National Transportation Center

Project ID: NTC2016-MU-R-01

URBAN TRANSPORTATION SYSTEM ANALYTICS AND OPTIMIZATION: A SENSOR DATA-DRIVEN APPROACH

Final Report

by

George F. List
gflist@ncsu.edu
Jeremy Addison
Elizabeth Byrom
919-515-8038
Department of Civil, Construction, and Environmental Engineering
North Carolina State University

And

Xuesong Zhou
xzhou74@asu.edu
Tony Liu
School of Sustainable Engineering and the Built Environment
Ira A. Fulton Schools of Engineering
Arizona State University

for

National Transportation Center at Maryland (NTC@Maryland)
1124 Glenn Martin Hall
University of Maryland
College Park, MD 20742

July 2018

ACKNOWLEDGEMENTS

The research team is deeply appreciative of the collaboration between the research groups at NC State University and Arizona State University. It helped shape and focus the effort. A debt of gratitude is also owed to the late colleague Dr. Mark Turnquist with whom Dr. List collaborated for more than a decade. In addition, the material builds on the work of current and former graduate students: especially Yi Chen, Dr. Tzu-Li Wu, and Dr. Hyunwoo Cho. The project was funded by the National Transportation Center @ Maryland (NTC@Maryland), one of the five National Centers that were selected in this nationwide competition, by the Office of the Assistant Secretary for Research and Technology (OST-R), U.S. Department of Transportation (US DOT).

DISCLAIMER

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

Sensor data, from vehicles and facilities, is revolutionizing how urban transportation systems operate. Pre-trip route choices can be informed by network status, en-route path choices can be predicated on evolving conditions, prices can influence path choice decisions, and more robust network operating conditions can be obtained. Careful placement of sensing equipment can enhance system observability in and controllability.

This study has explored new and creative ways to use sensor data to 1) enhance freight-related path choice, both pre-trip and en-route, and 2) improve the performance of urban networks more generally from a freight perspective. While a significant body of literature exists on both path choice and traffic assignment, this study presents new and creative ways to address these topics predicated on real-time data and a freight-first mentality. These new methods can lead to better freight-focused routing decisions and network operating conditions whose performance for freight is improved and can be assessed statistically.

Two specific research objectives have been targeted. The first is creation of new data-driven, truck-oriented path choice algorithms. The second is a data-informed, freight-focused traffic assignment model. Both these efforts have produced results that can enhance freight flows and at the same time mitigate congestion. The efforts build on previous research efforts in which the authors were involved plus findings from projects in which they have collaborated.

The path choice problem has been addressed using algorithms that deal with multiple objectives. One finds the k-shortest paths based on cost and risk, as illustrations of two objectives that often surface as being important in truck-related path choice. Another identifies routes on a probabilistic basis, with each route having a likelihood of being selected for use. The third explicitly finds the non-dominated multi-objective set of paths for any origin-destination (*OD*) pair. This third algorithm makes it possible for decision makers to look explicitly at the options available and select the path that seems to achieve the best compromise among the objectives.

The traffic assignment model uses pricing strategies to encourage the choice of “desired” paths. The network prices specifically facilitate truck flows. The prices focus on improving the quality of the truck trips; encouraging trucks to use facilities (freeways, arterials, etc.) that have high-quality performance. The pricing strategies are thus, multi-class vectors, with different prices by vehicle class. Moreover, with an interest in improving network performance robustness, the pricing strategies endeavor to reduce volume-to-capacity v/c ratios on the arcs. This improves network resilience; that is, the network’s ability to deal with unforeseen (and unpreventable) conditions caused by incidents, bad weather, unanticipated maintenance work, special events, etc. This resilience is critical to freight. It reduces the likelihood that door-to-door travel times will change dramatically if the network conditions that unfold are not exactly consistent with those used in making path choices.

