

Integration of the Load Matching and Routing Problem with Equipment Balancing for Small Package Carriers

Amy Mainville Cohn, Sarah Root, and Alex Wang
University of Michigan

Doug Mohr
United Parcel Service

July 17, 2005

Abstract

Small package delivery is a multi-billion dollar industry with complex planning decisions required to efficiently utilize costly resources and meet tight time requirements. The planning process is typically decomposed into sequential sub-problems to establish tractability. This decomposition can greatly degrade solution quality. In this paper, we therefore consider the integration of two closely related key sub-problems: *load matching and routing* and *equipment balancing*. First, we identify critical challenges faced in trying to solve these problems. Then, we present a novel modeling approach to address these challenges. Finally, we conclude with computational results from UPS, the world's largest package delivery company, demonstrating an improvement of approximately 5% over their existing methods for solving this pair of problems.

1 Introduction

The transportation industry plays a critical role in today's global economy, supporting nearly all aspects of everyday life. Each year, a staggering volume of goods is moved through the US freight transportation network. In 1998, more than 15 billion tons of goods worth over \$9 trillion were moved ([14]). Transportation-related goods and services accounted for approximately 11% of the US GDP in 2000, with only housing, health-care, and food accounting for a greater share ([25]). Small package carriers are an important sector of this industry, with the dominant players each transporting millions of packages per day, using tens of thousands of motor vehicles and hundreds of jet aircraft ([28], [13]).

Small package carriers, as well as many other players in the transportation industry, face an enormously complex planning process. They must allocate

multiple competing resources (drivers, vehicles, package sorting facilities, etc.) that are tightly constrained. The tight time windows associated with expedited package carriers serve to further increase this challenge. Additionally, cost functions in the transportation industry are often non-linear (for example, due to fixed charges or volume-based discounts), greatly impacting the tractability of optimization-based approaches to planning ([12], [5], [17], [19]).

As a result, the planning process is often decomposed into several sequential sub-problems, with the output from one sub-problem used as input to the next. Although such decompositions can greatly enhance tractability, they typically result in sub-optimality and can even lead to infeasibility for the overall problem.

Algorithmic advances and improvements in computing power have increased the size and complexity of large-scale transportation planning problems that can be solved. In many cases, it is therefore possible to more fully integrate existing planning processes. Recent research in integrated transportation network planning includes [9], [10], [11], [18], [20], [21], [23], and [27].

Given that full integration of the planning process is still beyond the scope of tractability for small package carriers, it is important when attempting to partially integrate the planning process to choose sub-problems that are both important as individual problems (that is, have a significant impact on cost and have a wide range of feasible solutions with substantially varying objective values) and that are tightly coupled with each other (that is, the decisions made in one problem substantially impact the feasible region of another). In this paper, we identify two such problems – *load matching and routing* and *equipment balancing* – and justify the importance of their integration. We introduce these problems in Section 2 and provide a high-level overview of the overall network planning process for small package carriers. In Section 3, we present a traditional modeling approach to the integrated load matching/routing and equipment balancing problem (*LMREB*), based on a *multi-commodity flow (MCF)* formulation, and identify challenges that prevent the tractability of this approach. We introduce an alternative modeling approach in Section 4, explain how this approach addresses the challenges posed by the *MCF* approach, and suggest ways to ensure the tractability of this alternative model. Section 5 presents our computational experiments, which were designed to evaluate our model’s tractability as well as its solution quality. We conclude in Section 6 with a summary and suggested areas for further research.

2 Network Planning for Small Package Carriers

Small package carriers transport millions of packages daily around the world ([13], [28]), with tight time windows that can span as little as 24 hours or less. The number of possible origin/destination (*O/D*) pairs, even within just the United States, is close to 2 billion, assuming 5 digit zip codes as the level of granularity. Most *O/D* pairs will not generate enough volume on a given day to justify the movement of a dedicated truck. Instead, packages are moved through intermediate *sorting facilities* (also known as *consolidation centers* or

hubs), where packages are grouped by common destinations to fill *trailers* more efficiently. A given package may move through several such facilities. For example, a package from Worcester, Massachusetts to Carlsbad, California might first be grouped with all other packages originating in Worcester and other surrounding towns, then loaded on a trailer to Boston. In Boston, this *load* might be *broken* and the package sorted and placed in a new load, traveling to San Diego. In San Diego, this load might again be broken and sorted, with the package being loaded onto a final trailer filled with Carlsbad-destined packages.

This type of consolidation operation allows for greater cost efficiencies. However, it also increases the travel time of packages, not only because of the increased circuitry in driving distance, but also because of the time incurred in sorting and handling packages, including time spent waiting for the arrival of other packages to comprise the loads. Thus, in order to meet the tight time windows offered, key trade-offs must be made between the use of intermediate handling to save cost and the more direct movement of packages to save time. Determining what path (i.e. series of intermediate handling points) each *O/D* pair should follow is often called *load planning* or *package routing* and is typically done at the tactical level, with all packages for a given *O/D* pair following the same load plan in order to improve operational simplicity.

Once the load plan has been established, the packages can be built into *loads* – groups of packages that will move together in a single trailer from one facility to another, such as Worcester to Boston in our earlier example. The loads are defined not only by origin and destination, but also by a time window. The earliest departure of the load corresponds to the latest available time of all the packages in that load. The latest arrival time of the load corresponds to the earliest due date of all its packages. When converting a group of packages into loads, one of the key determinations to be made is the number and type of trailers to be used. The two main trailer types are approximately 28' and 44' in length, each having different limits on both the weight and volume of packages they can hold. Although minimizing the number of trailers used is important, other factors are relevant in this decision-making process as well, such as how these loads will interact with one another when driver schedules are constructed. This process of converting packages to loads is done both for planning purposes at the tactical level, using average volume estimates, and for operational planning on the daily level.

After the loads have been built and assigned to trailer types, it must then be determined how these loads will move through the network. A load (trailer) from one facility to another will not necessarily be pulled by a single *tractor* directly from the load's origin to its destination. It may instead be combined with a second trailer with compatible time constraints moving between the same origin and destination. This is because, on many highways, a tractor can pull two trailers simultaneously. The cost of such a move is typically much lower than the cost of two separate tractors each pulling a single trailer. In fact, this per-mile savings may be significant enough to justify driving excess mileage to pickup and drop off a load with a different origin and/or destination so that at least part of the travel costs can be shared by two loads. We refer to this

as *load matching and routing*. Clearly, this cost structure provides significant opportunities but also results in a complex combinatorial challenge.

Given the construction of loads, it may also be necessary to *balance the network*. Package levels vary throughout the network, and different trailer types are assigned with different levels of utilization. Therefore, certain facilities may have more loads outgoing and others more loads incoming. This can result in equipment imbalances, with trailers accumulating at some locations and deficits occurring at others. Thus, *empty trailers* must often be moved throughout the network to regain balance. Because the physical trailers are a less heavily constrained resource, this *equipment balancing* does not have the same tight time considerations that are required of loaded moves.

