Efficient Circulation of Railway Rolling Stock

Arianna Alfieri

Politecnico di Torino, Dipt. Sistemi di Produzione ed Economia

Corso Duca degli Abruzzi 24, IT–10129, Torino, Italy

Rutger Groot

ORTEC Consultants

P.O. Box 490, NL–2800 AL Gouda, The Netherlands

Leo Kroon

NS Reizigers, Department of Logistics

P.O. Box 2025, NL–3500 HA Utrecht, The Netherlands

Rotterdam School of Management, Erasmus University Rotterdam

P.O. Box 1738, NL–3000 DR Rotterdam, The Netherlands

Alexander Schrijver

Centrum voor Wiskunde en Informatica

P.O. Box 94079, NL-1090 GB Amsterdam, The Netherlands

Abstract

Railway rolling stock is one of the most significant cost sources for operators of passenger trains. The *efficient* circulation of rolling stock is therefore one of the main objectives pursued in practice. This paper focuses on the determination of appropriate *numbers* of train units of different types together with their efficient *circulation* on a single line. In order to utilize the train units on this line in an efficient way, they are *coupled* to or *uncoupled* from the trains in certain stations according to the passengers' seat demand in peak hours and off-peak hours. Since coupling and uncoupling train units has to respect specific rules related to the shunting possibilities in the stations, it is important to take into account the *order* of the train units in the trains. This aspect strongly increases the complexity of the rolling stock circulation problem. This paper presents a solution approach based on an Integer Programming model. The approach is applied to a real life case study based on the timetable of NS Reizigers, the main Dutch operator of passenger trains.

1 Introduction

The efficient circulation of railway rolling stock is an important problem for operators of passenger trains, since the rolling stock is one of their most significant cost sources. The involved costs are mainly due to acquisition, power supply, and maintenance of the rolling stock. Since these costs are usually substantial, it has to be decided carefully how much rolling stock is necessary per scheduled train in order to provide a service to the passengers of a certain quality: an efficient circulation of rolling stock requires per trip a match between the provided rolling stock capacity and the passengers' demand for transportation.

In order to achieve this objective, the compositions of the trains may have to be changed during the operations by coupling or uncoupling rolling stock to or from the trains. For example, rolling stock may be uncoupled from the trains after the morning peak hours and it may be coupled again before the afternoon peak hours. In these *shunting* processes, which are usually carried out in the short time interval between the arrival of a train at a station and its subsequent departure, several practical rules related to the feasibility of the transitions of the train compositions are to be taken into account.

This paper focuses on the determination of appropriate *numbers* of *train units* of different types together with their efficient *circulation* on a single line of NS Reizigers, the main Dutch operator of passenger trains. Thereby the positions of the train units in the trains are taken into account, since these determine the feasibility of the transitions from one train composition into another.

This paper is organized as follows. A detailed description of the problem is given in Section 2. Section 3 gives an overview of earlier research in the area of rolling stock circulation. Section 4 describes two models that are used to solve our rolling stock circulation problem. The first model neglects the detailed compositions of the trains, and the second one takes these into account. Section 5 presents our solution approach based on a combination of a multi-commodity flow model and finding shortest paths in so-called *transition graphs*. The results of our computational experiments are presented in Section 6. Finally, conclusions are drawn in Section 7.

2 Problem description

2.1 Rolling stock: train units

NS Reizigers has a large variety of rolling stock available for its passenger transportation process. The major part of the rolling stock consists of *train units*. Each train unit consists of a certain number of carriages that cannot be split from each other during the daily operations. A main difference between a train unit and a single *carriage* is that each train unit has its own engines. As a consequence, a train unit can move individually in both directions without a locomotive. This is in contrast with a single carriage, which needs a locomotive for any movement. Figure 1 shows an example of a single deck train unit with 3 carriages, and a single carriage.

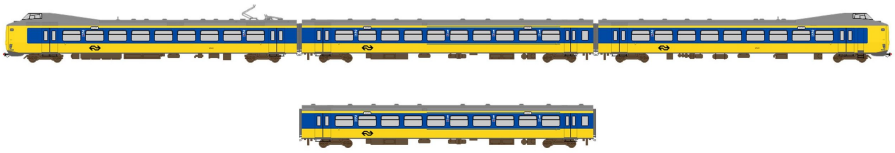

Figure 1: A single-deck train unit with 3 carriages and a single carriage

Single-deck train units can be subdivided into train units with 3 carriages and train units with 4 carriages. In this paper, we focus on the circulation of such single-deck train units. Each train unit has fixed capacities of first class and second class seats. These capacities are more or less linear in the number of carriages per train unit. Single deck train units with 3 or 4 carriages may be combined with each other into one longer train. The latter implies that for each type not only the number of train units in each train is to be determined, but also their order in the train. This is explained later in this section.

An ordered sequence of train units in a train is called a *composition*. For example, if "3" and "4" denote single-deck train units with 3 and 4 carriages, respectively, then "344" and "434" indicate different compositions. The notation here is such that both compositions have a train unit of type "4" in front (in

other words, the trains move from left to right).

Coupling train units onto a train or uncoupling train units from a train has to be carried out in the short period (typically just a few minutes) between a train's arrival at a station and its subsequent departure from that station. In order to minimize the time required for the involved shunting movements, train units are usually coupled onto the *front* end of an incoming train, and uncoupled from the *rear* end of a leaving train. Only if the difference between the departure time and the arrival time of a train is sufficiently large, more complicated shunting movements may be carried out.

From a capacity point of view, the earlier mentioned compositions "344" and "434" are identical. However, these compositions have different transition possibilities: the train unit of type "3" can be uncoupled easily from the composition "344", whereas the latter is not the case for the composition "434". In case of the composition "344", the train unit of type "3" can be uncoupled from the train and the remaining part of the train (with composition "44") can proceed without being disturbed by the uncoupled train unit. In case of the composition "434", uncoupling the train unit of type "3" would require several shunting operations which usually requires too much time. In this case, uncoupling the rear train unit of type "4" would be easy, but if one really wants to uncouple the train unit of type "3", then uncoupling the train unit of type "4" may conflict with the train's passenger demand later on.

The fact that train units can be subdivided into different types complicates both the planning process and the daily operations significantly. However, the *gain* is that per train a better match between the expected number of passengers (demand) and the provided number of seats (supply) can be realized. For example, with train units with a length of 4 carriages, only trains with 4, 8, or 12 carriages can be composed. However, a combination of train units with 3 and 4 carriages may give rise to trains with 3, 4, 6, 7, 8, 9, 10, 11 and 12 carriages. Note that each train should not be longer than the shortest platform of the stations along the train's route, and certainly not longer than 12 carriages.

2.2 Further details of the problem

The planning of the rolling stock circulation usually starts after the timetable has been completed, since the timetable is required as input for the rolling stock circulation planning. In this paper, the timetable is assumed to be cyclic.