The report is organized as follows. Section 1 provides an overview of the study and introduces basic concepts. Section 2 reviews pertinent literature. Section 3 presents the new, data-driven path choice algorithms that have been developed. There are three, one predicated on probabilistic assignment; a second developed from *k*-shortest path concepts; and a third based on multi-criteria path choice. Section 4 presents the new traffic assignment formulation that puts an emphasis on routing trucks first. It incorporates the new path choice algorithms from Section 3 and makes truck-focused alterations to how the traffic assignment problem is solved. Section 5 presents a study of network settings where Braess' paradox is operative. In such conditions, the addition of a new network link, and its associated capacity, makes the network's performance worse, not better. Seemingly counter-intuitive, these situations, while perhaps uncommon, can arise. If they do, pricing is a very important tool in redirecting the traffic assignment solutions toward more useful, better solutions. Section 6 takes an in-depth look at the use of real-time information to update the status of a transportation network and determine how it should best be operated. Section 7 summarizes the effort and identifies opportunities for future work.

1.0 INTRODUCTION

Sensor data, from vehicles and facilities, is revolutionizing the ways in which urban transportation systems operate. Pre-trip route choices are being informed by network status, en-route path choices are similarly being informed, prices are influencing path choice decisions, and more robust network operations are being obtained.

1.1 STUDY DESCRIPTION

The study has explored new and creative ways to use sensor data to 1) enhance freight-related path choice decision making, both pre-trip and en-route, and 2) improve freight-focused traffic assignment. Despite a significant body of literature that already focuses on both these topics, the study team perceived that new and creative ways were still needed to address freight-focused path choice predicated on statistical analysis of historical and real-time data.

Two specific research objectives have been targeted. The first is development of new, truck-oriented path choice algorithms. The second is incorporation of one or more of those algorithms into a new traffic assignment procedure that places first and primary priority on freight trips. Both these efforts have aimed to find ways of enhancing freight flows and at the same time mitigate congestion.

The path choice work has explored both pre-trip and en-route decision making. Truck routing decisions have always been predicated on network operating conditions. But, historically, the reliance has been on “hunches”, radio shows, and driver-shared information. However, those sources are being supplemented by real-time data from sensing systems. Since many freight trips involve multiple stops and can span entire days, network conditions change as trips progress. Hence, on a rolling horizon basis, tour plans need revision based on new information about the network’s status.

The path choice problem has been addressed using algorithms that deal with multiple objectives. One finds the *k*-shortest paths based on cost and risk, as illustrations of two objectives that often surface as being important in truck-related path choice. Another identifies routes on a probabilistic basis, with each route having a likelihood of being selected for use. The third explicitly finds the non-dominated multi-objective set of paths for any origin-destination (*OD*) pair. This third algorithm makes it possible for decision makers to look explicitly at the options available and select the path that seems to achieve the best compromise among the objectives.

The traffic assignment model uses pricing strategies to encourage the choice of “desired” paths. The network prices specifically facilitate truck flows. The prices focus on improving the quality of the truck trips; encouraging trucks to use facilities (freeways, arterials, etc.) that have high-quality performance. The pricing strategies are thus, multi-class vectors, with different prices by vehicle class. Moreover, with an interest in improving network performance robustness, the pricing strategies endeavor to reduce volume-to-capacity *v/c* ratios on the arcs. This improves network resilience; that is, the network’s ability to deal with unforeseen (and unpreventable) conditions caused by incidents, bad weather, unanticipated maintenance work, special events, etc. This resilience is critical to freight. It reduces the likelihood that door-to-door travel times will change dramatically if the network conditions that unfold are not exactly consistent with those used in making path choices.

As freight service providers know, high *v/c* ratios can lead to unstable operating conditions and high likelihoods of breakdowns and severe delays. Hence, truck-focused traffic assignment algorithms should focus on reducing these ratios. If they do, the network can more easily and readily accommodate variations in the traffic loading and respond more effectively to transients. Wu, List, and Adler (1999) considered single class, single time interval robust optimal (RO) formulations in which an additional objective was considered: minimizing the maximum *v/c* ratio among all or a select set of links. Cho (2006) then applied these ideas to stochastic networks. In both cases it was found that such a minimax objective when used in combination with the UO and SO objectives leads to more robust traffic assignment strategies that are better prepared to deal with surges in demand, incidents, other unanticipated events, and the daily stochasticity associated with emergent demands.

The network planning model combines freight-focused path choice algorithm with the freight-oriented pricing strategies. The combined model can explore the conjoint impacts of these two freight-focused network performance strategies. The resulting impacts on network performance are assessed.