Once the movement of loaded and empty trailers has been determined, drivers must be assigned to cover these movements. Because drivers have their own complex constraints, including government-mandated safety regulations and labor-negotiated rules or company policies for driver satisfaction, it may not be possible for a single driver to move an entire segment non-stop. Instead, the movement may need to be broken up into smaller segments. For example, a load moving from Boston to California might be transported by several different drivers, each stopping at an intermediate point and handing off the load to the next driver, possibly swapping loads before turning back towards home. Thus, planning decisions must be made at two different levels in this regard, how to schedule the drivers and how to break down the load route segments into even smaller pieces, so that each driver can pull a sequence of loads without violating driver restrictions and each load can be assigned to a sequence of drivers without violating time constraints.

Figure 1 provides a simplistic view of the hierarchy of network decisions. Here, we see a package originating at A and destined for B . It moves on load 1 from its origin to hub $H1$, then on load 2 from $H1$ to $H2$, then on load 3 from $H2$ to destination B . The next level provides detail about load 1. It is initially combined with load 4 as a double configuration and is driven by a single tractor to intermediate point C , where load 4 is dropped off. Load 1 then continues as a single trailer to its destination, $H1$. In the third level, we see greater detail on the movement of loads 1 and 4 from A to C . Here, we see that the tractor stops at two intermediate points, with three different drivers covering the three individual segments.

Within the planning process for small package carriers, a large number of tightly inter-related resources must be accounted for. These include packages, loads, trailers, tractors, package sorting facilities, and package handlers. Non-linear costs, tight time constraints, complex labor rules, and other factors make these individually challenging problems. As a whole, the global planning process is intractable as a single optimization problem. Nonetheless, improvements in modeling and algorithmic techniques, as well as advancements in computing capabilities, make partial integration of this planning process a possibility.

In this paper, we consider the integration of the *load matching and routing problem* with *equipment balancing*. We have chosen these problems for two reasons. First, both of these are important problems individually, with combi-

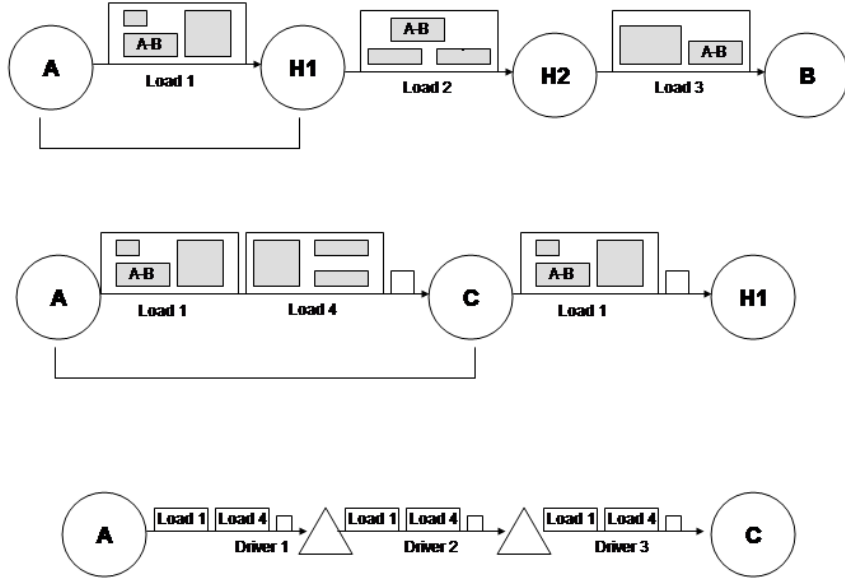


Figure 1: Network Structure

natorial complexity leading to a large set of feasible solutions and a significant variation in the quality (i.e. cost) of these solutions. Second, because loaded and empty trailers can be combined behind a single tractor, opportunities exist for substantial cost savings by making these decisions simultaneously.

In the following sections, we present two different models for the integrated problem (*LMREB*). For the sake of notational simplicity, and without loss of generality, the models are presented under the assumption that only 28' trailers have been used in building the loads which are an input to this problem. Furthermore, it is assumed that all arcs in the network permit the movement of two trailers behind a single tractor and that no other tractor configurations are permitted.

3 A MCF-Based Modeling Approach

Given the underlying structure of *LMREB*, it is logical to model it as a network flow problem ([1]); such formulations often have properties that can be exploited for very efficient solution techniques. More specifically, *LMREB* can be viewed as a variation of the *multi-commodity flow problem (MCF)* ([2], [6], [7], [15], [16]), in which multiple commodities must move through a network to satisfy

node-based supplies and demands without violating arc capacities. In this case, a commodity is defined as a set of loads that share a common origin, destination, earliest pickup time, and latest delivery time. The set of empty trailers moving through the network to regain balance also comprises a commodity. There are no arc capacities.

There are two key differences between *LMREB* and a pure *MCF* problem. First, we have to ensure that the commodities satisfy their time windows. Second, the arc flows are required to be integer, and the objective function is non-linear with respect to these flow values. This is because an even-valued arc flow x_A on arc A implies $\frac{x_A}{2}$ double trailers while an odd value implies $\lfloor \frac{x_A}{2} \rfloor$ double trailers plus one single trailer, with the cost of a single trailer being larger than one-half the cost of a double trailer.

It is possible to modify the traditional *MCF* formulation to address these issues; we do so in two ways. First, instead of a network in which nodes represent physical locations (in this case, carrier facilities), we use a *time-space network* (for other examples, see [8], [24], [27]), in which nodes represent both locations and points in time. An arc from node $\{f_1, t_1\}$ to node $\{f_2, t_2\}$ indicates the movement of trailers leaving facility f_1 at time t_1 and arriving at facility f_2 at time t_2 . An arc from node $\{f_1, t_1\}$ to node $\{f_1, t_2\}$ indicates trailers which remain at facility f_1 between times t_1 and t_2 . A supply is created at node $\{o, e\}$ for a commodity that originates at facility o with earliest pickup time e . A demand is created at node $\{d, a\}$ for a commodity that is destined for facility d with latest arrival time a . All other nodes are transshipment nodes for these commodities. The use of a time-space network, as seen in the formulation below, allows us to capture the time constraints implicitly within the *MCF* formulation. This comes, however, at the cost of a significant increase in the size of the network, in terms of the numbers of both nodes and arcs.

The second issue to be addressed is the non-linear cost function. We linearize the objective by defining two integer variables for each arc, which represent the numbers of double and single trailer configurations moving over this arc. We can then enforce the relationship between these variables and the total flow of trailers (i.e. commodities) over an arc, with only the configurations and not the commodity flows appearing in the objective function.