A line is a direct connection between two end points that is served with a certain frequency (e.g. once per hour or twice per hour). Each line is served by a fixed number of *trains*. During the day, each train runs up-and-down between the end points of the line. The number of trains on a line is determined by the line's circulation time, including the return times at the end points, and the line's frequency. This number equals the number of trains that can be observed on a picture of the line that is taken from above at any moment during the day.

The timetable is given in the form of a set of trips. A trip represents the movement of a train between two stations where the composition of the train can be changed. Each trip has a start time, an end time, an origin station and a destination station. For each trip, also the corresponding train is known. Conversely, each train corresponds to an ordered sequence of trips that are carried out by this train. In particular, for each trip also the next trip of the corresponding train is known.

Other input consists of the forecasts of the required capacities per trip. For collecting this input, the passengers are counted continuously by the conductors. The translation of these counts into the minimally required first and second class capacities per trip is based on a statistical procedure, which falls outside the scope of this paper. Since in the operations the number of passengers on a trip has a *stochastic* character and since NS Reizigers does not use a *seat reservation system*, it is impossible to *guarantee* a seat for all passengers, especially during the peak hours. However, outside the peak hours, the rolling stock capacity is usually sufficient to provide all passengers with a seat.

Fixed rolling stock costs are related to acquisition and depreciation. Variable rolling stock costs are related to power supply but also to maintenance of the rolling stock: after a certain number of kilometers, each train unit is directed to a maintenance station for a preventive check-up and possibly for a repair.

Note that, in the Netherlands, routing train units to a maintenance facility for a check-up is carried out only in the *operations*. Therefore, maintenance requirements can be ignored in our rolling stock circulation *planning* problem. Other variable rolling stock costs are related to the crew: each train requires at least one driver and one conductor, but if the length of a train exceeds a certain threshold, then a second (or third) conductor is required for safety reasons.

Given the timetable for a single line and the corresponding demand forecasts for a single day, the problem is to find appropriate numbers of train units of the different types, together with such a circulation of these train units that (*i*) all forecasted passengers can be seated, (*ii*) all practical constraints concerning the maximum lengths of the trains and the transitions of the train compositions are respected, and (*iii*) the relevant objective consisting of fixed and/or variable costs is minimized. Note that in some trains the allocated capacity may be larger than the capacity required by the demand. This is due to the demand on subsequent trips of the train, or for relocating train units.

2.3 Case study: the line 3000

The model and solution approach that are described later in this paper are illustrated with a case study based on the line 3000, one of the Intercity lines of NS Reizigers. This line provides twice per hour an Intercity connection from Den Helder (Hdr) to Nijmegen (Nm) and vice versa (see Figure 2 and [14]).

The timetable of the line 3000 is cyclic with a cycle length of 30 minutes. Only in the early morning and in the late evening, there are some exceptions from the cyclic timetable. For example, in the early morning there are some trains starting in Alkmaar (Amr), Amsterdam (Asd), Utrecht (Ut), and Arnhem (Ah). Similar exceptions exist in the late evening. In principle, the compositions of the trains can be changed only in Alkmaar and Nijmegen. In the late evening, there are again some exceptions from this rule, since then also some train units may be uncoupled in Den Helder, Amsterdam, Utrecht, or Arnhem.

The line 3000 contains 12 trains. This is caused by the fact that the circulation time on the line between Den Helder and Nijmegen and vice versa is

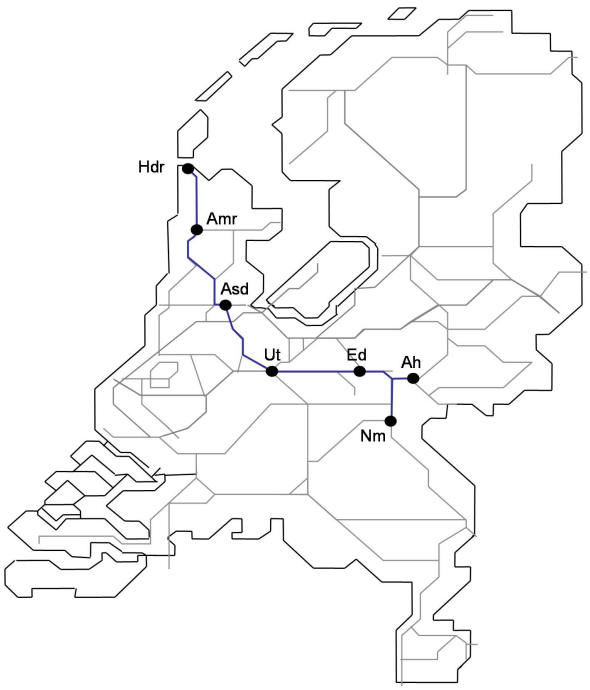

Figure 2: The line 3000 Den Helder (Hdr) - Nijmegen (Nm)

six hours and that there are two trips per hour in each direction. Hence every twelfth departure from, say, Nijmegen, can be covered by the same train.

Figure 3 shows a time-space diagram for part of the trips of this line. The numbers at the top of the figure indicate the time axis. The grey diagonal lines indicate the trips and the adjacent numbers are the corresponding train numbers. In order to keep Figure 3 as simple as possible, the compositions of the trains have been indicated for only a subset of the trains. Train units of type "3" are represented by dashed lines. Train units of type "4" are represented by solid lines. The train unit at the front of a train is represented by the rightmost line of a composition.

For example, the train on trip 3017 is operated with composition "43" where the train unit of type "3" is the front unit. This train arrives in Nijmegen at 7:11, where an additional train unit of type "3" is coupled. Nijmegen is a terminal

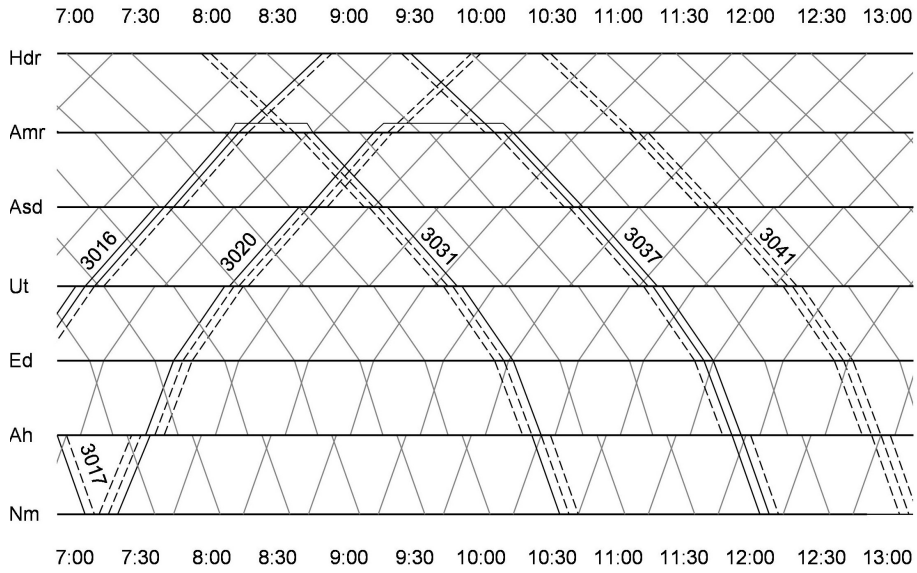

Figure 3: Part of rolling stock circulation for the line 3000

station, where the front end and the rear end of the train are exchanged. The coupled train unit becomes the front end of the incoming train, which is the rear end of the outgoing train. The train leaves at 7:20 from Nijmegen on trip 3024 with composition "334", and arrives in Arhnhem at 7:33.