Thus, the outcomes from the project are:

- An advance at the frontier of truck routing through new techniques that capitalizes on the sensor data becoming available from probes and instrumented infrastructure.
- An advance in network pricing strategies through a scheme that focuses explicitly on creating more robust networks that are more tolerant of unexpected heavy load conditions.
- An advance in network pricing strategies that focus explicitly on freight efficiency and productivity.

1.2 SETTING THE CONTEXT

As a prelude to presenting the study results, it seems useful to describe how truck-focused path choice decisions are made and how planning models typically treat trucks when doing traffic assignment.

1.2.1 Multiple Criteria Decision Making

Multiple criteria decision-making (MCDM) is the process of making choices when there are conflicting objectives. It is very relevant here because of the focus on truck-related path generation and traffic assignment. MCDM arises in many contexts including traffic assignment.

There are many ways to formulate multiple-objective decision-making algorithms. The best formulation depends on nature of the choices to be made, the objectives employed, the data available, and the extent to which the objective function values, themselves are important. As Zeleny (1982) indicates, the problem can be solved such that the objectives are considered separately, or they can be combined into (reduced to) a single objective through a set of weights.

Moreover, there are two variable spaces to track in a MCDM problem, not one. The first is the n-dimensional space of the choice variables. It is the vector of choice variables that determines the solution obtained and the objective function values. The second is the m-dimensional space of the objective functions. Like a consumer report assessment, or a "radar plot" in Excel, this space allows the analyst to compare the solutions in terms of the combinations of objective function values that they possess. One of those solutions is "best" in that it has the "optimal" combination of objective function values. But, that "optimal" combination is not necessarily obtained or ascertained by numerically combining the objective function values in some fashion (as with weights).

In addition, the non-dominated solutions are the ones that matter. In the m-dimensional multi-objective space, as illustrated in Figure 1.1, they are the ones for which no other solution has a better set of objective function values. To move from one non-dominated solution to another, a tradeoff is required where the "optimality" of one objective must be diminished to increase the "optimality" of another. As Zeleny (1982) indicates, this non-dominated surface is the set of Pareto optimal solutions.