Before presenting this modified *MCF* formulation, we introduce the following notation:

Variables

- s_{ij} = number of single trailers flowing on arc (i, j) (recall that node i is defined by both a facility and a time; this is suppressed for the sake of notational simplicity)
- d_{ij} = number of double trailers flowing on arc (i, j)
- x_{ij}^k = number of units of commodity k flowing on arc (i, j)
- y_{ij} = number of empty trailers flowing on arc (i, j)

Parameters

- c_{ij}^s = the cost of a single trailer configuration flowing on arc (i, j)
- c_{ij}^d = the cost of a double trailer configuration flowing on arc (i, j)
- b_j^k = supply of (> 0) or demand for (< 0) commodity k at node j

Sets

- V = the set of all time-space nodes j
- A = the set of all arcs (i, j)
- F = the set of all facilities f
- K = the set of all commodities k
- V_f = the set of all time-space nodes j corresponding to facility f

Given this notation, we can now state the *MCF* formulation for *LMREB*:

MCF-LMREB:

$$\text{Min} \quad \sum_{(i,j) \in A} (c_{ij}^d d_{ij} + c_{ij}^s s_{ij}) \quad (1)$$

st

$$\sum_{i:(j,i) \in A} x_{ji}^k - \sum_{i:(i,j) \in A} x_{ij}^k = b_j^k \quad \forall j \in V, k \in K \quad (2)$$

$$s_{ij} + 2d_{ij} = \sum_{k \in K} x_{ij}^k + y_{ij} \quad \forall (i, j) \in A \quad (3)$$

$$\sum_{j \in V_f} \left(\sum_{k \in K} b_j^k + \sum_{i:(j,i) \in A} y_{ji} - \sum_{i:(i,j) \in A} y_{ij} \right) = 0 \quad \forall f \in F \quad (4)$$

$$x_{ij}^k, y_{ij}, d_{ij}, s_{ij} \in Z^+ \quad \forall (i, j) \in A, k \in K \quad (5)$$

The objective function (1) sums the cost of all single and double trailer movements. Constraint set (2) contains the flow balance constraints, which ensure that the supply or demand for each commodity is met at each node. Constraint set (3) enforces the relationship between single and double trailer movements and the flow of commodities (i.e. loads) and empties. Constraint set (4) contains the constraints that enforce equipment balance in the network. These are simply the flow balance constraints for the empty trailer commodity

(note that empty trailers, unlike loaded trailers, do not have time constraints). Constraint set (5) enforces the integrality of all variable types.

Although this formulation is a valid model for *LMREB*, it is not a tractable approach for problem instances of realistic size. It suffers from two major sources of computational difficulty. First, this model is quite large. For example, a problem instance with only 1000 commodities and 100 facilities would have on the order of three million arcs, 3 billion variables, and 30 billion constraints, assuming time intervals of five minutes. This stems from the need to use a time-space network to capture the timing constraints for the loads.

The second problem is that the strength of the LP relaxation is extremely poor. This stems from the linearization of the cost function. Specifically, whenever the value

$$\sum_{k \in K} x_{ij}^k + y_{ij}$$

is odd for a given arc (i, j) , the optimal solution to the LP relaxation is to assign the fractional value

$$\frac{\sum_{k \in K} x_{ij}^k + y_{ij}}{2}$$

to d_{ij} and value 0 to s_{ij} . This results in a large number of nodes in the branch-and-bound tree. Each of these nodes in turn is a large LP, as discussed above. Computational results in Section 5 demonstrate the practical implications of these challenges.

4 A Cluster-Based Modeling Approach

Given the challenges posed by a *MCF* approach, how can we develop an alternative model that addresses these challenges? First, observe that any load can feasibly move from its origin to its destination as a single trailer configuration, leaving its origin no earlier than its earliest pickup time and arriving at its destination no later than its latest delivery time; otherwise, the problem would be infeasible. Furthermore, unless the load's time window is exactly the travel time between these two locations, then there are an infinite number of equivalent solutions, with the load leaving at any time between the earliest available and latest feasible pickup times. The *MCF-LMREB* model determines both whether the load travels direct as a single trailer, and if so, at what time it travels. However, the objective value does not depend on such timing decisions. Thus, the problem could be simplified by disaggregating these decisions – first finding an optimal matching and routing of trailers and empties (for which at least one time-feasible schedule must exist) and then selecting a valid schedule of departure times.

Second, observe that the weak LP relaxation in *MCF-LMREB* stems from the assignment of half of a double trailer configuration to an arc. If only one trailer moves over an arc, then a double trailer configuration should not be a permissible option, thereby preventing the fractional solutions.

Based on these observations, we have therefore focused not on the flow of commodities over arcs but instead on the *clustering of loaded and empty trailers* that interact with each other via sharing a common tractor for some portion of their routes. We define a *cluster* to be a set of loads, a set of empty trailers, the routes they take, and the tractor configurations that pull them. For example, Figures 2 through 4 show three possible clusters. Figure 2 shows a cluster made up of a single load that moves direct from origin to destination as a single trailer configuration. Figure 3 shows a cluster in which a single load moves direct from origin to destination, sharing a tractor with an empty trailer also moving between those two locations. Figure 4 shows a cluster in which the load from A to C is pulled along with an empty trailer from A to B . The empty trailer is then dropped off and a load from B to D is picked up. The $A - C$ and $B - D$ loads are combined and pulled by a single tractor from B to C . The $A - C$ load is dropped off and the $B - D$ load continues as a single trailer configuration to its destination.

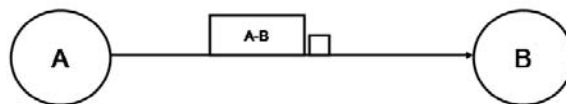


Figure 2: Single Load Moving Direct from Origin to Destination

It is trivial to compute the cost of a cluster by simply summing the costs of the tractor configurations over each segment in the cluster. It is also straightforward to determine whether a given cluster corresponds to at least one feasible time schedule. We can then define a variable x_c for each time-feasible cluster c which determines whether that cluster will be used in the solution. [Note that in most cases x_c could be represented by a binary variable. It is possible, however, to have multiple pairs of empties moving together between a given origin and

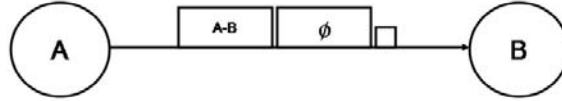


Figure 3: Load Moving Direct with an Empty Trailer

destination, and therefore such clusters can have values greater than one.] This leads to the following model, which is a variation of the classical *set partitioning* (*SP*) problem ([3], [22]).