In Arnhem, the physical composition of a train is usually not changed, but the front end and the rear end of a train are exchanged, as is shown in Figure 3. The latter is caused by the fact that each train leaves the station Arnhem from the same side as where it entered it. The train leaves Arnhem at 7:38 with composition "433".

The train arrives in Alkmaar at 9:19. Here the train unit of type "4" at the rear end of the train is uncoupled and remains in Alkmaar. The remaining train continues at 9:21 to Den Helder with composition "33". As a consequence, the train unit of type "4" should be uncoupled within 2 minutes, which prohibits a complex shunting process. The remaining train arrives at 9:56 in Den Helder. At 10:33 it returns to Nijmegen on trip 3041, again with composition "33". This process continues until the end of the day.

Note that in Den Helder there is also a departure already at 10:03 (see Figure 3). Thus the train might have returned already at that time. However, because of the robustness of the circulation, all trains have a standstill at Den Helder of at least 30 minutes. Obviously, the price that has to be paid for this robustness is the additional train that is required for the rolling stock circulation.

Train units that have been uncoupled from a certain train can be coupled onto another train later on. For example, Figure 3 shows how the train unit that was uncoupled from the rear end of the train on trip 3020 at 9:19 in Alkmaar is coupled onto the front end of another train on trip 3037 at 10:12. Note that in Alkmaar it is impossible to couple a train unit onto a northbound train, nor is it possible to uncouple a train unit from a southbound train.

3 Related literature

In the literature several rolling stock circulation problems have been described, which shows the importance of this kind of problems. However, most of the literature focuses on the circulation of *locomotive hauled carriages* and not on the circulation of *train units*. As far as we know, the following three papers are the only published papers that deal with the circulation of train units.

Schrijver [16] considers the problem of minimizing the number of train units of different types for an hourly train line in the Netherlands, given that the passengers' seat demand must be satisfied. The only restriction on the transition between two compositions on two consecutive trips is that the required train units must be available at the right time and station. (Un)coupling restrictions related to the feasibility of shunting movements are ignored.

Ben-Khedher et al. [5] study the problem of allocating *identical* train units to the French High Speed Trains. Their rolling stock allocation system is based on a capacity adjustment model that is linked to the seat reservation system. This system aims at maximizing the expected profit. Since here the train units are identical, shunting restrictions are less important in this paper.

Abbink et al. [1] present a model to allocate different rolling stock families

and types to different train lines. They present an Integer Programming model, that minimizes the seat shortages during the morning peak hours by allocating rolling stock families and types with different capacities to all the trains running simultaneously at 8:00 am, the busiest moment of the day. Their approach is applied to several scenarios of NS Reizigers that differ in the numbers of rolling stock families and types that can be allocated to a line.

Papers dealing with the efficient circulation of *locomotive hauled carriages*, possibly in combination with the required locomotives, are the following. Brucker et al. [6] study the problem of finding a circulation of railway carriages through a rail network, given a timetable. Since the required train compositions have been specified a priori, this paper focuses on finding appropriate repositioning trips of carriages from one station to another. Their solution approach is based on local search techniques like simulated annealing.

Also Van Montfort [13] focuses on the efficient circulation of railway carriages. He studies the assignment of carriages to trains, given a cyclic timetable and a *core standard structure* for train compositions on a combination of lines in the Netherlands. Van Montfort uses an Integer Programming model that is improved by the application of several types of valid inequalities.

Cordeau et al. [7] present a Benders decomposition approach for the locomotive and car assignment problem. Their approach is based on the concept of a *train consists*, i.e. a group of compatible units of rolling stock (locomotive(s), first and second class carriages) that travel along some part of a rail network. Computational experiments show that optimal solutions can be found in short computation times by applying column generation techniques. In a subsequent paper, Cordeau et al. [8] extend their model with various real-life constraints, for example dealing with maintenance of the rolling stock.

Lingaya et al. [11] study the problem of assigning carriages to trains at VIA Rail in Canada. They present a model to adapt a master plan to additional information concerning the expected numbers of passengers. They allow for coupling and uncoupling of carriages at various locations in the network and explicitly take the order of the carriages in the trains into account. Several

other real-life constraints, such as maintenance requirements, are considered as well. The solution approach is based on a Dantzig-Wolfe reformulation solved by column generation. Next, a branch-and-bound procedure is applied heuristically to obtain good integer solutions.

The problem described in the current paper is different from the problems in the above mentioned papers. First, the current paper deals with train units instead of locomotive hauled carriages. Shunting restrictions for train units are different from those for locomotive hauled carriages. Second, the current paper deals with train units of different types, which implies that the detailed orders of the train units in the trains are relevant. Further details of our problem and solution approach can be found in Groot [9].

4 Model formulation

This section describes the models that are used for solving the rolling stock circulation problem. In Section 4.1 we describe the definitions and notations that are used in this paper. In Section 4.2 we describe the model that neglects the details of the compositions of the trains. Thereafter, in Section 4.3 we present the model that takes into account the compositions of the trains. In Section 4.4 we describe the computational complexity of several variants of the rolling stock circulation problem.

4.1 Definitions and notation

We assume to have a single line connecting two end stations with a given frequency. The timetable on this line is cyclic. The number of trains that are running on this line depends on the line's circulation time and frequency. All trains are assumed to be operated by train units that are possibly of different types. The latter implies that the problem is a complex variant of a multi-commodity flow problem. In particular, the problem could be called an *ordered* multi-commodity flow problem, since the order of the train units in the trains is as important as their number.

