ROUTING OF MULTIMODAL FREIGHT TRANSPORTATION USING A CO-SIMULATION OPTIMIZATION APPROACH

Yanbo Zhao
Department of Electrical Engineering, University of Southern California
3740 McClintock Ave. Los Angeles, CA 90089
Tel: 213-740-2244; Email: yanbozha@usc.edu

Petros Ioannou, Corresponding Author
Department of Electrical Engineering, University of Southern California
3740 McClintock Ave. Los Angeles, CA 90089
Tel: 213-740-4452; Email: ioannou@usc.edu

Maged Dessouky
Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089
Tel: 213-740-4891; Email: maged@usc.edu

Word count: 5962 words text + 6 tables/figures x 250 words (each) = 7462 words

Submission Date: November 15, 2016
ABSTRACT
The complexity and dynamics of multimodal freight transportation with the unpredictability of the effects of incidents, disruptions and demand changes make the optimum routing of freight traffic a challenging task. Making routing decisions in a multimodal transportation environment to minimize a certain objective relies on estimating the dynamical states of the multimodal traffic network. The purpose of this paper is to formulate a multimodal freight routing problem and then propose a COSMO (CO-Simulation Optimization) approach that can lead to more efficient decisions in freight routing by exploiting the availability of powerful computational software tools in the state estimations of complex and dynamic multimodal transportation networks. In the proposed approach, we develop a novel load balancing methodology to optimize routing decisions from an overall system perspective. A simulation testbed consisting of a road traffic simulation model and a rail simulation model for the Los Angeles/Long Beach Port area has been developed and applied to demonstrate the efficiency of the proposed approach.

Keywords: Freight Routing, Multimodal, Simulation, Optimization
1. INTRODUCTION

Efficient freight movement is an essential factor not only in urban transportation but also in social and economic development as well as environmental considerations (1-3). The growth of worldwide trade will significantly increase traffic congestion and air pollution due to existing congestion in the urban transportation infrastructure especially in metropolitan areas with major ports such as Los Angeles/Long Beach where there is a high concentration of both freight and passenger traffic that share the same infrastructure. One of the biggest challenges for freight transport efficiency in such a multimodal environment arises from the fact that the same rail and road networks are used for moving people in addition to goods which leads to non-homogeneous traffic. This non-homogeneity has a detrimental impact on the transportation system performance because of the differences of vehicle sizes and dynamics between passenger and freight transport. The freight vehicles such as freight trains and trucks take longer distances to stop and time to accelerate from a stopping position, consume more fuel and generate more air pollution compared to passenger vehicles. The situation becomes even worse during incidents and disruptions that lead to network changes such as road or railway closures that require rapid response and distribution of freight traffic across the multimodal network. Without efficient routing of the freight transport, the transportation network will face severe capacity shortages, inefficiencies, and route load imbalances across the network in space and time. Therefore a more efficient freight routing system could save transport costs and contribute to sustainability and efficiency of the entire urban transportation network.

Due to the important role of freight transportation, numerous researchers have addressed the issue of multimodal freight transportation routing and scheduling (1-27). Jourquin and Beuthe presented a multimodal freight model based on a digitized geographic network (1). Southworth and Peterson developed a multi-layer intermodal shipment routing model in (2). The intermodal freight transport between rail and road has been described in (4-6). As a fundamental issue for optimum routing, the multicriteria shortest path problem in a multimodal network has been studied by many researchers. Modesti and Sciomachen applied a link utility measure approach to solve the multiobjective shortest path problem (7). Lozano and Storchi considered the impact of modal transfer costs when finding the shortest multimodal path (8). Dynamic and stochastic routing for a multimodal transportation environment was studied in (9) and (10). Speed-up techniques for the shortest path algorithm have also been analyzed including Core-Based routing (11), label-setting and label-correcting methods (12) and the improved label setting algorithms (13). Optimization techniques have been commonly used to solve the multimodal transport routing and scheduling problem such as in (14-18). Guelat and Florian proposed a linear approximation algorithm to solve the multimodal multiproduct freight assignment problem (14). Castelli et al. used a Lagrangian-based heuristic procedure to solve the signal line and general network scheduling problem (15). Ham, Kim and Boyce showed the application of Wilson’s iterative balancing method in interregional multimodal shipments (16). Zografos et al. developed a dynamic programming based algorithm for multimodal time-scheduling with a shortest path algorithm (17). Mocchia et al. solved a multimodal routing problem with timetables and time windows by integrating heuristics and a column generation algorithm (18). The main difficulty is that these classical approaches of using mathematical models breaks down when faced with the control and optimization of complex networks such as multimodal freight networks that exhibit nonlinear travel time and cost functions which are difficult to mathematically represent with respect to the routing decisions and to find a closed form solution. The availability of fast computers and
software tools open the way for new approaches that go beyond the limitations of network complexity. The traffic flows and states can be better predicted using simulation models that are more complex and can capture phenomena that cannot be formulated with simple models (28). Some researchers tried to solve the multimodal routing problem from the aspect of user equilibrium in dynamic traffic assignment as in the unimodal road scheduling problem (19-25). The common idea of user equilibrium based traffic assignment is to search the routing decision such that the trip costs on the used routes for the same demand are equal and the costs of the used routes are less than the costs of the other unselected routes. Peeta and Mahmassani formulated and developed a simulation-based method for the dynamic traffic assignment problem in which the trip costs for each origin-destination demand are in equilibrium (19). The proposed simulation model based method was also applied in (20)-(22) to solve the intermodal routing problems. The equilibrium model between supply and demand has been also studied in which all trips for all suppliers share the same trip cost (23)-(24). A solution algorithm based on the cross entropy model was proposed and compared to the moving successive average method that was widely used in the user equilibrium problem (25). Afshin et al. developed a coordinated multimodal load balancing system in which the suppliers share route information so that it could lead to better solutions for the overall system (26). Moreover, Russ et al. and Yamada et al showed the applications of routing and scheduling approach in multimodal freight network design (27)-(28).