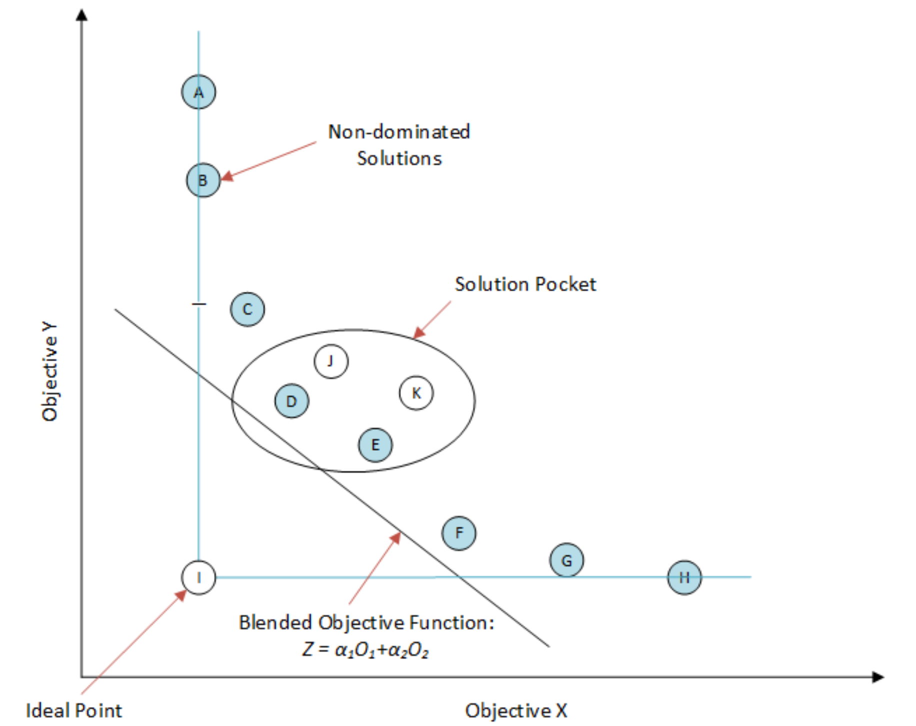


Figure 1.1: The objective space and its attributes

Moreover, since there are multiple objectives, there is an "ideal point" where all the objectives are at their optimal values. (In a single objective setting, this is the lowest/highest/best value for the objective that can be obtained. The quality of solutions can be measured in terms of their distances from this ideal point. A variety of norms can be used. The L1 and L2 norms are common (a linearly combined sum and the Euclidean distance), as illustrated by Zeleny (1982). But higher order norms up to and including the L∞ norm are useful. In the latter case, only the largest objective function deviation from the ideal point value is important.

Normalizing the objective function values is often, also important. This means ascertaining the range of values that are possible (probable) and normalizing those values so that the minimum is 0 and the maximum is 1. Then, when the analyst says two objectives are of equal importance, equal weight values (e.g., both 0.5 or both 1) really does result in equal weight between the objectives. Otherwise, the objective function values also contribute to the relative numerical importance of the objectives.

All these ideas are important because both the truck-focused path choice problem and the traffic assignment problem are inherently multi-objective in nature. Single objective formulations are possible, but they mask the fact that there are important tradeoffs among the objectives.

1.2.2 Freight-Focused Path Choice

Path choice has always been focused on finding the "best" way to go from "A" (the origin) to "B" (the destination). The definition of "best" varies, however; and finding the "best" path depends on what data are available.

For trucks, path choice is more complicated than for autos. And, "best" can have many different interpretations. It can mean avoiding the police; ducking around truck inspection stations; avoiding facilities that charge tolls; or detouring around intersections and interchanges that have tight geometries. Cost tends to be more important than either distance or time. Moreover, truck path choice tends to be multi-objective in nature. Trucks want to find a path that strikes an "optimal" balance among the objectives.

For trucks, too, using all the links in the network is not an option. In a network like the one shown in Figure 1.2, some links may not be available, as in the link from S to U. A sub-network must be used in finding paths. Government agencies often guide, or control, the path choices that trucks have by prohibiting the use of some links. New York, for example, has a freight route network (see New York City, 2018) that trucks must use. These are not routes as in the sense of paths, but rather links in the highway network that trucks are encouraged (or required) to use, except for local deliveries.

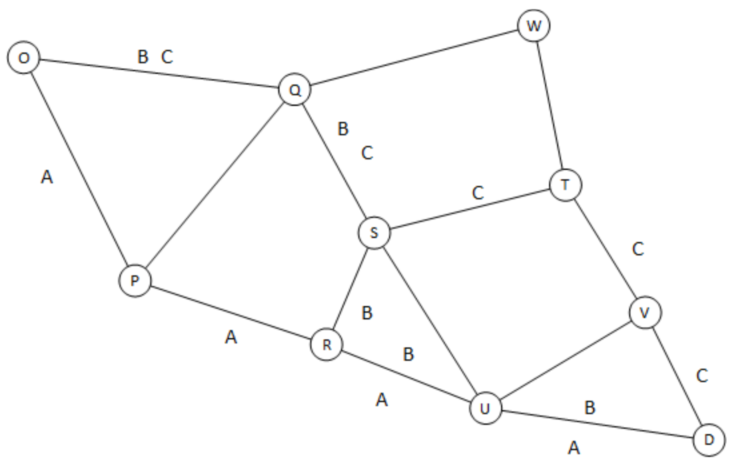


Figure 1.2: A hypothetical network with three paths, A, B, and C

Government agencies can also place restrictions on the types of trucks that can use certain facilities (above and beyond just being a truck). These restrictions most often pertain to attributes like class, length, width, height, weight, and cargo carried (explosives or hazardous cargoes). New York City (2018), for example, prohibits trucks longer than 48' to use the city street network, even though much of the current truck fleet has a length of 53'. Some states, like California (2018), place separate speed limits on trucks. Rarely, agencies will place time-of-day restrictions on trucks (Port Authority of New York and New Jersey (2018). People who engage in "city logistics" research are, in part, trying to find ways for trucks to accomplish their pick-up and delivery tasks, including traversing the network, using routes that are effective and inexpensive.