In this paper, we use the following notation. First, we have a set T of trips, a set S of stations, and a set J of types of train units. Each trip $t \in T$ is represented by an origin station O_t, a destination station D_t, a start time S_t, and an end time E_t. For each train τ, the set T_τ denotes the ordered set of trips that is operated by train τ. In particular, for each trip t also the next trip $n(t)$ that is carried out by the same train is known. The parameter L_t represents the length of trip t. Furthermore, the parameter $d_{c,t}$ $(c = 1, 2)$ denotes the expected number of passengers in class c on trip t. The set T_t^a is the subset of trips t' arriving in station O_t before the departure of trip t from O_t. That is, $T_t^a = \{\ t' \in T \mid D_{t'} = O_t, E_{t'} < S_t\ \}$. Similarly, T_t^d is the subset of trips t' departing from station O_t before the departure of trip t. Thus $T_t^d = \{\ t' \in T \mid O_{t'} = O_t, S_{t'} < S_t\ \}$. The length of the shortest platform along trip t is denoted by P_t. The length and the number of carriages of each train unit of type j are denoted by l_j and N_j, respectively. Finally, $C_{j,c}$ represents the capacity in class c $(c = 1, 2)$ of each train unit of type j. Finally, F_j and V_j denote the fixed and variable costs of each train unit of type j, respectively.

4.2 Model 1: neglecting the compositions

If the order of the train units in the trains is neglected, then the rolling stock circulation problem can be represented by a multi-commodity flow model with several additional constraints. This model is represented by a *flow graph* (also called a *time-space graph*, see Figure 4), whose nodes correspond to events (an arrival or departure of a train in a station) and whose arcs are connections between events. An arc is a trip arc if the two connected nodes belong to different stations (represented by dashed lines in Figure 4), and it is a station arc if the two events are consecutive events of the same station (represented by solid lines). A station arc represents a connection between an arrival (or departure) in a station and the next arrival (or departure) in the same station.

In each node of the flow graph, we have to guarantee the flow balance. In Figure 4, for example, the number of train units in Alkmaar (Amr) before the arrival of the train on trip 3031 plus the number of train units arriving in the

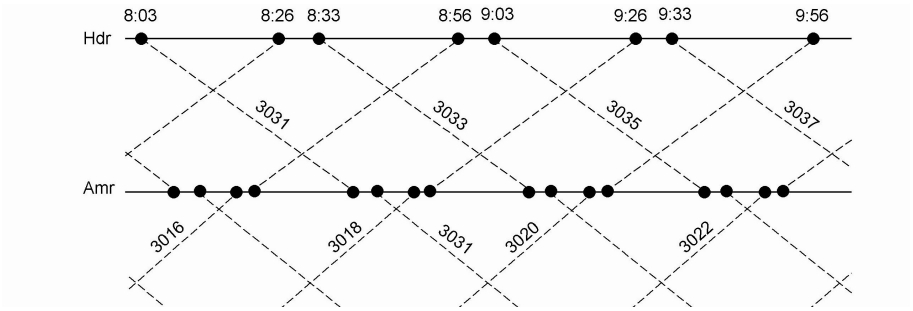

Figure 4: Part of the flow graph for the line 3000

train on trip 3031 equals the number of train units leaving from Alkmaar in the train on trip 3031 plus the number of train units in Alkmaar after the departure of the train on trip 3031. This relation holds separately for each train unit type.

The model is expressed in terms of the decision variables $x_{t,j}$ indicating the number of train units of type j that is allocated to trip t. Additional decision variables are the variables $y^0_{s,j}$ denoting the number of train units of type j that is stored in station s during the night, and y^{tot}_j denoting the total number of available train units of type j.

Now Model 1 (neglecting the compositions) can be stated as follows:

$$\min F(x,y) \qquad \text{subject to}$$

$$y^{tot}_j = \sum_s y^0_{s,j} \qquad \forall j \in J \tag{1}$$

$$x_{t,j} \leq y^0_{s,j} + \sum_{t' \in T^a_t} x_{t',j} - \sum_{t' \in T^d_t} x_{t',j} \qquad \forall j \in J; t \in T, s = O_t \tag{2}$$

$$\sum_j l_j x_{t,j} \leq P_t \qquad \forall t \in T \tag{3}$$

$$\sum_j C_{j,c} x_{t,j} \geq d_{c,t} \qquad \forall t \in T; c = 1,2 \tag{4}$$

$$x_{n(t),j} \leq x_{t,j} \qquad \forall t : O_t = \text{Ah}, D_t = \text{Amr}; j \in J \tag{5}$$

$$x_{n(t),j} \geq x_{t,j} \qquad \forall t : O_t = \text{Hdr}, D_t = \text{Amr}; j \in J \tag{6}$$

$$x_{t,j} \in \{\ 0, \ldots, M_{t,j}\ \} \qquad \forall t \in T, j \in J \tag{7}$$

$$y_j^{tot}, y_{s,j}^0 \in \mathcal{Z}^+ \qquad \forall s \in S, j \in J \tag{8}$$

In this model, $F(x, y)$ represents one of the following objective functions:

1. PB1: minimize the fixed costs of the train units ($\min \sum_j F_j y_j^{tot}$).
2. PB2: minimize the variable costs of the train units ($\min \sum_t L_t(\sum_j V_j x_{t,j})$.
3. PB3: minimize the variable costs of the train units taking into account an upper bound on the number of carriages ($\min \sum_t L_t(\sum_j V_j x_{t,j})$ subject to $\sum_j N_j y_j^{tot} \leq UB$).

The total number of required train units of each type equals the total number of train units $y_{s,j}^0$ that stay in the various stations during the night, as is represented by constraints (1). Constraints (2) describe that the number of train units of type j that is allocated to trip t should not exceed the number of train units of this type that is available in station O_t just before the start time of trip t. The latter equals the number of such train units by the start of the day plus the number of train units that have arrived in this station until this time instant, minus the number of train units that have departed from there until then. Note that it is not difficult to take into account a certain minimum re-assignment time between the uncoupling of a train unit from a train and the subsequent coupling of this train unit onto another train. However, this re-assignment time has been omitted here. On each trip t, a train should not be longer than the length of the shortest platform P_t along the trip. The latter is guaranteed by constraints (3). Constraints (4) are the demand satisfaction constraints for first class and second class seat demands on each trip. Constraints (5) and (6) describe that in Alkmaar train units cannot be coupled onto a northbound train nor uncoupled from a southbound train. Finally, constraints (7) and (8) specify the integer character of the decision variables. Here $M_{t,j}$ is an appropriate upper bound for the variable $x_{t,j}$.