The simulation models can also be integrated with control and optimization techniques to provide better and more robust decisions (26). The purpose of this paper is to formulate and solve the dynamical multimodal freight routing problem by exploiting the availability of powerful computational software tools. We therefore propose a method we refer to as COSMO (CO-Simulation Optimization) as a potential innovation in dealing with multimodal transportation routing that cannot be handled by the traditional way. We are proposing a novel load balancing algorithm in the COSMO approach with estimates of the route states from outputs of the system simulators based on work in (26).

The paper is organized as follows. Section 2 gives a formulation of the optimum routing problem for multimodal freight transport. Section 3 proposes the COSMO approach and demonstrates how to solve the formulated problem with the proposed approach. Section 4 shows the experimental results of the proposed approach on portions of the transportation network in Southern California. Finally the conclusions are discussed in Section 5.

2. PROBLEM FORMULATION

A multimodal freight transportation network $G$ can be represented as a directed graph network consisting of a set of nodes ($N$) with a set of directed links ($L$) connecting the nodes. A link in the network could be one segment of a roadway or railway track while a node with zero length connects multiple links. All passenger and freight traffic start and end at certain network nodes. Let $I$ and $J$ be the sets of origin nodes and destination nodes respectively. Both $I$ and $J$ are a subset of the node set $N$. In this paper, we are dealing with the routing of freight traffic flows that are container flows between origin nodes and destination nodes. The analysis time horizon is discretized into $|K|$ small time intervals to formulate the problem. The notations that will be used throughout the paper are defined as follows:

$i$ The index of an origin node, $i \in I$;

$j$ The index of a destination node, $j \in J$;

$k$ The index of time, $k \in K$ where $K = \{0, 1, \ldots, |K|\}$;
The index of a link in the network $G$, $l \in L$;

The set of all paths from an origin $i$ to a destination $j$;

The index of a path from an origin $i$ to a destination $j$, $p \in P_{i,j}$;

The length of a time interval (unit: hour);

The total demand in number of containers departing from origin node $i$ to destination node $j$;

The number of containers departing from origin node $i$ to destination node $j$ using path $p$ with a departure time of $k$;

The traffic volume of link $l$ at time $k$;

The travel time of link $l$ at time $k$;

We next describe the constraints for our multimodal freight transport routing problem.

The first constraint ensures that the total amount of shipped containers throughout the analysis time horizon for each origin/destination pair equals the required demand amount.

$$\sum_{k} \sum_{p \in P_{i,j}} x_{i,j}^p (k) = d_{i,j}, \quad \text{for } \forall i \in I, \forall j \in J$$  \hspace{1cm} (1)

Let $Y(k) = [y_1(k), y_2(k), \ldots, y_{|L|}(k)]'$ be the vector of traffic volumes (unit: vehicles/hour) on links 1 to $|L|$ at time $k$. Then the relationship of the traffic volume on a link $l$ with the departure container traffic and other parameters in the network can be expressed as a nonlinear dynamical equation:

$$y_l(k+1) = f_l(y_l(k), a_l(k), X(k), k), \quad \text{for } \forall l \in L, \forall k \in K$$ \hspace{1cm} (2)

