using a stochastic dynamic approach. Dans H. E. Bradley, rédacteur, *Operational Research '90*, pages 473–486. Pergamon Press, New York.

CRAINIC, T. G., GENDREAU, M. et DEJAX, P. (1993). Dynamic and stochastic models for the allocation of empty containers. *Operations Research*, 41:102–126.

CRAINIC, T. G. et ROUSSEAU, J.-M. (1986). Multicommodity, multimode freight transportation: A general modeling and algorithmic framework for the service network design problem. *Transportation Research*, 20B:225–242.

CRANE, R. R., BROWN, F. B. et BLANCHARD, R. O. (1955). An analysis of a railroad classification yard. *Operations Research*, 3:262–271.

DAGANZO, C. F. (1986). Static blocking at railyards: Sorting implications and track requirements. *Transportation Science*, 20:189–199.

DAGANZO, C. F. (1987a). Dynamic blocking for railyards: Part I. Homogeneous traffic. *Transportation Research*, 21B:1–27.

DAGANZO, C. F. (1987b). Dynamic blocking for railyards: Part II. Heterogeneous traffic. *Transportation Research*, 21B:29–40.

DAGANZO, C. F., DOWLING, R. G. et HALL, R. W. (1983). Railroad classification yard throughput: The case of multistage triangular sorting. *Transportation Research*, 17A:95–106.

DANTZIG, G. B. et WOLFE, P. (1960). Decomposition principle for linear programming. *Operations Research*, 8:101–111.

DEJAX, P. J. et CRAINIC, T. G. (1987). A review of empty flows and fleet management models in freight transportation. *Transportation Science*, 21:227–247.

DESAULNIERS, G., DESROSIERS, J., IOACHIM, I., SOLOMON, M.M., SOUMIS, F. et VILLENEUVE, D. (1998). A unified framework for deterministic time constrained vehicle routing and crew scheduling problems. Dans T.G. Crainic et G. Laporte, rédacteurs, *Fleet Management and Logistics*, pages 57–93. Kluwer, Norwell, MA.

DUMAS, Y., DESROSIERS, J. et SOUMIS, F. (1991). The pickup and delivery problem with time windows. *European Journal of Operational Research*, 54:7–22.

FARVOLDEN, J. M. et POWELL, W. B. (1994). Subgradient methods for the service network design problem. *Transportation Science*, 28:256–272.

FISCHETTI, M. et TOTH, P. (1997). A package for locomotive scheduling. Rapport technique DEIS-OR-97-16, University of Bologna, Italy.

FLISBERG, P., HOLMBERG, K., JOBORN, M. et LUNDGREN, J. T. (1996). A model for dimensioning of transportation capacities in a railway network. Rapport technique LiTH-MAT-R-1996-25, Department of Mathematics, Linköping University, Sweden.

FLORIAN, M., BUSHELL, G., FERLAND, J., GUÉRIN, G. et NASTANSKY, L. (1976). The engine scheduling problem in a railway network. *INFOR*, 14:121–138.

FORBES, M. A., HOLT, J. N. et WATTS, A. M. (1991). Exact solution of locomotive scheduling problems. *Journal of the Operational Research Society*, 42:825–831.

FORD, L. R. et FULKERSON, D. R. (1962). *Flows in Networks*. Princeton University Press, Princeton, NJ.

FRANK, M. et WOLFE, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110.

FRANK, O. (1966). Two-way traffic on a single line of railway. *Operations Research*, 14:801–811.

GEOFFRION, A. M. (1974). Lagrangean relaxation for integer programming. *Mathematical Programming Study*, 2:82–113.

GLICKMAN, T. S. et SHERALI, H. D. (1985). Large-scale network distribution of pooled empty freight cars over time, with limited substitution and equitable benefits. *Transportation Research*, 19B:85–94.

GOHRING, K. W. (1971). Application of network flow theory to the distribution of locomotives and cabooses. Présenté au Institute of Management Sciences Southeastern Chapter Winter Symposium.

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

GREENBERG, B. S., LEACHMAN, R. C. et WOLFF, R. W. (1988). Predicting dispatching delays on a low speed, single track railroad. *Transportation Science*, 22:31–38.