Valid Inequalities

This subsection describes how the constraints on the maximum train length (3) and demand satisfaction (4) can be made more tight by adding certain *valid inequalities*. For example, the model may contain the following constraints (9) and (10) for a certain trip t:

$$3x_{t,3} + 4x_{t,4} \quad \leq 12 \tag{9}$$

$$166x_{t,3} + 224x_{t,4} \quad \geq 510 \tag{10}$$

These constraints represent the fact that a train should have a length of at most 12 carriages, and that the second class seat demand is to be satisfied. Here the number 510 equals the required number of second class seats on trip t, and the numbers 166 and 224 represent the second class capacities of train units with 3 and 4 carriages, respectively. Due to the integrality of the variables $x_{t,3}$ and $x_{t,4}$, constraint (10) can be sharpened as follows:

$$x_{t,3} + 2x_{t,4} \quad \geq 4 \tag{11}$$

$$x_{t,3} + x_{t,4} \quad \geq 3 \tag{12}$$

The above example is shown in Figure 5. Here the grey and dark areas correspond to the continuous feasible region for the train composition on trip t obtained by considering constraints (9) and (10). The black dots represent the feasible combinations of the two types of train units (Unit 3 and Unit 4, respectively). In this example, the feasible combinations are (4,0), (2,1), (1,2), and (0,3). The dark area represents the convex hull of the feasible region.

We certainly do not claim that valid inequalities such as (11) and (12) give a complete description of the convex hull of the integer feasible region of the complete problem. Nevertheless, the improved local description of the convex hull turned out to give an improved performance of our solution approach in many cases (see Section 6).

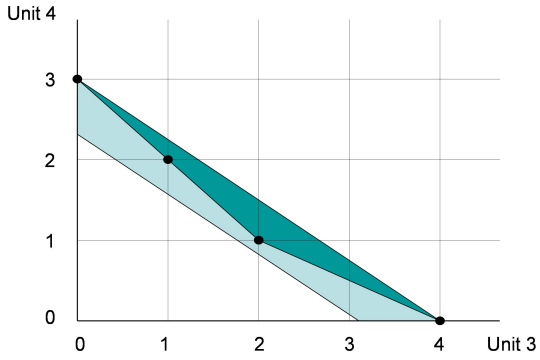

Figure 5: Reduced feasible region

4.3 Model 2: taking into account the compositions

In order to also take into account the compositions of the trains, the following definitions are used. First, the set of all feasible compositions is denoted by K, and the set of compositions that are feasible for trip t is denoted by K_t. This set is determined based on the required first and second class capacities and on the maximum train length for trip t. The parameter $g_{k,j}$ denotes the number of train units of type j in composition k. The feasible transitions from one composition of a train to another are described in the sets $A_{t,k}$ for each trip t and composition k. That is, the set $A_{t,k}$ contains the compositions that are feasible on trip $n(t)$ if the train has composition k on trip t.

For each train, the feasible compositions per trip and the feasible transitions from one composition to another are represented in a so-called *transition graph* (see Figure 6). The set of nodes of the transition graph of train τ is the set $\bigcup_{t \in T_\tau} \{ (t,k) \mid k \in K_t \}$. Here the union is taken over all trips t that are carried out by train τ. The set of arcs of this graph is the set $\bigcup_{t \in T_\tau} \bigcup_{k \in K_t} A_{t,k}$. Here the union is again taken over all trips t that are carried out by train τ and over all feasible compositions $k \in K_t$. Furthermore, each transition graph has a source node, connected to all nodes of the first trip of the corresponding train, and a sink node, connected to all nodes of the last trip of this train.

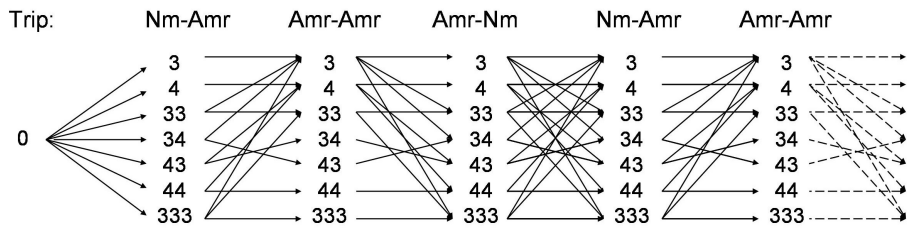

Figure 6: Part of the transition graph for one train of the line 3000

An example of (part of) a transition graph is shown in Figure 6. In order to keep the figure as clear as possible, we assume that the maximum train length is 9 carriages on all trips. The involved train starts with a trip from Nijmegen to Alkmaar. On this first trip, each of the 7 compositions with at most 9 carriages can be chosen. The feasible compositions on the second trip depend on the composition that was chosen on the first trip. The feasible transitions in Alkmaar represent the fact that the order of the train units in the train was changed in Arnhem (e.g. composition 34 changes to 43) and the fact that in Alkmaar train units can only be *uncoupled* from a northbound train. Note that the second trip runs from Alkmaar to Alkmaar since it is assumed that in Den Helder no train units are (un)coupled. Therefore, the two train movements from Alkmaar to Den Helder and vice versa have been aggregated into a single trip from Alkmaar to Alkmaar. The feasible transitions between the second and the third trip represent the fact that the order of the train units in the train was changed in Den Helder and the fact that in Alkmaar train units can only be *coupled* onto a southbound train. The transition graph shows that in Nijmegen train units can be coupled or uncoupled (but not both). The rest of the transition graph can be explained similarly.

In each transition graph, we have to find a path from its source node to its sink node, which altogether minimize the objective function $F(x, y)$. The objective functions that are used for Model 2 are the same as for Model 1.

Since there is a transition graph for each train, this seems to give a decomposition of the problem across the trains at first sight. However, this is not the

case, since the paths through the transition graphs interact with each other via the inventories of train units in the stations. Train units uncoupled from a train can be coupled onto another one later on, as was shown in Figure 3.

In the model, binary decision variables are associated with the *nodes* in the transition graphs. That is, a decision variable $a_{t,k}$ assumes the value 1 if and only if composition k is selected for trip t. Now Model 2 (taking into account the compositions) can be represented as follows:

$$\min F(x,y) \qquad \text{subject to (1), (2), (5)-(8), and}$$

$$\sum_{k \in K_t} a_{t,k} = 1 \qquad \forall t \in T \tag{13}$$

$$a_{t,k} \leq \sum_{k' \in A_{t,k}} a_{n(t),k'} \qquad \forall t \in T, k \in K_t \tag{14}$$

$$x_{t,j} = \sum_{k \in K_t} g_{k,j} a_{t,k} \qquad \forall t \in T, j \in J \tag{15}$$

$$a_{t,k} \in \{0,1\} \qquad \forall t \in T, \forall k \in K_t \tag{16}$$

Note that constraints (3) and (4) of Model 1 do not appear in Model 2, since these are handled now by constraints (13). These constraints represent the fact that for each trip exactly one *appropriate* composition is to be selected. Constraints (13) guarantee that the composition that is selected for trip t is compatible with the composition that is selected for the next trip $n(t)$ of the same train. Next, constraints (15) link the flow graph and the transition graphs with each other: for each type, the number of train units on a trip follows directly from the composition that is used on that trip. In fact, these constraints make the flow variables $x_{t,j}$ superfluous, since in all occurrences of these variables they can be removed by substituting (15). If this substitution is applied consequently, then (15) itself becomes superfluous as well. Finally, constraints (16) specify the binary character of the composition variables $a_{t,k}$.