These aspects of the truck environment all must be considered in creating truck-focused paths. For hazardous materials trucks, for example, multiple objectives must be used because of the importance of population exposure to risk (see, for example, List *et al*., 1991). Time-of-day variations in the network operating conditions may also be important, since truck routing is far more about developing tours than it is just finding paths from A to B (see, for example, Nozick *et al*., 1997).

Because freight-focused path choice is inherently multi-objective, it is important to have a path-building procedure that identifies multiple paths, with different levels of achievement among the objectives. The paths identified must be non-dominated (Pareto optimal) and they should span the space of possible objective combinations. These ideas were described in List *et al*, (1991).

This project presents three ways to generate these paths. The first treats the problem as being explicitly multi-objective. The objectives are placed in lexicographic order and then an extended version of Dijkstra's (1959) algorithm is employed to identify solutions. A drawback to this procedure is that the only paths identified lie on the non-dominated (Pareto optimal) frontier. Other options lying just behind that frontier are not found. The latter paths are helpful because they do not involve making significant changes to the objective function weights. Their relative importance can remain unchanged. The second method uses a multi-objective *k*-shortest path algorithm. Each successive path has a larger objective function value, but the objective weights are unchanged. An extended version of Dijkstra's (1959) algorithm is again employed. The third method finds paths using Dial's (1971) probabilistic assignment algorithm. It not only finds candidate, multi-objective paths, but it also provides a suggested proportional sharing of the traffic among those paths.

One problem with the path choice problem is that the travel times are sensitive to the traffic flow rate. The BPR function (Vythoulkas, 1990) suggests that:

$$t = t_0(1+\alpha * (f / C)^{\beta} \qquad (1\text{-}1)$$

where t is the travel time, t_0 is the free-flow (no-flow) travel time, f is the flow rate and C is the capacity. The values for α and β depend on conditions, but typical values are $\alpha = 0.15$ and $\beta = 4$.

This illustrates a way in which a data-driven approach to path choice can add value. Rather than working with the free-flow travel times, or some other vector of static values, cumulative distribution functions (CDFs) for observed traffic management channel (TMC)-based travel times (or travel rates) can be employed. From a trucking firm's perspective, these can be obtained from public sources such as the NPMRDS data set (see Federal Highway Administration, 2013) or from commercial vendors (e.g., PC-Miler, 2018).C The values can be for specific time intervals (e.g., every 5 minutes) or more aggregate. In the disaggregate case, stitched TMC-based travel times can be developed by synthesizing spatial-temporal trajectories through the network (see Chase *et al.* 2012).

1.2.3 Freight-Focused Traffic Assignment

Traffic assignment is the process of assigning trips to a transport network. Most notably, here, it is part of the process used by transportation planners to assess the performance of system networks. And, in a world of automated vehicles, it will be the way in which paths are chosen. (Unlike today, where people make those real-time choices.) Most often, researchers think about traffic assignment in the context of urban highway networks, but the idea pertains to any type of transport system. The traffic assignment process finds "the best" way for trips to traverse a capacitated network, across space and time, from origins to destinations. The idea of "best" focuses on optimizing one or more measures of network (system) performance.

For example, the triangle regional model (TRM), used by the Raleigh, NC metropolitan planning organizations (MPOs), does traffic assignment using a "user equilibrium" objective. The AM and PM peaks are the primary focus, but other time periods are also studied (see for example, Capital

Area Metropolitan Planning Organization, 2018). Figure 1.2, for example, shows the assignment of single unit trucks (SUTs) to the Raleigh area network for the AM peak hour in 2010.