Variables

- x_c = number of times cluster c appears in the solution

Parameters

- v_c = the cost of cluster c
- $\delta_{cl} = 1$ if cluster c includes load l , else 0
- η_{cf} = the net flow of trailers (both loaded and empty) through facility f relative to cluster c ; a positive value indicates a surplus and a negative value indicates a deficit

Sets

- C = the set of time-feasible clusters c
- L = the set of loads l
- F = the set of all facilities f

Given this notation, we can now state the cluster-based formulation for *LMREB*:

C-LMREB:

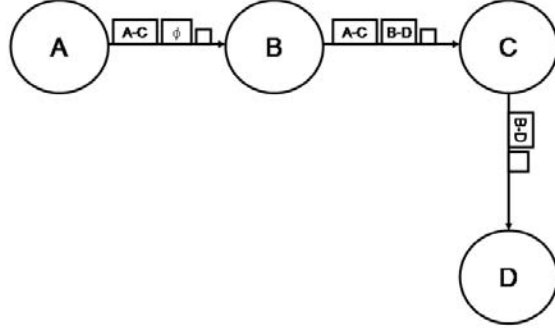


Figure 4: Complex Cluster Example

$$\text{Min} \quad \sum_{c \in C} v_c x_c \quad (6)$$

st

$$\sum_{c \in C} \delta_{cl} x_c = 1 \quad \forall l \in L \quad (7)$$

$$\sum_{c \in C} \eta_{cf} x_c = 0 \quad \forall f \in F \quad (8)$$

$$x_c \in Z^+ \quad \forall c \in C \quad (9)$$

The objective function (6) sums the cost of all chosen clusters. Constraint set (7) ensures that every load is contained in exactly one chosen cluster. Constraint set (8) looks at the net flow of trailers (both loaded and empty) through facility f for each chosen cluster c . These must collectively sum to zero to ensure balance in the network. Constraint set (9) enforces the integrality of the variables.

This new formulation provides a number of benefits. First, the model has far fewer constraints than *MCF-LMREB* (just one for each load and for each facility) and is more simply defined. Second, the LP relaxation is strengthened, because there is no opportunity to move half of a pair of loads between two facilities unless both loads exist. Third, the cluster-based model allows us to capture the time associated with reconfiguring tractors as loads are put together and broken apart, which cannot be captured in the *MCF*-based model. Given the tight time constraints of an expedited package carrier, this improved accuracy in computing timing issues can impact the feasibility of the final solution.

Finally, *C-LMREB* provides substantial flexibility in expanding the problem scope. For example, it might be more cost-effective for two loads to meet and be combined at a non-facility location, such as a highway rest area, which leads to less circuitous mileage. Such a possibility is valid for the cluster definition and would not change the structure of *C-LMREB*, whereas the *MCF-LMREB* network would have to be greatly expanded with additional nodes for these opportunities to be permitted. A second example is the use of additional tractor configurations. In some parts of the United States, three trailers can be pulled behind a single tractor, for example. This can be captured in the structure of *C-LMREB* via the existing variable definition but would require one new set of variables in *MCF-LMREB* for each new tractor configuration. A third example is the use of new modes (for example, railroad movements), which can again be captured implicitly within the cluster variables, rather than having to explicitly incorporate them in the formulation.

These benefits, of course, come at a cost – theoretically, there are an exponential number of clusters to be considered. There are several ways, however, to address this obstacle. These include *feasibility*, *dominance*, and *indifference* to eliminate clusters a priori that are guaranteed not to be part of an optimal solution. *Delayed column generation* can leverage the dual information to help identify the optimal cluster set. Finally, *heuristics* can be used to find high-quality solutions quickly by exploiting problem structure.

Feasibility Although there are theoretically an infinite number of ways in which loads and empty trailers can be combined, many of these combinations will not be time-feasible. This is particularly true for complex clusters, comprised of many loads and multi-stop routings, because the time consumed by reconfiguring tractors and driving circuitous miles quickly can lead to the violation of a load’s time window. Clearly, only clusters for which at least one valid time schedule exists need to be included in the model.

Dominance For a given set of loads and empty trailers, there may be many different ways in which these trailers can be combined to form a cluster. In fact, if non-facility meets are permitted, this number is infinite. However, each of these clusters will correspond to the same column in the constraint matrix. This is because the column is made up of the loads that are covered, which doesn’t depend on the way in which the loads are matched or routed, and the impact on node balance. Note that node balance is unaffected because all loads are routed from origin to destination – thus, they contribute to a deficit at their origin, a surplus at their destination, and for all intermediate nodes, they must flow both in and out, netting to a contribution of zero. Thus, for a given set of loads and empty trailers, we only need to include the cluster with the lowest cost, as this would always be chosen over the others. We refer to this cluster as *dominant*; all dominated clusters can be excluded from the model without impacting solution quality. An example of dominance is given in Figure 5.

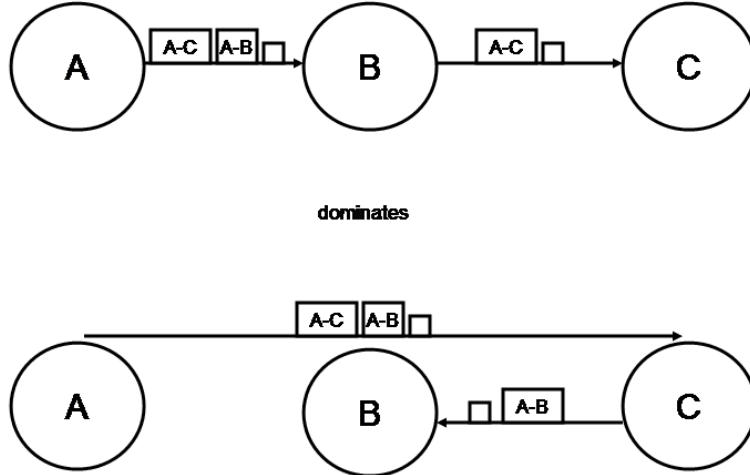


Figure 5: Two clusters carrying the same loads

Indifference Consider the three clusters shown in Figure 6. Clearly, we are *indifferent* as to whether the solution contains cluster A or clusters B and C . In fact, they correspond to the same real-world solution, because cluster A can be decomposed into two pieces, B and C , each of which itself is a valid cluster. Because the cost function is additive with respect to the individual tractor movements and because timing feasibility depends on the matching and routing of individual loads, we can limit ourselves to clusters that are *minimally independent* – that is, valid clusters that cannot be defined as the union of other valid clusters. Another way to define this is to say that a cluster is minimally independent if and only if, for any strict subset of the loads covered by that cluster, there exists at least one tractor movement pulling both a load from that set and a load from its complement. By considering only minimally independent clusters, the size of the problem can greatly be reduced.