HAGHANI, A. E. (1987). Rail freight transportation: A review of recent optimization models for train routing and empty car distribution. *Journal of Advanced Transportation*, 21:147–172.

HAGHANI, A. E. (1989). Formulation and solution of a combined train routing and makeup, and empty car distribution model. *Transportation Research*, 23B:433–452.

HALLOWELL, S. F. et HARKER, P. T. (1996). Predicting on-time line-haul performance in scheduled railroad operations. *Transportation Science*, 30:364–378.

HARKER, P. T. (1989). Use of advanced train control systems in scheduling and operating railroads: Models, algorithms, and applications. *Transportation Research Record*, 1263:101–110.

HARKER, P. T. (1995). Services and technology: Reengineering the railroads. *Interfaces*, 25:72–80.

HARKER, P. T. et HONG, S. (1990). Two moments estimation of the delay on a partially double-track rail line with scheduled traffic. *Journal of the Transportation Research Forum*, 31:38–49.

HARKER, P. T. et HONG, S. (1994). Pricing of track time in railroad operations: An internal market approach. *Transportation Research*, 28B:197–212.

HERREN, H. (1973). The distribution of empty wagons by means of computer: An analytical model of the Swiss Federal Railways. *Rail International*, 4:1005–1010.

HERREN, H. (1977). Computer-controlled empty wagon distribution on the SBB. *Rail International*, 8:25–32.

HIGGINS, A., KOZAN, E. et FERREIRA, L. (1996). Optimal scheduling of trains on a single line track. *Transportation Research*, 30B:147–161.

HIGGINS, A., KOZAN, E. et FERREIRA, L. (1997). Modelling the number and location of sidings on a single line railway. *Computers and Operations Research*, 3:209–220.

HOLMBERG, K., JOBORN, M. et LUNDGREN, J. T. (1998). Improved empty freight car distribution. *Transportation Science*, pages 163–173.

HOLT, J. (1973). Locomotive scheduling by computer "bashpeak". *Rail International*, 4:1053–1058.

HUNTLEY, C. L., BROWN, D. E., SAPPINGTON, D. E. et MARKOWICZ, B. P. (1995). Freight routing and scheduling at CSX Transportation. *Interfaces*, 25(3):58–71.

JOBORN, M. (1995). *Empty Freight Car Distribution at Swedish Railways – Analysis and Optimization Modeling*. Thèse de doctorat, Department of Mathematics, Linköping University, Sweden.

JORDAN, W. C. et TURNQUIST, M. A. (1983). A stochastic, dynamic network model for railroad car distribution. *Transportation Science*, 17:123–145.

JOVANOVIĆ, D. et HARKER, P. T. (1990). A decision support system for train dispatching: An optimization-based methodology. *Journal of the Transportation Research Forum*, 31:25–37.

JOVANOVIĆ, D. et HARKER, P. T. (1991). Tactical scheduling of rail operations: The SCAN I system. *Transportation Science*, 25:46–64.

KEATON, M. H. (1989). Designing optimal railroad operating plans: Lagrangian relaxation and heuristic approaches. *Transportation Research*, 23B:415–431.

KEATON, M. H. (1991). Service-cost tradeoffs for carload freight traffic in the U.S. rail industry. *Transportation Research*, 25A:363–374.

KEATON, M. H. (1992). Designing railroad operating plans: A dual adjustment method for implementing lagrangian relaxation. *Transportation Science*, 26:263–279.

KIKUCHI, S. (1985). Empty freight car dispatching model under freight car pool concept. *Transportation Research*, 19B:169–185.

KOPECKY, MAURICE (1998). Quel avenir pour le rail dans les diverses régions du monde? *Rail International*, Mars:2–8.

KRAAY, D. R. et HARKER, P. T. (1995). Real-time scheduling of freight railroads. *Transportation Research*, 29B:213–229.

KRAAY, D. R., HARKER, P. T. et CHEN, B. (1991). Optimal pacing of trains in freight railroads: Model formulation and solution. *Operations Research*, 39:82–99.

KRAFT, E. R. (1987). A branch and bound procedure for optimal train dispatching. *Journal of the Transportation Research Forum*, 28:263–276.