4.4 Computational complexity

The rolling stock circulation problem (both neglecting and taking into account the compositions) is NP-hard. This result is proved here for the case where the total fixed costs are to be minimized. For other objectives a similar result holds. The NP-hardness is proved by a reduction from the problem Numerical 3-Dimensional Matching (N3DM), which is defined as follows:

N3DM

Instance: $3n$ positive integers a_i, b_i and c_i and a positive integer d satisfying $\frac{d}{4} < a_i, b_i, c_i < \frac{d}{2}$ and $\sum_{i=1}^{n}(a_i + b_i + c_i) = nd$.

Question: Do there exist permutations σ and τ of the set $\{1, \ldots, n\}$ such that $a_{\sigma(i)} + b_{\tau(i)} + c_i = d$ for $i = 1, \ldots, n$?

Theorem 1. *Finding a feasible rolling stock circulation with minimum total fixed costs for the train units is NP-hard.*

Note that the proof of Theorem 1 applies to the rolling stock circulation problem both if the compositions are neglected and if these are taken into account. Hence both variants of the rolling stock circulation problem are NP-hard.

Proof. Let I be an instance of N3DM as described above. Then the following instance I' of the rolling stock circulation problem is constructed.

The instance I' contains $3n$ trains and $6n$ trips. Each train carries out a trip from station A to station B in the time interval (0,1) and a trip back from station B to station A in the time interval (2,3).

For $i = 1, \ldots, n$, the first trip of train i has a second class demand of d and the second trip of this train has a second class demand of $2d - c_i$. The first class demand of the latter trips equals 1. For *all* other trips the first class demand equals 0. For $i = 1, \ldots, n$, the first trip of train $i + n$ has a second class demand of $d + a_i$ and the second trip of this train has a second class demand of d. For $i = 1, \ldots, n$, the first trip of train $i + 2n$ has a second class demand of $d + b_i$ and the second trip of this train has a second class demand of d.

There are $2n+1$ different train unit types. For $i = 1, \ldots, n$, train unit type i has first class capacity 1 and second class capacity a_i. For $i = 1, \ldots, n$, train unit type $i+n$ has first class capacity 0 and second class capacity b_i. The last train unit type has first class capacity 0 and second class capacity d. For each train unit type, the fixed costs are equal to the second class capacity.

Now we claim the following: there exists a feasible solution for I if and only if there exists a feasible solution for I' with total fixed costs $4nd - \sum_{i=1}^{n} c_i$.

First, suppose there exists a feasible solution for I' with total fixed costs $4nd - \sum_i c_i$. Note that both in the time interval (0,1) and in the time interval (2,3), the total second class demand equals the total fixed costs. Due to the fixed cost structure of the train units, it follows that in both time intervals there is an exact match of the second class demand and the provided second class capacity. Thus in the time interval (0,1), each trip with second class demand $d + a_i$ is covered by a train consisting of a train unit with second class capacity d and a train unit with second class capacity a_i. Similarly, each trip with second class demand $d + b_i$ is covered by a train consisting of a train unit with second class capacity d and a train unit with second class capacity b_i. Finally, each trip with second class demand d is covered by a train consisting of a train unit with second class capacity d.

In the time interval (2,3), the same train units are assigned to the trips running in this interval. Each trip with first class demand 1 (and second class demand $2d - c_i$) is covered by at least one train unit with first class demand 1 (and second class capacity in $\{\ a_i \mid i = 1, ..., n\ \}$). Hence, each of these trips is covered by *exactly* 1 of these train units. The $2n$ trips with second class demand d are covered by $2n$ train units with second class capacity d. The remaining train units are assigned to the trips with second class demand $2d - c_i$ in such a way that on each trip there is an exact match of capacity and demand. Exactly one train unit with second class capacity in $\{\ b_i \mid i = 1, ..., n\ \}$ and one train unit with second class capacity d is assigned to each of these trips.

Hence, if $\sigma(i)$ and $\tau(i)$ are defined as the train unit types that are assigned to the trip with second class demand $2d - c_i$, then it follows that $a_{\sigma(i)} + b_{\tau(i)} + d =$

$2d - c_i$ for $i = 1, \ldots, n$. Thus σ and τ are the requested permutations, and it follows that I has a feasible solution.

Conversely, if I has a feasible solution, then the construction can be reversed to find a feasible solution for I' with total fixed costs $4nd - \sum_i c_i$. Since N3DM is NP-complete, and the reduction is polynomial, it follows that the rolling stock circulation problem is NP-hard. □

The proof of Theorem 1 is based on the fact that the number of trains and the number of train unit types are not fixed a priori. However, if these numbers are fixed a priori, then the problem can be solved in polynomial time, as is shown in Theorem 2. Note that this result is hardly relevant from a computational point of view, due to the huge size of the involved network.

Theorem 2. *If the number of trains, the number of stations, the number of train unit types, and the maximum train length are fixed, then a feasible rolling stock circulation with minimum total fixed costs can be found in an amount of time that is polynomial in the number of trips.*

Proof. This theorem is proved by solving the problem as a shortest path problem in a network with numbers of nodes and arcs that are polynomial in the number of trips. Here we only give a rough sketch of the proof. For similar proofs, see Arkin and Silverberg [3] or Kroon et al. [10].

Without loss of generality, all arrivals and departures take place at different time instants. If an event is defined as the arrival or departure of a train at a station, then each node in the network represents a feasible state of the system between two consecutive events. These nodes represent both the compositions of all trains that are running in the corresponding time interval and the inventories at the stations. Each arc in the network represents a feasible transition between two successive nodes in the network. If the number of trains, the number of stations, the number of train unit types, and the maximum train length are fixed, then it is not difficult to see that the numbers of nodes and arcs in the network are polynomial in the number of trips. A shortest path in this network corresponds with a rolling stock circulation with minimum total fixed costs. □

5 Solution Approach

For single commodity flow problems many algorithms are available (Ahuja et al. [2]), but for multi-commodity flow problems this is not the case. Furthermore, the available algorithms usually assume that variables may have fractional values (Ahuja et al. [2], McBride [12]). Solution approaches for integer multi-commodity flow problems were studied by Barnhart et al. [4], based on branch-and-price techniques. However, their paper does not deal with *ordered* multi-commodity flow problems. The solution approach that we propose for solving the latter problems can be summarized as follows.

1. Solve Model 1. This provides a (usually strong) lower bound for the solution of Model 2.

2. Reduce the transition graphs by successively applying:
   - Node elimination, and
   - Disconnection elimination.

3. If each reduced transition graph still contains a path from source to sink, then solve Model 2 on the reduced transition graphs. Otherwise go back to Step 2 with relaxed elimination conditions.