Column Generation Although the number of feasible, dominant, and minimally independent clusters may still be quite large, the number of clusters that actually appear in the optimal solution cannot exceed the number of loads plus the number of empty trailers being redistributed throughout the network, and in fact will typically be far smaller than this as loads and empties are combined. Therefore, most clusters will be irrelevant to the problem. When solving the individual LP relaxations in the C - $LMREB$ model, the dual information can therefore be leveraged, via delayed column generation, to identify promis-

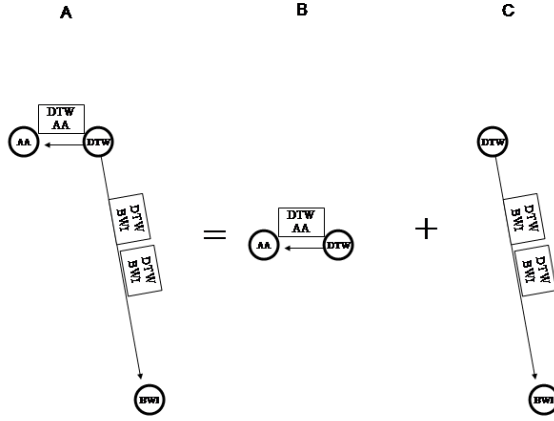


Figure 6: Cluster A versus Clusters B and C

ing clusters. At the IP level, this column generation approach can be imbedded within the branch-and-bound tree via *branch-and-price* ([6], [4], [26]).

Heuristics Finally, we observe that it is often less important in practice to find a provably optimal solution than it is to find a high-quality solution quickly. An added benefit of the cluster-based model is that the time taken to find feasible solutions can be controlled by limiting the set of clusters to consider. These solutions can then be improved upon as time permits by providing additional clusters. The careful selection of the initial clusters, based on exploiting problem structure and leveraging user expertise, can lead to high-quality solutions quickly, as demonstrated in the following section.

5 Computational Results

The purpose of our computational experiments was two-fold. First, we wanted to assess the tractability of the cluster-based modeling approach. Second, we wanted to evaluate the solution quality that could be gained through the integration of load matching and routing with equipment balancing, even when limiting the number of clusters considered.

All experiments were conducted on data from a regional sub-network provided by UPS. This sub-network contains 263 nodes (i.e. facilities), 2,644 arcs, and 2,067 loads. The table below shows the distribution of node imbalances.

Surplus (+) / Deficit (-)	Number of Nodes
< -10	7
-10 to -6	6
-5 to -1	37
Balanced	90
1 to 5	118
6 to 10	5

Table: System balance

Multi-Commodity Flow Based Modeling Approach We began by revisiting the *MCF*-based modeling approach to verify our hypothesis that this approach would not be tractable.

To assess our theory that the LP relaxation would be intolerably weak, we ignored the constraints that require loads to be picked up and delivered within their available time windows. This enabled us to use a facility-based network rather than a time-space network. The resulting model, even with this relaxation of the time constraints, had over 1.2 million variables and had constraints numbering in the hundreds of thousands. We solved the LP relaxation of this model using AMPL running CPLEX 9.1 on a Sun Fire 280R computer with an UltraSPARC III Cu 1.2 GHz Superscalar SPARC V9 processor and 8 GB of main memory. This single LP took more than a minute to solve. As expected, the value for all variables s_{ij} , corresponding to the movement of single trailers, was zero. Furthermore, all of the variables x_{ij}^k and y_{ij} had integer values as well. This is also not surprising – because the network does not have capacity constraints on the arc flows, there is no incentive to split commodities over multiple paths. The variables d_{ij} , however, showed significant fractionality – 745 of the 1,296 arcs used had a value of $n + 0.5$, where n is some integer value of 0 or greater. This is to be expected given the cost structure, in which half of a double trailer configuration is less costly than a single trailer configuration.

Now consider the impact of branching within this model. Choose a given arc (i, j) for which d_{ij} is fractional. On one branch d_{ij} will be set to be less than or equal to 0.5 less than its current value and on the other branch it will be set to be greater than or equal to 0.5 more than its current value. As the flow is redistributed, all remaining arcs will continue to have incentive to split single trailers. Thus, the number of fractional variables in these new nodes is unlikely to decrease significantly and may even increase. This suggests that the branch-and-bound tree could contain on the order of $2^{1,000}$ nodes, each of which is a sizeable linear program.

It is certainly true that heuristics could be used to find good initial integer-feasible solutions and that specialized branching strategies might be developed to improve performance. Nonetheless, the fundamental challenge remains that the non-linear cost structure, in which the cost of pulling half a double trailer is less expensive than pulling a single trailer, will always encourage fractional values for variables associated with the flow of odd numbers of trailers over a given arc, resulting in both a poor lower bound and a large branch-and-bound

tree. Furthermore, recall that these estimates are based on the problem *without any time constraints*.

We next estimated the size of the formulation once timing considerations were added into the model. The number of variables increased to more than 600 million and the number of constraints increased to over 72 million. Again, this might be improved by careful exploitation of the problem structure – for example, by removing nodes that have only one arc in and one arc out and replacing the two arcs with a single “super arc”. Nonetheless, we believe that the problem will remain quite large and that this, in conjunction with the weak LP relaxation resulting from the non-linear cost structure, will render a multi-commodity flow based approach intractable.

Cluster-Based Modeling Approach The remainder of our experiments therefore focused on the cluster-based approach. We used the same regional sub-network and set of loads. In all experiments on this model, we limited ourselves to a pre-defined set of “cluster templates,” enumerating all clusters that matched these templates and solving explicitly, rather than using column generation to find a provably optimal solution. In all remaining experiments, the number of constraints was 2,330 – one cover constraint for each of the 2,067 loads and one balance constraint for each of the 263 nodes.

Cluster Experiment One In the first cluster-based experiment, we found an upper bound by disaggregating the problems and by requiring all loads to move direct from origin to destination. Specifically, we limited the model to 7,124 clusters derived from four templates (see Figure 7):

- a load moving as a single trailer direct from origin to destination;
- two loads with the same origin and destination and compatible time windows moving as a double trailer direct from origin to destination;
- a single empty trailer moving from node A to node B ;
- two empty trailers moving as a double trailer from node A to node B .

The IP solver was allowed to run for 20,000 nodes, taking approximately 15 minutes. An optimality gap of 1.4% was achieved within a few seconds (recall that this optimality gap is relative to the restricted problem for the given cluster templates, not for the true problem with all possible clusters included). The final optimality gap was 0.36%, with 15,590 nodes pending. The total solution cost was \$614,476.