KRAFT, E. R. (1988). Analytical models for rail line capacity analysis. *Journal of the Transportation Research Forum*, 29:153–162.

KRAFT, E. R. (1998). *A Reservations-Based Railway Network Operations Management System*. Thèse de doctorat, Department of Systems Engineering, University of Pennsylvania, Philadelphia, PA.

KROON, L. G., ROMEIJN, H. E. et ZWANEVELD, P. J. (1997). Routing trains through railway stations: Complexity issues. *European Journal of Operational Research*, 98:485–498.

LEBLANC, L. J. (1976). Global solutions for a nonconvex nonconcave rail network model. *Management Science*, 23:131–139.

MAGNANTI, T. L. et WONG, R. T. (1981). Accelerating Benders decomposition: Algorithmic enhancement and model selection criteria. *Operations Research*, 29:464–484.

MAHEY, P. (1986). Méthodes de décomposition et décentralisation en programmation linéaire. *RAIRO Recherche opérationnelle*, 20:287–306.

MANSFIELD, E. et WEIN, H. H. (1958). A model for the location of a railroad classification yard. *Management Science*, 4:292–313.

MARÍN, A. et SALMERÓN, J. (1996a). Tactical design of rail freight networks. Part I: Exact and heuristic methods. *European Journal of Operational Research*, 90:26–44.

MARÍN, A. et SALMERÓN, J. (1996b). Tactical design of rail freight networks. Part II: Local search methods with statistical analysis. *European Journal of Operational Research*, 94:43–53.

MARKOWICZ, B. P. et TURNQUIST, M. A. (1990). Applying the lp solution to the daily distribution of freight cars. Présenté au TIMS/ORSA National Meeting, Las Vegas.

MARTINELLI, D. R. et TENG, H. (1996). Optimization of railway operations using neural networks. *Transportation Research*, 4C:33–49.

MARTLAND, C. D. (1982). PMAKE analysis: Predicting rail yard time distributions using probabilistic train connection standards. *Transportation Science*, 16:476–506.

MARTLAND, C. D. et SUSSMAN, J. M. (1995). A perspective on rail systems modelling: What works and what doesn't. Rapport technique 95-1, MIT Center for Transportation Studies, Cambridge, MA.

MCDANIEL, D. et DEVINE, M. (1977). A modified Benders' partitioning algorithm for mixed integer programming. *Management Science*, 24:312–379.

MCGAUGHEY, R. S., GOHRING, K. W. et MCBRAYER, R. N. (1973). Planning locomotive and caboose distribution. *Rail International*, 4:1213–1218.

MENDIRATTA, V. B. et TURNQUIST, M. A. (1982). A model for management of empty freight cars. *Transportation Research Record*, 838:50–55.

MISRA, S. C. (1972). Linear programming of empty wagon disposition. *Rail International*, 3:151–158.

MORIN, M.-H. (1993). *Modélisation et décompositon des problèmes de transbordement dynamiques: application à la répartition des wagons Fret/SNCF*. Thèse de doctorat, Université J. Fourier, Grenoble, France.

MORLOK, E. K. et PETERSON, R. B. (1970). Final report on a development of a geographic transportation network generation and evaluation model. *Journal of the Transportation Research Forum*, 11:71–105.

NACHTIGALL, K. (1995). Time depending shortest-path problems with applications to railway networks. *European Journal of Operational Research*, 83:154–166.

NACHTIGALL, K. (1996). Periodic network optimization with different arc frequencies. *Discrete Applied Mathematics*, 69:1–17.

NACHTIGALL, K. et VOGET, S. (1996). A genetic algorithm approach to periodic railway synchronization. *Computers and Operations Research*, 23:453–463.

NACHTIGALL, K. et VOGET, S. (1997). Minimizing waiting times in integrated fixed interval timetables by upgrading railway tracks. *European Journal of Operational Research*, 103:610–627.

NEMHAUSER, G. L. (1969). Scheduling local and express trains. *Transportation Science*, 3:164–175.

NEWTON, H. N. (1996). *Network Design Under Budget Constraints with Application to the Railroad Blocking Problem*. Thèse de doctorat, Auburn University, Auburn, AL.