4. If Model 2 has a feasible solution, then STOP. Otherwise go back to Step 2 with relaxed elimination conditions.

The lower bound obtained in Step 1 by solving Model 1 turns out to be quite strong usually. Therefore, after Model 1 has been solved, it is checked in Steps 2 to 4 if there exists a solution for Model 2 with the *same value* of the objective function L. The latter is done by first reducing the numbers of nodes and arcs of the transition graphs by the application of *node elimination* and *disconnection elimination*. Next, these elimination methods are described in more detail.

Node elimination

The node elimination process consists of solving a series of subproblems. Each of these subproblems corresponds to solving the Linear Programming relaxation

of Model 1, where (i) the constraint $F(x, y) \leq L$ has been added, and (ii) a certain variable $x_{t,j}$ has been fixed to one of its a priori feasible values V. These a priori feasible values for a variable $x_{t,j}$ are determined based on the passenger demand for trip t and on the maximum train length for trip t.

If a Linear Programming instance in which a variable $x_{t,j}$ has been fixed to the value V does *not* have a feasible solution, then obviously the corresponding Integer Programming instance is infeasible as well. Therefore, all nodes and arcs in the transition graph corresponding to trip t and each composition $k \in K_t$ with $g_{k,j} = V$ can be eliminated from this transition graph. This node elimination process iterates over all variables $x_{t,j}$ and over all a priori feasible values V.

For example, if a Linear Programming instance in which a certain variable $x_{t,4}$ has been set to the value $V = 2$ turns out to be infeasible, then all nodes and arcs corresponding to this trip and a composition with 2 train units of type "4" (that is, "44", "344", "434", and "443") are eliminated from the involved transition graph.

Disconnection elimination

The node elimination process may have the effect that some nodes in a transition graph become disconnected, since their neighbors have been removed. Then obviously no path in this transition graph from the source node to the sink node may pass through these nodes. We then perform *disconnection elimination* by eliminating all nodes that have become disconnected in this way. During the disconnection elimination, other nodes may become disconnected. Thus this process may be carried out until no more disconnected nodes are eliminated.

Once the disconnection elimination process terminates, another round of node elimination could be started. This process could be carried out until no more nodes are eliminated at all. However, each iteration has a certain cost in terms of the required CPU time. Since this cost turns out to be high, in particular in comparison with the number of additionally eliminated nodes, Step 2 is carried out only once per major iteration.

Iteration and termination

If, at the end of Step 2, there is still at least one connected node (feasible composition) for each trip, then the remaining transition graphs are used as input for Model 2 with the additional restriction $F(x, y) \leq L$.

Otherwise, if there exists at least one trip whose nodes have all been eliminated, then obviously a feasible solution for Model 2 with objective function value L does not exist. In this case, we go back to Step 2, where the elimination process is relaxed slightly. That is, in Step 2 we set $L := L + \varepsilon$, where the actual value of ε depends on the objective function $F(x, y)$. Note that by relaxing the elimination conditions, the number of eliminated nodes and arcs decreases, which increases the probability of obtaining a feasible solution.

If Model 2 with the additional restriction $F(x, y) \leq L$ has a feasible solution for the remaining transition graphs, then this solution is an optimal solution for Model 2. Otherwise, if a feasible solution for Model 2 with objective function value L does not exist, then we also go back to Step 2, where the relaxed elimination process is carried out after setting $L := L + \varepsilon$.

6 Computational Results

6.1 Experiments

This section presents the computational results that were obtained by applying the algorithm described in Section 5 to the case of the Intercity line 3000 (see Section 2.3). The data that were used for our experiments correspond to a standard Tuesday from the timetable of NS Reizigers. For the solution of the Integer Programming models and their Linear Programming relaxations we used CPLEX (version 6.6) on a RISC 6000 workstation.

In our experiments, we consider single-deck train units for which two types are available: train units with 3 carriages (type "3") and train units with 4 carriages (type "4"). On all trips, the maximum train length is 12 carriages. In order to evaluate the influence of the valid inequalities of the capacity constraints instead of the capacity constraints themselves, we experimented with

both model formulations.

We further solved the instances for a complete day, but also for just the trips that start before 10:30 (called the Morning Peak Hours, or MPH). In fact, the appropriate number of train units is mainly determined by the passengers' seat demand during the morning peak hours. Note that this is usually the busiest period of the day, since the peak in the morning is usually higher than the peak in the afternoon, which lasts longer.

6.2 Results

In Table 1, the dimensions of the case study are reported in terms of the numbers of columns, rows and non-zeros in the constraint matrix. The labels *complete* and *MPH* indicate the problems with all the trips in the timetable and the problem reduced to only trips leaving during the Morning Peak Hours.

	Model 1, complete	Model 1, MPH
columns	1544	494
rows	1755	544
non-zeros	4132	1290
	Model 2, complete	Model 2, MPH
columns	14142	3716
rows	6380	1894
non-zeros	66017	16182

Table 1: Dimensions of the line 3000 problem

Obviously, in Model 1 the number of decision variables is less than the number of constraints, while in Model 2 the situation is the opposite. For Model 2, the numbers of decision variables and constraints representing the compositions depend on the number of nodes eliminated from the transition graphs, and thus also on the objective function used. The values in Table 1 are average values observed in our computational experiments.

Tables 2 to 4 show the results of our computational experiments. The row *CPU* indicates the CPU-time used for the complete solution process. Obviously, the additional valid inequalities such as (11) and (12) often speed up the so-

lution process significantly, due to the improvement of the value of the Linear Programming lower bound. Only in some cases (e.g. in the case of the objective PB1) the total CPU time is higher if these valid inequalities are taken into account. The latter is due to the larger number of major iterations during the solution process of Model 2, as is represented in the row *iter*.

The row *objval* denotes the obtained objective function value. Here the fixed cost per train unit are assumed to be equal to the number of carriages of the train unit. The obtained numbers of train units of type "3" and "4" are represented in the rows y_3^{tot} and y_4^{tot}, respectively.