The initial root node had an objective value of \$573,448, suggesting a relatively tight LP bound, as the current best solution is within 7% of this. The slow convergence to provable optimality is largely due to the fact that it is still possible to benefit from replacing a single empty trailer movement with one half

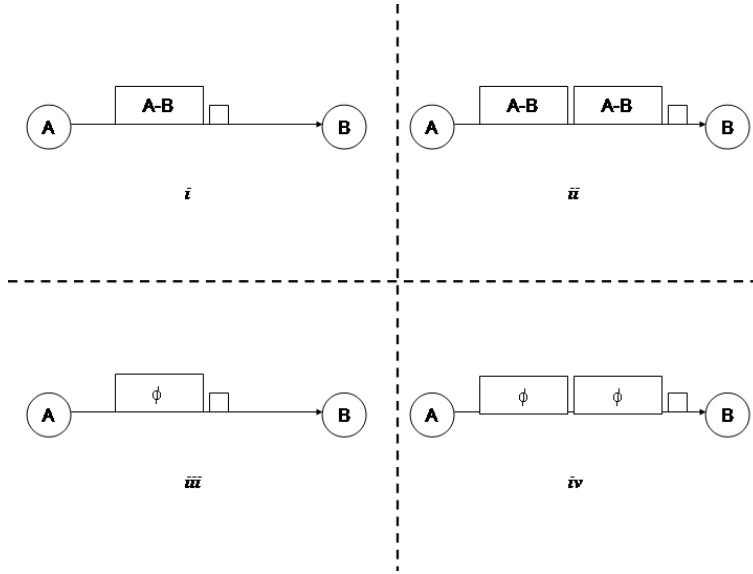


Figure 7: Templates for Cluster Experiment 1

of a double empty trailer movement, but this has limited impact on cost and a near-optimal solution was found in minutes.

The following tables show the distribution of clusters used in this solution, as well as their impact on cost. Observe that the movement of empty trailers constitutes approximately 14% of the cost, and that 51% of the clusters are single trailer configurations.

	# Clusters	Percent
Cluster i	713	44%
Cluster ii	677	42%
Cluster iii	111	7%
Cluster iv	127	8%
Total	1628	

Table: Cluster distributions for cluster experiment 1

	Cost	Percent
Cluster i	\$272,299	44%
Cluster ii	\$256,412	42%
Cluster iii	\$37,080	6%
Cluster iv	\$48,685	8%
Total	\$614,476	

Table: Costs for cluster experiment 1

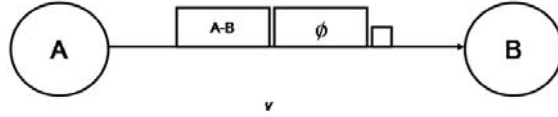


Figure 8: Additional Template for Cluster Experiment 2

Cluster Experiment Two In the second cluster-based experiment, we allowed loaded and empty trailers to be combined, but still required all loads to move direct from origin to destination. In other words, we added the cluster template shown in Figure 8:

- a load moving direct from origin to destination in conjunction with an empty trailer.

This resulted in 2,067 additional clusters, for a total number of 9,191 clusters.

The IP solver was again allowed to run for 20,000 nodes, taking approximately 15 minutes. An optimality gap of 0.28% was achieved within a few seconds. The final optimality gap was 0.2%, with 18,569 nodes pending.

The best known integer solution had an objective value of \$588,486, a 4.2% improvement over the prior solution. The percent of loads that move as single configurations decreased from 34% to 26%, and the percent of empty trailers that move as single configurations decreased from 30% to 12%. Overall, 38% of the tractors now pull a single trailer, as opposed to 51% in the previous experiment. Additionally, note that 185 clusters combining a loaded and empty trailer are selected, suggesting the role of integration in these improved results.

	# Clusters	Percent
Cluster <i>i</i>	540	35%
Cluster <i>ii</i>	671	43%
Cluster <i>iii</i>	52	3%
Cluster <i>iv</i>	96	6%
Cluster <i>v</i>	185	12%
Total	1544	

Table: Cluster distributions for cluster experiment 2

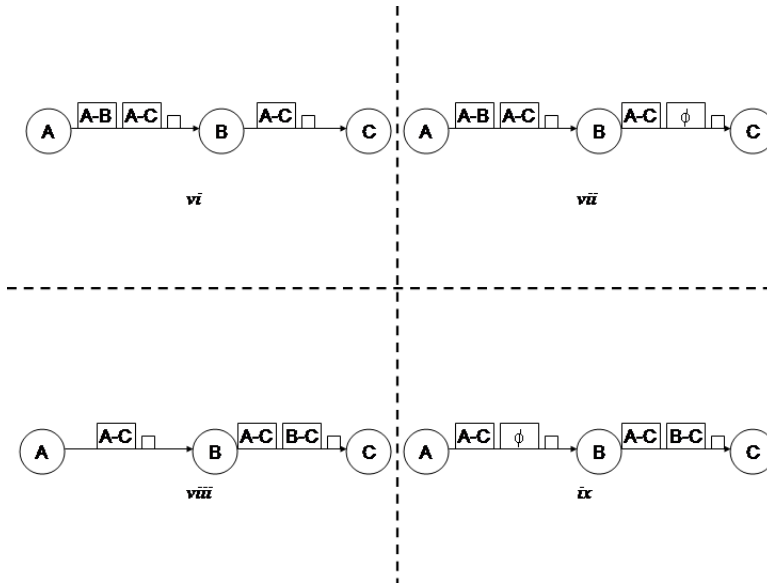


Figure 9: Additional Templates for Cluster Experiment 3

	Cost	Percent
Cluster <i>i</i>	\$206,252	35%
Cluster <i>ii</i>	\$255,121	43%
Cluster <i>iii</i>	\$17,488	3%
Cluster <i>iv</i>	\$30,687	5%
Cluster <i>v</i>	\$78,938	13%
Total	\$612,950	

Table: Costs for cluster experiment 2

Cluster Experiment Three In the third cluster-based experiment, we included additional cluster templates to allow some circuitous mileage in order to match trailers. The new cluster templates are illustrated in Figure 9. 35,516 new clusters were added based on these templates, resulting in a total of 44,707.

The IP solver was again allowed to run for 20,000 nodes, taking approximately one hour. An optimality gap of 1.38% was achieved within a few seconds. The final optimality gap was 0.66%, with 18,243 nodes pending. Notice that the convergence is slower than that of the prior two cluster-based experiments, due to the significant increase in number of variables and thus the solution time for the individual LP's.

The following tables show the results of the current best integer solution. The objective value shows an 8.9% improvement over the first cluster set and a 4.9% improvement over the second cluster set. In this solution, the percent of

loads moving as single configurations fully from origin to destination dropped from 26% to 17%. Another 5% loads move over one leg as a single and over a second leg as part of a double. The remaining 78% move fully from origin to destination as part of a double configuration. Approximately 10% of the empty trailer movements now travel as singles. Overall, 31% percent of the tractor movements are now made up of a single trailer. 14% of the clusters (193) now contain both loaded and empty trailers on at least one leg. For those loads incurring circuitous miles, the average excess distance was approximately 43 miles.