NÕU, A. (1997). Railway timetabling - Lagrangian heuristics. Rapport technique TRITA/MAT-97-OS12, Royal Institute of Technology, Stockholm, Sweden.

NÕU, A., DESROSIERS, J. et SOUMIS, F. (1997). Weekly locomotive scheduling at swedish state railways. Rapport technique G-97-35, GERAD, École des Hautes Études Commerciales de Montréal, Canada.

NOZICK, L. K. et MORLOK, E. K. (1997). A model for medium-term operations planning in an intermodal rail-truck service. *Transportation Research*, 31A:91–107.

ODIJK, M. A. (1996). A constraint generation algorithm for the construction of periodic railway timetables. *Transportation Research*, 30B:455–464.

ÖZEKICI, S. (1987). Average waiting times in queues with scheduled batch services. *Transportation Science*, 21:55–61.

ÖZEKICI, S. et ŞENGÖR, S. (1994). On a rail transportation model with scheduled services. *Transportation Science*, 28:246–255.

PETERSEN, E. R. (1974). Over-the-road transit time for a single track railway. *Transportation Science*, 8:65–74.

PETERSEN, E. R. (1975). Interference delays on a partially double-tracked railway with intermediate signalling. *Journal of the Transportation Research Forum*, 16:55–62.

PETERSEN, E. R. (1977a). Railyard modeling: Part I. Prediction of put-through time. *Transportation Science*, 11:37–49.

PETERSEN, E. R. (1977b). Railyard modeling: Part II. The effect of yard facilities on congestion. *Transportation Science*, 11:50–59.

PETERSEN, E. R. et TAYLOR, A. J. (1982). A structured model for rail line simulation and optimization. *Transportation Science*, 16:192–205.

PETERSEN, E. R. et TAYLOR, A. J. (1983). Line block prevention in rail line dispatch and simulation models. *INFOR*, 21:46–51.

PETERSEN, E. R. et TAYLOR, A. J. (1987). Design of a single-track rail line for high speed trains. *Transportation Research*, 21A:47–57.

PETERSEN, E. R., TAYLOR, A. J. et MARTLAND, C. D. (1986). An introduction to computer aided train dispatch. *Journal of Advanced Transportation*, 20:63–72.

PHILIP, C. E. et SUSSMAN, J. M. (1977). Inventory model of the railroad empty-car distribution process. *Transportation Research Record*, 656:52–60.

POWELL, W. B. (1986). A stochastic model of the dynamic vehicle allocation problem. *Transportation Science*, 20:117–129.

POWELL, W. B. (1987). An operational planning model for the dynamic vehicle allocation problem with uncertain demands. *Transportation Research*, 21B:217–232.

POWELL, W. B. (1995). Making the solution fit the problem. *INFORMS Rail Applications Newsletter (Spring)*, pages 6–9.

POWELL, W. B., JAILLET, P. et ODONI, A. (1995). Stochastic and dynamic networks and routing. Dans M. O. Ball, T. L. Magnanti, C. L. Monma et G. L. Nemhauser, rédacteurs, *Network Routing*, pages 141–295. North-Holland, Amsterdam.

POWELL, W. B. et SHEFFI, Y. (1989). Design and implementation of an interactive optimization system for network design in the motor carrier industry. *Operations Research*, 37:12–29.

RAMANI, K. V. (1981). An information system for allocating coach stock on Indian Railways. *Interfaces*, 11(3):44–51.

RAMANI, K. V. et MANDAL, B. K. (1992). Operational planning of passenger trains in Indian Railways. *Interfaces*, 22(5):39–51.

RATCLIFFE, L. L., VINOD, B. et SPARROW, F. T. (1984). Optimal prepositioning of empty freight cars. *Simulation*, 42:269–275.

RIVIER, R. E. et TZIEROPOULOS, P. (1984). Interactive graphics models for railway operational planning. Dans M. Florian, rédacteur, *The Practice of Transportation Planning*, pages 245–259. Elsevier Science Publishers, Amsterdam.

RIVIER, R. E. et TZIEROPOULOS, P. (1987). Computer-aided planning of railway networks, lines and stations. Dans T. K. S. Murthy, L. S. Lawrence et R. E. Rivier, rédacteurs, *Computers in Railway Management*, pages 3–16. Computational Mechanics Publications, Berlin.