The rows *tot-nodes* and *elim-nodes* contain the total number of nodes in the transition graphs and the number of nodes eliminated in the elimination phase, respectively. The difference between these two numbers gives the total number of nodes in the reduced transition graphs.

	complete, no-cut	complete, cut	MPH, no-cut	MPH, cut
CPU (sec)	812.1	2132.1	39.3	93.1
objval	112	112	112	112
y_3^{tot}	20	20	20	20
y_4^{tot}	13	13	13	13
iter	1	4	1	3
tot-nodes	8604	8604	2498	2498
elim-nodes	515	1950	298	1597

Table 2: Results for the objective function PB1

	complete, no-cut	complete, cut	MPH, no-cut	MPH, cut
CPU (sec)	4122.8	478.5	36.7	29.7
objval	1735	1735	617	617
y_3^{tot}	23	23	23	23
y_4^{tot}	14	14	11	11
iter	1	1	1	1
tot-nodes	8604	8604	2498	2498
elim-nodes	416	416	144	144

Table 3: Results for the objective function PB2

	complete, no-cut	complete, cut	MPH, no-cut	MPH, cut
CPU (sec)	4266.9	483.6	38.1	30.3
objval	1769	1769	621	621
y_3^{tot}	20	20	20	24
y_4^{tot}	13	13	13	10
iter	1	1	1	1
tot-nodes	8604	8604	2498	2498
elim-nodes	449	2766	177	1597

Table 4: Results for the objective function PB3

6.3 Comments

For all objective functions, most of the CPU time is spent on the solution of Model 1. The presence of different solutions with the same objective function value (*degeneracy*), combined with a weak continuous relaxation, makes it difficult to certify optimality. Often, the first integer solution found is the optimum, but then it may take quite some time to close the integrality gap.

When solving Model 2, the solution of each Linear Programming instance during the node elimination phase takes usually very little time, but the large number of subproblems that has to be solved here makes the node elimination process rather time consuming.

After the elimination phase, Model 2 does not seem to be very difficult to solve, at least not more difficult than Model 1. This is mainly due to the excellent lower bound that is provided by the solution of Model 1. Due to this excellent lower bound, quite often only one major iteration of the solution process is needed.

In all cases, the results obtained by solving the MPH problems are rather similar to the results obtained by solving the *complete* problems in terms of the numbers of train units of the different types. This was to be expected, since the numbers of train units needed to satisfy a certain seat demand are mainly determined by the demand for seats during the morning peak hours. Anyway, the fixed costs of a certain MPH problem provide a strong lower bound for the fixed costs of the corresponding complete problem.

7 Conclusions and further research

In this paper we studied the problem of determining optimal numbers of train units together with their efficient circulation on a single line, thereby taking into account the fact that trains can be composed of train units of different types. The latter implies that not only the *number* of train units of the different types in the trains, but also their *order* in the trains is to be modeled.

We proved that the rolling stock circulation problem is NP-hard in its most general form. However, if the number of trains, the number of stations, the number of train unit types, and the maximum train length are fixed, then an optimal rolling stock circulation can be found in an amount of time that is polynomial in the number of trips.

We described a model and an algorithm for solving the rolling stock circulation problem, The model uses the concept of *transition graphs*. The algorithm starts with reducing the sizes of the transition graphs as much as possible by applying node elimination and disconnection elimination. Then the remaining problem is solved by CPLEX. We applied the algorithm to the Intercity line 3000 of NS Reizigers. Based on our results, it can be concluded that the proposed solution approach is a powerful scheme, which succeeds to find optimal solutions within an acceptable amount of time.

In our further research, we will study cases that have an even more complex structure than the one presented in this paper. Such cases may involve several train lines at the same time, more than one family of train units (for example, both single-deck and double-deck train units), more than two types of train units, or trains splitting and combining underway.

In addition to the *strategic* problem of determining appropriate numbers of train units to be operated on a certain set of lines, we will also focus on the *operational* problem of optimally circulating a given number of train units along these lines. In that case, the problem is to find a balance between the conflicting objectives of minimizing (*i*) the shortages of seats (service), (*ii*) the number of train unit or carriage kilometers (efficiency), and (*iii*) the number of shunting movements (robustness). The latter is relevant since shunting movements are

potential sources of disruptions of the railway process. Therefore, avoiding these shunting movements may be beneficial for the punctuality.

In our further research, we will also focus on alternative algorithmic approaches for solving the rolling stock circulation problem. In particular, we will experiment with a solution approach based on column generation. Here the column generation mechanism generates appropriate paths through the transition graphs based on shortest path algorithms. This column generation mechanism takes into account *dual cost* information obtained from the master problem. The latter handles the coordination between the paths in the different transition graphs, mainly by taking into account the inventories of train units in the different stations. Preliminary results of the application of this approach can be found in Peeters and Kroon [15].

References

[1] E.J.W. Abbink, B.W.V. van den Berg, L.G. Kroon, and M. Salomon, "Allocation of Train Units to Passenger Trains". *Transportation Science*, 38(1) (2004): 33-41.

[2] R.K. Ahuja, T.L. Magnanti, J.B. Orlin, "Network Flow - Theory, Algorithms, and Applications", Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[3] E.M. Arkin and E.L. Silverberg. "Scheduling Jobs with Fixed Start and End Times". *Discrete Applied Mathematics*, 18 (1987): 1-8.

[4] C. Barnhart, A. Hane, P.H. Vance, "Using Branch-and-Price to Solve Origin-Destination Integer Multi-Commodity Flow Problems", *Operations Research*, 32 (1998): 208-220.

[5] N. Ben-Khedher, J. Kintanar, C. Queille, W. Stripling, "Schedule Optimization at SNCF: From Conception to Day of Departure", *Interfaces*, 28 (1998): 6-23.

[6] P. Brucker, J. Hurink, T. Rolfes, "Routing of Railway Carriages: A Case Study", *Osnabrücker Schriften zur Mathematik*, Reihe P, Heft 205 (1998).

[7] J.F. Cordeau, F. Soumis, J. Desrosiers. "A Benders Decomposition Approach for the Locomotive and Car Assignment Problem". *Transportation Science*, 34 (2000): 133-149.

[8] J.F. Cordeau, F. Soumis, J. Desrosiers, "Simultaneous Assignment of Locomotives and Cars to Passenger Trains", *Operations Research*, 49 (2001): 531-548.

[9] R. Groot, "Minimum Circulation of Railway Stock: an Integer Programming Algorithm", Master's Thesis, University of Amsterdam, 1996.

[10] L.G. Kroon, H.E. Romeijn, and P.J. Zwaneveld. "Routing trains through railway stations: complexity issues". *European Journal of Operational Research*, 98 (1999): 485-498.

[11] N. Lingaya, J.F. Cordeau, G. Desaulniers, J. Desrosiers, F. Soumis, "Operational Car Assignment at VIA Rail Canada", *Transportation Research B*, 36 (2002): 755-778.

[12] R.D. McBride, "Advances in Solving the Multi-Commodity-Flow Problem", *Interfaces*, 28 (1998): 32-41.

[13] J. van Monfort, "Optimizing Railway Carriage Circulation with Integer Linear Programming", Master's Thesis, University of Amsterdam, 1997.

[14] NS homepage: www.ns.nl

[15] M. Peeters and L.G. Kroon, "Circulation of Railway Rolling Stock: a Branch-and-Price Approach", ERIM Report Series 2003-055-LIS, Erasmus University Rotterdam (2003),

[16] A. Schrijver, "Minimum Circulation of Railway Stock", *CWI Quarterly*, 6 (1993): 205-217.