In (2), $f_l$ is a nonlinear and time-dependent function of the traffic volume of a link $l \in L$. The impact of the traffic volumes from adjacent links at time $k$ is denoted by $a_l(k)$ and $X(k)$ is the vector of departure freight traffic volumes from all origin nodes at time $k$ as in (3). Since $y_l(k)$ and $a_l(k)$ contain the impact of the previous departure container traffic before time $k$ (i.e., $X(r)$ for $\forall r < k$) so only $X(k)$ is included in equation (2). The link volumes in the transportation network are time-dependent due to various factors such as time-dependent passenger traffic, network changes, accidents and incidents.

Let $W(k) = [w_1(k), w_2(k), \ldots, w_{|L|}(k)]'$ be the vector of travel time (unit: $\Delta t$) of links 1 to $|L|$ at time $k$. $w_l(k)$ specifies the time length that a container takes to travel on link $l$ if it enters link $l$ at time $k$. The link travel time is a function of the link volume at time $k$ which is time-dependent because of the impact of the time-dependent passenger traffic, network incidents and railway dispatching decisions. The travel time of a link is dependent on not only the link flow on itself but also on the flows of the other links, therefore,

$$W(k) = g(Y(k), k), \quad \text{for } \forall k \in K$$ \hspace{1cm} (4)

Let $T_{i,j}^p(k)$ be the travel time (unit: $\Delta t$) of a path $p$ from an origin node $i$ to destination node $j$ if a container departs from origin $i$ at time $k$. Assume a path $p$ contains links...
$l_{p,1} \rightarrow \cdots \rightarrow l_{p,N_p}$, where $N_p$ is the number of links on this path $p$. Define $e_{p,n_p}(k)$ as the entering time at link $l_{p,n_p}$ for a container on path $p$ with a departure time of $k$ at the origin. Then the path travel time can be computed as follows:

$$T_{i,j}^p(k) = \sum_{n_p=1}^{N_p} w_{i,n_p} \left( e_{i,n_p}(k) \right)$$

where

$$e_{i,1}(k) = k, \quad \text{and} \quad e_{i,n_p}(k) = e_{i,n_p-1}(k) + w_{i,n_p} \left( e_{i,n_p-1}(k) \right), \quad \text{for } n_p = 1, \ldots, N_p - 1$$

Let $S_{i,j}^p(k)$ be the average cost for shipping one container from origin $i$ to destination $j$ with a departure time of $k$, then the objective function can be expressed as

$$\sum_{k \in K} \sum_{i \in I} \sum_{j \in J} \sum_{p \in P_{i,j}} S_{i,j}^p(k) x_{i,j}^p(k)$$

where

$$S_{i,j}^p(k) = C_{i,j}^p(k) + K_p T_{i,j}^p(k), \quad \text{for } \forall i \in I, \forall j \in J, \forall p \in P_{i,j}, \forall k \in K$$

In (8), $C_{i,j}^p(k)$ is the non-time independent cost (unit: dollar) generated by vehicle setup, distance cost, etc. per container from origin $i$ to destination $j$ at departure time $k$ with path $p$. This cost could be obtained directly based on the information such as path distance and used vehicle type. $K_p$ is the weight value of the travel time on path $p$.

In summary, the multimodal routing problem can be expressed as follows.

$$\min \sum_{k \in K} \sum_{i \in I} \sum_{j \in J} \sum_{p \in P_{i,j}} S_{i,j}^p(k) x_{i,j}^p(k)$$

subject to constraints (1) – (6) and

$$x_{i,j}^p(k) \geq 0 \quad \text{for } \forall i \in I, \forall j \in J, \forall p \in P_{i,j}, \forall k \in K$$

$$0 \leq y_{i,l}(k) \leq u_i(k) \quad \text{for } \forall l \in L, \forall k \in K$$

given $d_{i,j}, u_i(k)$ for $\forall i \in I, \forall j \in J, \forall l \in L, \forall k \in K$

where $u_i(k)$ is the capacity of link $l$ at time $k$, i.e. the maximum amount of vehicles that could be on link $l$ at time $k$.

The explicit forms of the dynamical functions in (2) and (4) are difficult to mathematically express directly due to the nonlinearities and complex variable interactions. Therefore, we propose a COSMO approach in which the traffic network simulation models are used to replace the mathematical functions of (2) and (4) to generate more accurate link volumes and link/path travel times to solve the above optimization problem iteratively.