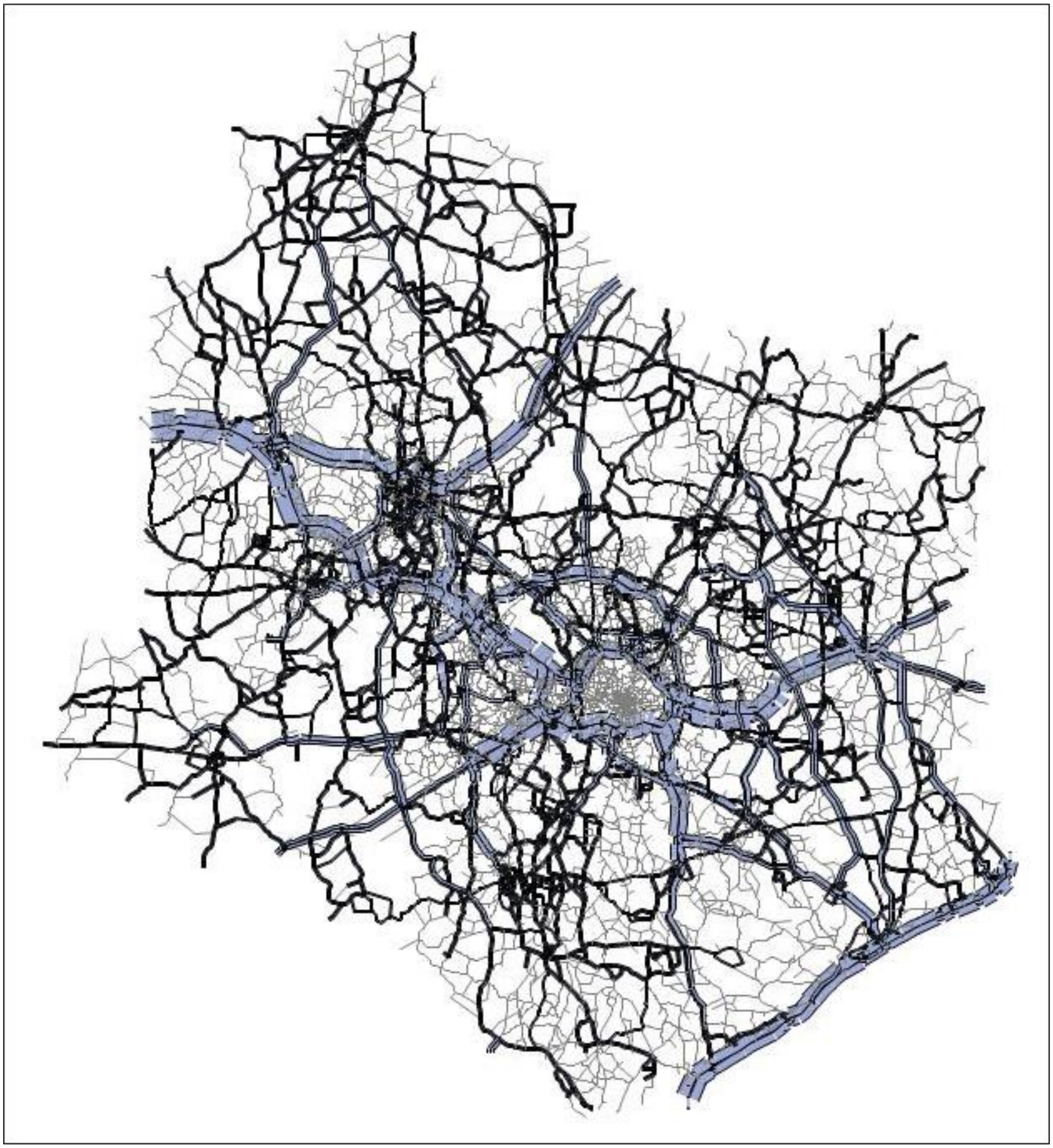


Figure 1.3: Assignment of Single Unit Trucks (SUTs) to the Raleigh, NC metropolitan area network in the AM peak for 2010. Source: List (2017).

The figure shows how traffic passes over the network. The widths of the lines indicate the relative flow rates. Two lines are shown for each link, one in each direction.

Traffic assignment helps network planners make good capital (capacity) investment decisions and operational control decisions. It shows which links are congested, where delays occur, and when, and what routes are used. Through alternatives analyses, it indicates where the addition of capacity or the introduction of tolls or some other operational change will help alleviate congestion. Journey to work trips (or home from work trips) are often the primary focus. Freight trips are often of secondary concern, or they are ignored altogether.