SALZBORN, F. J. M. (1969). Timetables for a suburban rail transit system. *Transportation Science*, 3:297–316.

SALZBORN, F. J. M. (1970). The minimum fleetsize for a suburban railway system. *Transportation Science*, 4:383–402.

SAUDER, R. L. et WESTERMAN, W. M. (1983). Computer aided train dispatching: Decision support through optimization. *Interfaces*, 13(6):24–37.

SHERALI, H. D. et TUNCBILEK, C. H. (1997). Static and dynamic time-space strategic models and algorithms for multilevel rail-car fleet management. *Management Science*, 43:235–250.

SIDDIQEE, M. W. (1972). Investigation of sorting and train formation schemes for a railroad hump yard. Dans G. F. Newell, rédacteur, *Traffic Flow and Transportation*, pages 377–387. Elsevier, New York.

SMITH, M. E. (1990). Keeping trains on schedule: On-line planning systems for the advanced railroad electronics system (ARES). *Journal of the Transportation Research Forum*, 31:17–24.

SMITH, S. et SHEFFI, Y. (1988). Locomotive scheduling under uncertain demand. *Transportation Research Record*, 1251:45–53.

SPIECKERMANN, S. et VOSS, S. (1995). A case study in empty railcar distribution. *European Journal of Operational Research*, 87:586–598.

SUZUKI, S. (1973). A method of planning yard pass trains on a general network. Dans M. Ross, rédacteur, *Operational Research '72*, pages 353–361. North Holland, Amsterdam.

SZPIGEL, B. (1973). Optimal train scheduling on a single track railway. Dans M. Ross, rédacteur, *Operational Research '72*, pages 343–352. North-Holland, Amsterdam.

THOMET, M. A. (1971). A user-oriented freight railroad operating policy. *IEEE Transactions on Systems, Man, and Cybernetics*, 1:349–356.

TURNQUIST, M. A. et DASKIN, M. S. (1982). Queuing models of classification and connection delay in railyards. *Transportation Science*, 16:207–230.

VAN DYKE, C. D. (1986). The automated blocking model: A practical approach to freight railroad blocking plan development. *Transportation Research Forum*, 27:116–121.

VAN DYKE, C. D. (1988). Dynamic management of railroad blocking plans. *Transportation Research Forum*, 29:149–152.

WHITE, W. W. et BOMBERAULT, A. M. (1969). A network algorithm for empty freight car allocation. *IBM Systems Journal*, 8:147–169.

WRIGHT, M. B. (1989). Applying stochastic algorithms to a locomotive scheduling problem. *Journal of the Operational Research Society*, 40:187–192.

YAGAR, S., SACCOMANNO, F. F. et SHI, Q. (1983). An efficient sequencing model for humping in a rail yard. *Transportation Research*, 17A:251–262.

YU, G. (1998), rédacteur. *Operations Research in the Airline Industry*. Kluwer Academic Publishers, Boston.

ZIARATI, K. (1997). *Affectation des locomotives aux trains*. Thèse de doctorat, École Polytechnique de Montréal, Canada.

ZIARATI, K., SOUMIS, F. et DESROSIERS, J. (1997a). Locomotive assignment using train delays. Rapport technique G-97-27, GERAD, École des Hautes Études Commerciales de Montréal, Canada.

ZIARATI, K., SOUMIS, F., DESROSIERS, J., GÉLINAS, S. et SAINTONGE, A. (1997b). Locomotive assignment with heterogeneous consists at CN North America. *European Journal of Operational Research*, 97:281–292.

ZIARATI, K., SOUMIS, F., DESROSIERS, J. et SOLOMON, M. M. (1998). A branch-first, cut-second approach for locomotive assignment. Rapport technique G-98-11, GERAD, École des Hautes Études Commerciales de Montréal, Canada.

ZWANEVELD, P. J., KROON, L. G., ROMEIJN, H. E., SALOMON, M., DAUZÈRE-PÉRÈS, S., HOESEL, S. P. M. VAN et AMBERGEN, H. W. (1996). Routing trains through railway stations: Model formulation and algorithms. *Transportation Science*, 30:181–194.