3. SOLUTION TECHNIQUE

3.1 The COSMO Approach Introduction
The proposed COSMO approach as shown in Figure 1 works as follows: The freight network represents the physical network used for multimodal freight transport, i.e. the transportation network including road network and rail network as well as their interactions. Traffic data including link states, network incident and disruption information, and expected passenger traffic demands are fed into the simulation models that are developed to describe the characteristics of the actual freight transportation network. The predicted future freight network states from the running of the simulation models with the traffic data inputs and the candidate routing decision are used to update the states of the links and network. The link states include traffic volumes, travel time, congestion status, fuel consumption, pollution emission etc. Then the optimization block finds the new candidate routes that can reduce the most total cost. If one of the stopping criteria is not satisfied, the load balancing block redistributes the freight flows from the current used routes to the new found routes to reduce the total cost. This load balancing operation leads to new freight network states, updated link states and possible new minimum cost routes. Therefore this iterative procedure continues until one of following stopping criteria is satisfied: 1) The maximum number of iterations is reached; 2) The change of the total cost is less than a predefined value between two consecutive iterations. Once the stopping criterion is satisfied the final solution (i.e. the routing results) are given to the actual freight network to implement. We next discuss the approach of Figure 1 in detail.

Let \( \varphi_{i,j,k}^p (X) \) be the predicted average cost per container of \( x_{i,j}^p (k) \) in a routing decision \( X \) from the updated link states based on the simulation outputs where \( X \) is a routing decision from all the origins to the destinations for the entire analysis horizon, i.e.

\[
X = \left[ x_{i,j}^p (k), \forall i \in I, j \in J, p \in P_{i,j}, k \in K \right]'
\]  

Since the network simulation models can predict time-dependent link volumes and travel times under the constraints of the traffic flow dynamics (2) - (6) and link capacities (11), the original optimization problem can be rewritten as:

\[
\min \ TC (X) = \sum_{k \in K} \sum_{p \in P_{i,j}} \sum_{j \in J} \varphi_{i,j,k}^p (X) x_{i,j}^p (k)
\]  

subject to

\[
\sum_{k \in K} x_{i,j}^p (k) = d_{i,j}, \quad \forall i \in I, \forall j \in J
\]

\[
x_{i,j}^p (k) \geq 0, \quad \forall i \in I, \forall j \in J, \forall p \in P_{i,j}, \forall k \in K
\]

Assume that the cost function \( TC \) is differentiable with respect to the routing decision, the first-order necessary conditions for an optimal solution \( X^* \) of the above problem are:
\[
\sum_{k \in K} \sum_{i \in I} \sum_{j \in J} \sum_{p' \in P_{ij}} \frac{\partial TC(X)}{\partial x_{ij}^p(k)} \bigg|_{X = X^*} \left(x_{ij}^{p'}(k') - x_{ij}^p(k)^*\right) \geq 0,
\]
\[
\text{for } \forall x_{ij}^{p'}(k') \in \Omega
\]

where \(\Omega\) is the feasible solution set given by constraints (15)-(17). The first order derivative in (18) gives the change in the total cost by adding one more unit of containers on path \(p'\) with a departure time of \(k'\) from origin \(i'\) to destination \(j'\), i.e. the marginal cost of a route. The conditions (18) mean that the total cost cannot be reduced further by changing the optimal solution \(X^*\) to another solution locally.

The conditions (18) also state that at the optimal solution, the marginal route costs of any used paths connecting the same OD pair are equal or less than the marginal cost of any other unused paths connecting this OD pair. Otherwise, there exists another solution such that the total cost can be reduced further. In other words, by redistributing some containers from the routes having greater marginal route costs to other routes having smaller marginal path costs, the total cost may be reduced. The idea of the proposed COSMO approach is to iteratively search the routes with most reduced cost then conduct the load balancing from the current used routes to the new found routes that could reduce the most total cost. The overall steps of the COSMO approach can be described as follows:

**Step 1: Obtain an initial solution**

Set the iteration counter \(m = 0\). Assign the given freight demands to a subset of predefined routes in the transportation network, and obtain an initial routing decision \(X^{(0)}\). As an example, the predefined route for each OD pair can be the route with minimum shipping cost without adding the freight demands. \(P_{i,j}(k)^{(0)}\) is the set of the used routes in the initial decision at time \(k\).

**Step 2: Update the link states**

Set the current routing decision \(X^{(m)}\) into the freight network simulation models and run the simulation models to obtain updated link volumes, travel times as well as other link states.