	# Clusters	Percent
Cluster <i>i</i>	341	24%
Cluster <i>ii</i>	625	45%
Cluster <i>iii</i>	44	3%
Cluster <i>iv</i>	93	7%
Cluster <i>v</i>	126	9%
Cluster <i>vi</i>	52	4%
Cluster <i>vii</i>	33	2%
Cluster <i>viii</i>	56	4%
Cluster <i>ix</i>	34	2%
Total	1404	

Table: Cluster distributions for cluster experiment 3

	Cost	Percent
Cluster <i>i</i>	\$114,877	21%
Cluster <i>ii</i>	\$233,617	42%
Cluster <i>iii</i>	\$15,911	3%
Cluster <i>iv</i>	\$32,229	6%
Cluster <i>v</i>	\$51,233	9%
Cluster <i>vi</i>	\$32,067	6%
Cluster <i>vii</i>	\$23,041	4%
Cluster <i>viii</i>	\$35,711	6%
Cluster <i>ix</i>	\$21,039	4%
Total	\$559,725	

Table: Costs for cluster experiment 3

Cluster Experiment Four In the final cluster-based experiment, we provided six additional cluster templates, as illustrated in Figures 10 and 11. 119,385 new clusters were added based on these templates, resulting in a total of 164,092.

The IP solver was again allowed to run for 20,000 nodes, taking approximately 5.5 hours. An optimality gap of 2.24% was achieved within a few minutes. The final optimality gap was 1.4%, with 18,322 nodes pending.

The following tables show the results of the current best integer solution. The objective value shows a 0.6% improvement over the third cluster set, a

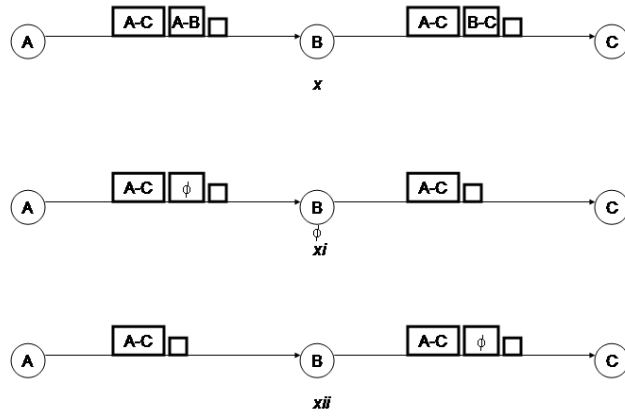


Figure 10: Additional Templates for Cluster Experiment 4

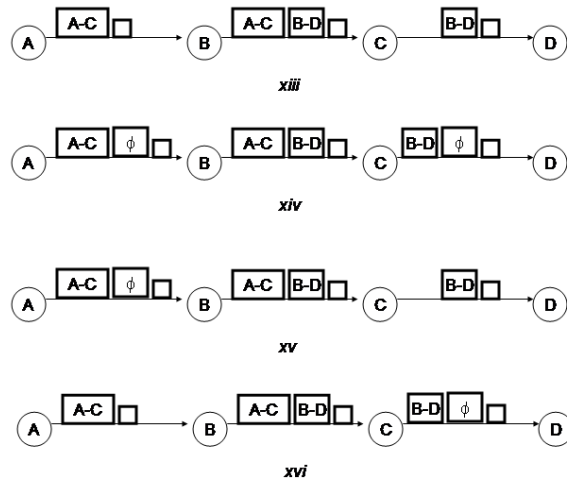


Figure 11: Additional Templates for Cluster Experiment 4

5.5% improvement over the second cluster set, and a 9.5% improvement over the third cluster set. In this solution, the percent of loads moving as single configurations fully from origin to destination dropped to 15%. Another 2% loads move partially as a single and partially as part of a double. The remaining 83% move fully from origin to destination as part of a double configuration. Overall, only 26% percent of the tractor movements are now made up of a single trailer. For those loads incurring circuitous miles, the average excess distance was approximately 54 miles.

	# Clusters	Percent
Cluster <i>i</i>	318	24
Cluster <i>ii</i>	541	41
Cluster <i>iii</i>	45	3
Cluster <i>iv</i>	108	8
Cluster <i>v</i>	113	9
Cluster <i>vi</i>	18	1
Cluster <i>vii</i>	18	1
Cluster <i>viii</i>	13	1
Cluster <i>ix</i>	17	1
Cluster <i>x</i>	138	10
Cluster <i>xi</i>	2	0
Cluster <i>xii</i>	0	0
Cluster <i>xiii</i>	2	0
Cluster <i>xiv</i>	0	0
Cluster <i>xv</i>	1	0
Cluster <i>xvi</i>	0	0
Total	1334	

Table: Cluster distributions for cluster experiment 4

	Cost	Percent
Cluster <i>i</i>	\$108,868	20%
Cluster <i>ii</i>	\$207,908	37%
Cluster <i>iii</i>	\$16,288	3%
Cluster <i>iv</i>	\$37,732	7%
Cluster <i>v</i>	\$44,643	8%
Cluster <i>vi</i>	\$13,486	2%
Cluster <i>vii</i>	\$13,449	2%
Cluster <i>viii</i>	\$8,427	2%
Cluster <i>ix</i>	\$10,798	2%
Cluster <i>x</i>	\$88,622	16%
Cluster <i>xi</i>	\$2,541	0%
Cluster <i>xii</i>	\$0	0%
Cluster <i>xiii</i>	\$2,514	0%
Cluster <i>xiv</i>	\$0	0%
Cluster <i>xv</i>	\$909	0%
Cluster <i>xvi</i>	\$0	0%
Total	\$556,185	

Table: Costs for cluster experiment 4

Upper and Lower Bounds Although the fourth cluster-based experiment was solved to within 1.4% of optimality, this is only with respect to the given set of clusters. It does not provide us with a true lower bound for the problem. Such a lower bound can be constructed by computing the cost of each load moving direct from origin to destination at one-half the cost of a double trailer movement over that leg. This is then added to a lower bound on the cost of equipment balancing, which can be found by solving a transportation problem, with the cost of each arc again being one-half the cost of a double trailer movement over that leg. With respect to this bound (which yields a value of \$482,849 in the current problem instance), the solution from experiment 4 has an optimality gap of approximately 15%. However, this is clearly a weak lower bound, based on the assumption that all loads can be paired with another trailer without incurring any circuitous mileage or violating time constraints. We hope to develop tighter bounds in the future to establish a more realistic optimality gaps.

As an alternative method for assessing the quality of our solution, we have also compared it to the upper bound provided by UPS' existing approach to solving the problem. UPS used their current methodology to solve the same data set as was considered in our experiments. Their solution had an objective value of \$585,128, which is approximately 5% higher than the solution value found in experiment four. Additionally, they have suggested that the particular sub-network they provided to us underestimates the potential opportunities for savings, because it considers a small geographical area and contains loads that typically have very tight time windows. When applying the cluster-based

modeling approach to the full national network, they anticipate greater potential savings because longer movements (for example, between the East and West coasts) will have far more opportunities to combine loads.

6 Conclusions and Future Research

Load matching/routing and equipment balancing are two combinatorially challenging problems faced by small package carriers. In each of these problems, cost savings can be realized by exploiting the fact that two trailers pulled together behind a single tractor travel at lower cost than two trailers being pulled separately. The integration of these two problems can yield even greater opportunities for cost savings, because loaded and empty trailers may also be paired with each other. However, this integrated planning problem is even more challenging, with its non-linear cost function, tight time constraints, and problem size presenting significant computational challenges for traditional network flow-based approaches.