**Step 3: Search for new minimum route**

With the updated link states, find new time-dependent routes with the minimum marginal cost for the OD pairs. Then determine the candidate route that can reduce the most total cost by evaluating the minimum cost routes of all OD pairs. Let \(p_r\) be a new found route with the most reduced cost for an OD pair \((i, j)\) and its departure time be \(k_r\).

\[
P_{i,j}(k)^{(m+1)} = P_{i,j}(k)^{(m)} \cup p_r \text{ if } k_r = k
\]

**Step 4: Check for Convergence**

Check whether the convergence criteria is satisfied, stop the algorithm if the cost reduction between two consecutive iterations is less than a predefined threshold value or the maximum number of iterations has been achieved. Otherwise, go to the load balancing step 5.

**Step 5: Perform load balancing**

For the OD pair with the new minimum route, conduct load balancing by redistributing the freight loads from the current used routes connecting this OD pair to the new route with the minimum marginal cost.

Considering the fact that it is difficult to find the explicit functional form of \(TC(X)\), the marginal costs of the different routes cannot be computed directly. A possible way to find the route marginal cost is via running simulation models by adding one unit of container to the objective route and checking the change in total cost. However this is impracticable for a large scale
transportation network considering the fact the amount of possible routes grows exponentially with respect to the network size. An alternative way is to estimate the route marginal costs with the updated link states from the simulation models. We next describe the estimation model of the marginal route cost from the simulation results and how to find the route with most reduced cost.

3.2 Finding the Minimum Route using Simulation Models

Assume we have a current routing decision \( X \) and its corresponding \( \varphi_{i,j,k}^p (X) \) can be obtained from running the simulation models. Then, the marginal costs of the different routes can be computed by the following equation:

$$
\frac{\partial TC(X)}{\partial x_{i,j}^p (k')} = \frac{\partial \sum_{k \in K} \sum_{i \in I} \sum_{j \in J} \sum_{p \in P_{i,j}} \varphi_{i,j,k}^p (X) x_{i,j}^p (k)}{\partial x_{i,j}^p (k')}
$$

$$
= C_{i,j}^p (k') + K_p T_{i,j}^p (k') + \sum_{k \in K} \sum_{i \in I} \sum_{j \in J} \sum_{p \in P_{i,j}} K_p x_{i,j}^p (k) \frac{\partial T_{i,j}^p (k)}{\partial x_{i,j}^p (k')}
$$

Equation (20) shows the change of total cost if \( x_{i,j}^p (k') \) is changed by one unit of container. The first term in (20) is the non-travel time cost from the vehicle usage, distance cost, etc. which is available directly for a given route without running the simulation models and the second term is the cost of travel time of route \( p' \) with a departure time of \( k' \) that can be computed using model (5) and (6) with the predicted link volumes and travel times that are the outputs of the simulation models. Thus, the values of these two terms are obtained from the route information and the outputs of the simulation models. The third term describes the change in the total travel time cost when changing one unit of container on path \( p' \) with a departure time of \( k' \) from origin \( i' \) to \( j' \), which is difficult to mathematically express directly since the travel time of a given route is a complicated function of the link traffic volumes on this route and other network factors due to the nonlinear dynamical characteristics and route interactions in the traffic network.

From the simulation outputs of the current routing decisions, we obtain the link states \( y_i (k), w_i (k) \), and path travel time \( T_{i,j}^p (k) \) for \( \forall i \in I, \forall j \in J, \forall p \in P_{i,j}, \forall k \in K' \). By the derivative chain rule and equations (4)-(6),

$$
\frac{\partial T_{i,j}^p (k)}{\partial x_{i,j}^p (k')} = \sum_{n_p=1}^{N_p} \frac{\partial w_{i,j,p}^p (e_{i,j,p}^p (k))}{\partial x_{i,j}^p (k')}
$$

$$
= \sum_{n_p=1}^{N_p} \frac{\partial w_{i,j,p}^p (e_{i,j,p}^p (k))}{\partial y_{i,j,p}^p (e_{i,j,p}^p (k))} \frac{\partial y_{i,j,p}^p (e_{i,j,p}^p (k))}{\partial x_{i,j}^p (k')}
$$

for \( \forall i' \in I, \forall j' \in J, \forall p' \in P_{i,j}, \forall k' \in K \)

The term \( \frac{\partial w_{i,j}^p}{\partial y_{i,j}^p} \) in (21) is the derivative of the link travel time with respect to the link volume. It can be approximately determined using the simulated link traffic volumes \( y_i (k) \) and the link capacities \( u_i (k) \). The calculations of the derivative of the link travel time term for road links and railway links are different due to their different characteristics. .

For the road links, the travel time derivative can be obtained using the fundamental
diagram of traffic flow (29) with the observed link volume and average travel time. The travel time
derivative can also be determined using a road travel time model such as the Bureau of Public
Roads (BPR) function in (30) or other estimated functions in (31). Take the BPR function as an
example,

\[ w_i = w_{i,\text{free}} \times \left( 1 + \alpha \left( \frac{y_i}{u_i} \right)^\beta \right) \]  

(22)

where \( w_i \) is the link travel time, \( w_{i,\text{free}} \) is the link free-flow travel time that is determined by the link
length and speed limit, \( y_i \) is the link volume and \( u_i \) is the link capacity. \( \alpha \geq 0, \beta \geq 0 \) are model
parameters that can be estimated from historical data. With this function, the link travel time
derivative in (21) for a road link can be computed by the following equation,

\[ \frac{\partial w_i}{\partial y_i}(k) = \left[ \frac{\alpha \beta w_{i,\text{free}}}{u_i(k)} \left( \frac{y_i(k)}{u_i(k)} \right)^{\beta-1} \right] \]  

for \( \forall l \in L, \forall k \in K \)  

(23)

For the railway links, considering the impact of the passenger train schedule and the
freight dispatching decisions, the travel time of a link is not an explicit function of link volumes so
the corresponding travel time derivative cannot be estimated easily. Therefore the travel time
derivatives of the rail links are estimated from running the rail simulation models repeatedly or
using historical operational data.

\[ \frac{\partial y_{l, p, \alpha, \beta}}{\partial x_{l, j}^p}(k') \text{ in (21) describes the derivative of the traffic volume of link} \]
\[ l_{p, \alpha} \text{ at time } e_{l_{p, \alpha}}(k) \text{ when } x_{l, j}^p(k') \text{ changes by one unit of container}. \] Ignoring the link
interactions, we can estimate this term using the following simple equation,

\[ \frac{\partial y_{l, p, \alpha, \beta}}{\partial x_{l, j}^p}(k') \approx \begin{cases} \frac{1}{\Delta t}, & l_{p, \alpha} = l_{p', \alpha'} \text{ and } e_{l_{p, \alpha}}(k) = e_{l_{p', \alpha'}}(k') \\ 0, & \text{otherwise} \end{cases} \]  

for \( \forall i' \in I, \forall j' \in J, \forall p' \in P_{i', j'}, \forall k' \in K \)  

(24)

Finally, the marginal costs of a route by (20)-(24) can be approximately computed by,

\[ \frac{\partial TC(X)}{\partial x_{l, j}^p}(k') \approx \sum_{\alpha, \beta} \left( c_{l_{p, \alpha}} \left( e_{l_{p, \alpha}}(k') \right) + \kappa_{l_{p, \alpha}} \left( e_{l_{p, \alpha}}(k') \right) \right) + \frac{z_{l_{p, \alpha}} \left( e_{l_{p, \alpha}}(k') \right)}{\Delta t} \frac{\partial w_{l_{p, \alpha}}}{\partial e_{l_{p, \alpha}}}(k') \]  

for \( \forall i' \in I, \forall j' \in J, \forall p' \in P_{i', j'}, \forall k' \in K \)  

(25)

where \( c_{l_{p, \alpha}} \left( e_{l_{p, \alpha}}(k') \right) \) is the non-travel time cost (unit: dollar) of link \( l_{p, \alpha} \) at time \( e_{l_{p, \alpha}}(k') \) 
generated by vehicle usage, distance cost, etc. \( \kappa_{l_{p, \alpha}} \) is the value of travel time on link \( l_{p, \alpha} \). 
\( z_{l_{p, \alpha}} \left( e_{l_{p, \alpha}}(k') \right) \) is the total number of containers on link \( l_{p, \alpha} \) at time \( e_{l_{p, \alpha}}(k') \). All the data
required in (25) can be obtained directly or computed approximately from the simulation model.