We have developed an alternative cluster-based modeling approach in which the variable definition specifically addresses these challenges, implicitly capturing the time constraints and simultaneously improving the quality of the LP relaxation. To the best of our knowledge, this is the first published approach to integrating these two problems. Although the cluster-based model has a very large number of variables, we discuss methods by which this set of variables can be reduced. Furthermore, we provide computational experiments to demonstrate that even with only a limited set of clusters being taken into consideration, high-quality solutions can be obtained. We believe that other transportation arenas (for example, *less-than-truckload* carriers) can leverage these results as well. In addition, this model is structurally flexible and can easily be modified to incorporate other travel modes (for example, rail), other tractor configurations (for example, triples where permitted), and out-of-network meet points. We also believe that it can naturally be extended to further integrate other aspects of the planning process such as the allocation of packages to trailer types, which currently serves as an input to the integrated problem considered here.

There are a number of extensions to this research that we hope to address in the future. In order to improve solution quality, new cluster templates can be identified and evaluated, and dual information can be leveraged within a branch-and-price framework to generate new clusters. Computational performance can be improved through the development of tighter lower bounds, better branching strategies, and a stronger LP relaxation. In particular, replacing the notion of a load with that of a *commodity* (that is, a set of loads sharing the same origin, destination, and time window) can improve performance by reducing redundancies within the solution space. Beyond this, the further integration of other planning decisions within this modeling framework can lead to additional reductions in system cost.

Acknowledgment

We gratefully acknowledge the assistance of Holly Shoals, Shirin Mehraban, and Anthony Celmer.

References

- [1] R. Ahuja, T. Magnanti, and J. Orlin. (1993.) *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, NJ.
- [2] A. Assad. (1978). "Multicommodity Network Flows—a Survey." *Networks* 8: 37–91.
- [3] E. Balas and M. Padberg. (1976). "Set Partitioning: A Survey" *SIAM Review* 18: 710 - 760.
- [4] C. Barnhart, E. Johnson, G. Nemhauser, M. Savelsbergh, and P. Vance. (1998). "Branch-and-Price: Column Generation for Solving Huge Integer Programs." *Operations Research* 46: 316 - 329.
- [5] C. Barnhart, N. Krishnan, K. Daeki, and K. Ware. (2002). "Network Design for Express Shipment Delivery." *Computational Optimization and Applications* 21: 239-262.
- [6] C. Barnhart, C. Hane, and P. Vance. (2000). "Using Branch-and-Price-and-Cut to Solve Origin-Destination Integer Multicommodity Flow Problems." *Operations Research* 48: 318 - 236.
- [7] J. Castro and N. Nabona. (1996). "Implementation of Linear and Non-linear Multicommodity Network Flows." *European Journal of Operational Research* 92: 37-53.
- [8] P. Chardaire, G. McKeown, S. Verity-Harrison, and S. Richardson. (2005). "Solving a Time-Space Network Formulation for the Convoy Movement Problem." *Operations Research* 53: 219 - 230.
- [9] A. Cohn and C. Barnhart. (2003). "Improving Crew Scheduling by Incorporating Key Maintenance Routing Decisions." *Operations Research* 51: 387 - 396.
- [10] J. Cordeau, G. Stojković, F. Soumis, and J. Desrosiers. (2001). "Benders Decomposition for Simultaneous Aircraft Routing and Crew Scheduling." *Transportation Science* 35: 375 - 388.
- [11] J. Cordeau, F. Soumis, and J. Desrosiers. (2000). "A Benders Decomposition Approach for the Locomotive and Car Assignment Problem." *Transportation Science* 34: 133 - 149.
- [12] J. Eckstein and Y. Sheffi. (1987). "Optimization of Group Line-Haul Operations for Motor Carriers Using Twin Trailers." *Transportation Research Record* 1120: 12 - 23.

- [13] <http://www.fedex.com/us/about/today/companies/ground/facts.html?link=4>.
- [14] Federal Highway Administration. (2005). “The Freight Story: A National Perspective on Enhancing Freight Transportation.” *Freight Management and Operations* <http://ops.fhwa.dot.gov/freight/freight_analysis/freight_story/>.
- [15] I. Ghamlouche, T. Crainic, and M. Gendreau. (2003). “Cycle-Based Neighborhoods for Fixed-Charge Capacitated Multicommodity Network Design.” *Operations Research* 51: 655 - 667.
- [16] J. Kennington. (1978). “A Survey of Linear Cost Multicommodity Network Flows.” *Operations Research* 2: 209-236.
- [17] D. Kim and C. Barnhart. (1997). “Multimodal Express Shipment Service Design: Models and Algorithms.” *Computers and Industrial Engineering* 33: 685-688.
- [18] D. Klabjan, E. Johnson, G. Nemhauser, E. Gelman, and S. Ramaswamy. (2002). “Airline Crew Scheduling with Time Windows and Plane-Count Constraints.” *Transportation Science* 36: 337 - 348.
- [19] M. Kuby, and R. Gray. (1993). “The Hub Network Design Problem with Stopovers and Feeders: The Case of Federal Express.” *Transportation Research A* 27: 1-12.
- [20] M. Lohatepanont and C. Barnhart. (2004). “Airline Schedule Planning: Integrated Models and Algorithms for Schedule Design and Fleet Assignment.” *Transportation Science* 38: 19 - 32.
- [21] M. Lübbecke and U. Zimmermann. (2003). “Engine Routing and Scheduling at Industrial In-Plant Railroads.” *Transportation Science* 37: 183 - 197.
- [22] G. Mitra and E. El-Darzi. (1990). “Set Covering and Set Partitioning: A Collection of Test Problems.” *OMEGA* 18: 195 - 201.
- [23] A. Newman and C. Yano. (2000). “Scheduling Direct and Indirect Trains and Containers in an Intermodal Setting.” *Transportation Science* 34: 256 - 270.
- [24] B. Rexing, C. Barnhart, and T. Kniker. (2000). “Airline Fleet Assignment with Time Windows.” *Transportation Science* 34: 1 - 20.
- [25] United States Department of Transportation: Bureau of Transportation Statistics. (2003). “US International Trade and Freight Transportation Trends.” <<http://www.bts.gov>>.
- [26] F. Vanderbeck. (2000). “On Dantzig-Wolfe Decomposition in Integer Programming and Ways to Perform Branching in a Branch-and-Price Algorithm” *Operations Research* 48: 111 - 128.

- [27] P. Wu, J. Hartman, and G. Wilson. (2005). “An Integrated Model and Solution Approach for Fleet Sizing with Heterogeneous Assets” *Transportation Science* 39: 87 - 103.
- [28] Yahoo! Finance. (2005). UPS Company Profile. <<http://finance.yahoo.com/q/pr?s=UPS>>.