Then the marginal costs of a route is the sum of the time-dependent link costs if a link cost
is set as,
\[
c_i(k') + \kappa_i w_i(k') + \kappa_j \frac{z_j(k')}{\Delta t} \nabla y_i(k')
\]
where \(c_i(k')\) is the non-travel time cost (unit: dollar) of link \(l\) at time \(k'\) generated by vehicle usage, distance cost, etc. Therefore the problem of finding the routes with minimum marginal costs can be converted into an elementary time-dependent shortest path problem where the link costs are set as in (26). The time-dependent shortest path algorithms in references (8-11) can be applied to find the shortest routes with minimum marginal cost for all OD pairs. Then, the \(p_i\) in equation (19) (i.e. the route with the most reduced) cost can be determined.

### 3.3 Load Balancing Algorithm
For an OD pair where a new route was found, the load balancing algorithm redistributes the freight loads among \(P_{i,j}^{(m+1)}\) containing the current used routes and the new route based on the marginal costs for the OD pairs with the new found routes. One possible way to do the loading balancing is moving the freight loads between two routes iteratively until the marginal costs of all the routes are equal as done in (28). However this load balancing algorithm faces a slow convergence problem for large demand sizes or large network sizes. Here we propose a load balancing algorithm with quicker convergence based on solving a linear programming problem using an auxiliary routing solution \(X_{\text{Aux}}^{(m)}\),

\[
\min \ TC\left(X_{\text{Aux}}^{(m)}\right)
\]

where \(p_{i,j,k}^{\text{Aux}}\left(X_{\text{Aux}}^{(m)}\right) = p_{i,j,k}^{\text{Aux}}\left(X_{\text{Aux}}^{(m)}\right)\) for \(i \in I, j \in J, \forall p \in P_{i,j}, \forall k \in K\)

subject to
\[
\sum_{k} \sum_{p \in P_{i,j}^{(m+1)}} x_{i,j}^{p}(k) = d_{i,j}, \quad \text{for } i \in I, j \in I
\]
\[
x_{i,j}^{p}(k) \geq 0, \quad \text{for } i \in I, j \in J, \forall p \in P_{i,j}^{(m+1)}, \forall k \in K
\]
given \(d_{i,j}\) for \(i \in I, j \in J\)

Then the new routing decision \(X^{(m+1)}\) can be generated using a step size method \(\alpha_m \in [0,1]\),

\[
X^{(m+1)} = X^{(m)} + \alpha_m \left(X_{\text{Aux}}^{(m)} - X^{(m)}\right)
\]

The most widely applied method of step size selection is the Method of Successive Averages (MSA) in which the step size is selected as \(1/(m+1)\) as in the user equilibrium algorithms such as in (19,21). Although MSA works well for small scale networks, its convergence becomes slow for large networks. In this paper, we decide the optimal step size by solving the following optimization problem (32) in which the total cost of a new possible routing decision is evaluated by running the simulation model. Due to the fact it is time consuming to evaluate all the possible step sizes in the feasible set, we build a discrete set of candidate step sizes to be evaluated in each iteration.

\[
\alpha_m = \arg \min_{\alpha_m \in [0,1]} TC\left(X^{(m)} + \alpha_m \left(X_{\text{Aux}}^{(m)} - X^{(m)}\right)\right)
\]

We next compare three different step size selection algorithms (i.e., enumeration as in Abadi et al. (28), MSA, and the optimization approach using (32)) on a simple example.
three possible routes connecting one OD pair whose characteristics and conditions during three
time intervals (one-hour each in length) are shown in Table 1.

TABLE 1 Route Characteristics and Traffic Conditions of Simple Example

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>Route</th>
<th>Length (mile)</th>
<th>Capacity (veh/hour)</th>
<th>Current Demand (veh/hour)</th>
<th>Travel Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>12</td>
<td>1100</td>
<td>1200</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>10</td>
<td>1000</td>
<td>1000</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11</td>
<td>1050</td>
<td>1500</td>
<td>31</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>12</td>
<td>1100</td>
<td>1000</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>10</td>
<td>1000</td>
<td>950</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11</td>
<td>1050</td>
<td>1100</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>12</td>
<td>1100</td>
<td>800</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>10</td>
<td>1000</td>
<td>900</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11</td>
<td>1050</td>
<td>700</td>
<td>15</td>
</tr>
</tbody>
</table>

The number of containers between this OD pair is 1200 and the total cost is the sum of the
travel times in minutes of all the vehicles. Figure 2 shows the convergence of the three step size
algorithms. The x-axis is the iteration number and the y-axis is the total cost for that iteration. The
required numbers of iterations to stop for the three algorithms are about 750 for the enumeration
algorithm of Abadi et al., 20 for the MSA algorithm, and 6 for the optimal step size algorithm.
Therefore, the optimal step size method provides the best convergence speed although the three
algorithms find the nearly same reduced cost.

FIGURE 2 Performances of different load balancing algorithms
a) Enumeration Algorithm in Abadi et al. (25); b) MSA Algorithm; c) Optimal Step Size
Algorithm
4. SIMULATION MODELS AND CASE STUDY

4.1 Simulation Models

The simulation models used in the COSMO approach of this paper consist of a macroscopic road network model and a rail simulation model. In this paper we use the macroscopic traffic simulator VISUM to develop the road network model to achieve fast network state predictions computationally. The simulator parameters such as lane number, length and capacity etc. of the links are configured based on a practical transportation network. The inputs including passenger and freight traffic for the road network are expressed as number of trips between zones that are origins and destinations within the road network. The passenger traffic data will be filled into the network based on available historical data obtained from the Southern California Association of Governments.

For the rail simulator, we use the railway simulation system of Lu et al. in (32) which was developed based on the ARENA simulation software. The rail simulator is used to evaluate the dynamical train movements for a complex rail network. The track network is divided into different segments based on their speed limits, length, and locations. Then, an abstract track graph is constructed with these segments. The inputs for the rail simulator are the passenger and freight train schedules including their planned departure times, origin stations, and destinations. Then the train movements in the track network are simulated to calculate the travel times and delays of all involved trains.

The integration of the two models has been realized by sharing the OD demands and simulation outputs. The road network simulator sends the freight traffic that will be delivered through trains to the rail simulator. Then, the rail simulator creates the freight train schedule based on the train capacity and simulates the train movements with the planned passenger trains together to output the predicted train arrival times. After receiving the outputs of the rail simulator, the road network simulator will generate necessary truck flows to dispatch containers from the rail stations to their final destinations.

4.2 Case Study and Results

We evaluated the routing between six main destinations (D1 – D6) and three terminals (A, B, C) in the region with different demands as shown in Figure 3. We assume that there are five trains with homogenous capacities of 50 containers between the port terminals and two rail stations nearby the destinations. The average weight of all the containers is assumed to be 25 tons and transportation costs per unit (price/ton-mile) are assumed to be 8 cents for road and 3 cents for railway (34). The time values of a road link and a rail link are set to be 40 dollars and 100 dollars per hour respectively. The values of travel time are estimated based on truck/train usage costs, driver salaries and possible loading/unloading costs. The demands of the six destinations are provided in Table 2. Three shippers communicate their load demands to a coordinator who runs the COSMO approach to generate routes for their demands by minimizing the overall cost.
Three traffic conditions are evaluated including normal traffic, congested traffic and traffic with road closure. In the normal traffic, the passenger traffic is set as the average daily volumes obtained from the Southern California Association of Governments while in the congested condition the passenger traffic is increased by 50% above the average daily volumes. In the third case, the lane closures are introduced at two locations on the main freeways I-710 and I-110 causing the capacities of the two freeway segments to reduce by a half. Figure 4 shows the average cost of transferring all the containers from their origins to the assigned destinations via the multimodal transportation network for the three traffic cases (normal traffic, congested traffic, and congested traffic with road closure). As shown in the figure, the final solutions after using the COSMO approach reduces the average costs compared to the initial solutions in which the minimum cost route before adding the freight is selected as the route to transfer the containers for each OD pair. The main reason for the cost improvements is the proposed approach reduces the congestions of the main freeway bottlenecks by balancing the loads across the network. Moreover,
the COSMO approach provides a better optimization performance during congested traffic and when a lane closure on the freeway occurs than normal traffic conditions.

\[ \text{FIGURE 4 Evaluation of the results of the three traffic conditions} \]

5. CONCLUSION

In this paper, we formulated a multimodal freight routing problem and then developed a COSMO approach that can lead to more efficient decisions in freight routing by using simulation models in the state estimations of complex and dynamic transportation networks. A load balancing methodology that optimizes routing decisions from an overall system perspective is proposed based on estimations of route marginal costs from the simulation outputs. The performance of the proposed approach is demonstrated with an example in the Los Angeles/Long Beach area. The computational results showed that the effectiveness of the proposed approach in reducing the average costs under different traffic conditions. The proposed methodology relies on the estimation model of route marginal costs. As future work, the performance of the estimation model under different traffic conditions will be evaluated and the scalability of the proposed approach will also be analyzed.

6. ACKNOWLEDGEMENT

The research reported in this paper was partially supported by the National Science Foundation under grant 1545130.
REFERENCES


Zhao, Ioannou, Dessouky