Mathematically, the problem can be described as follows. There is a network of nodes and links that form a graph. The links are edges in the graph. The nodes are either junctions where links connect (interchanges and intersections) or places where traffic goes to and from. (These are the origins and destinations, often called centroids or traffic generators.) Combinations of origins and destinations are called *OD* pairs. The links are connections between the nodes. They are categorized by facility type as being freeways, ramps, arterials, local streets, etc.). A typical network model has 10-20 different link types.

Arcs are one-way (directional) links. The traffic flows over them. For example, a local street link that allows flow in both directions has two arcs. The same is true for a freeway link between interchanges. A a one-way street has arcs in only the direction in which flow can take place. The symbol a refers to a specific arc, A is the set of all arcs.

Paths are sequences of arcs. A_p is the set of arcs belonging to path p. Paths exist between nodes; and, the paths that are important (useful) from a traffic assignment perspective are the ones that connect the origins to the destinations. If p is a path (any path) on the network, and P is the set of all paths, then P_{od} is the set of paths pertaining to *OD* pair *od*. For want of a better terminology, P_{od} is the "path portfolio" for *od*.

The notion of a "market" m can be used to designate a specific sub-class of trips traveling across the network. This simplifies the notation when there are different trip types involved, like autos and trucks. A good example is the "market" m for auto travel between O and D, or the "market" for truck travel for that same *OD* pair (or buses, taxis, garbage trucks…). The set M encompasses all these markets. It can have as many members as the number of *OD* pairs multiplied by the number of trip classes. If there are 100 origins/destinations, meaning there are 10,000 *OD* pairs; and there are two classes of trips, autos and trucks; then there are 20,000 markets. (The flow for some of these markets may be zero.) With this terminology, P_m is the set of paths that pertain to market m. Also, A_m is the set of arcs that lie on the paths in P_m. That is:

$$A_m = \bigcup_{p \in P_m} A_p \tag{1-2}$$

Flows (traffic) in market m must use one or more of the paths in P_m. Path choice algorithms determine which p values in P_m are used for each market and what flow rates f_p pertain.

Often, market flows are time varying. That is, f_p becomes f_{pk}, a flow rate vector which is disaggregate across time intervals k as well as paths p. More technically, f_{pk} is the flow rate for trips on path p that start their trips at O destined for D in time interval k. By extension, F_{mk} is the set of flows on all paths for market m in time interval k, F_k is the set of all flows in time interval k for all markets, and $\boldsymbol{F}$ is the superset of flows for all paths and time intervals. (The traffic assignment algorithm's task is to identify the numerical values for $\boldsymbol{F}$.)

Algorithms that solve the traffic assignment problem treat it as a constrained, non-linear optimization (Sheffi, 1985). In that regard, one or more objectives pertain, like minimizing total vehicle miles or hours; or achieving user equilibrium. There are also constraints, like not exceeding the arc capacities.

Some constraints ensure that the solutions are feasible (implementable in the real world). Others ensure that all the traffic that wants to leave *O* destined for *D* for a given market *m* in a time interval *k* gets assigned to one or more of the paths available. Additional equations allow computation of arc flow rates based on the path flow rates. The objective function(s) and the constraints guide the selection of the paths.

A typical objective is to minimize the total cost of moving the traffic for all markets and time intervals. This results in identifying the "system optimal" solution. That is, for $p \in P$, which is the set of all paths, and for which P_{mk} is the set of paths for market *m* in time interval *k*, the system optimal solution is a vector of flows, f_{pk}, that minimizes the total transportation cost, Z_s:

$$Z_s = \sum_k \sum_m \sum_{p \in P_{mk}} f_{pk} c_{pk} \tag{1-3}$$

where c_{pk} is the cost of using path *p* for traffic that leaves the origin in time interval *k*. This is motivated by Wardrop's (1952) first principle.

This objective can also be computed based on the arc flows. Let f_{ak} be the flow on arc *a* in time interval *k*. That flow rate can be computed from the path flow rates as follows:

$$f_{ak} = \sum_p f_{pk} \delta_{apk} \quad or \quad f_{ak} = \sum_{p \in P_{ak}} f_{pk} \tag{1-4}$$

where δ_{apk} is 1 if path *p* traverses arc *a* in time interval *k*. (In the second notation, P_{ak} is the subset of all paths that traverse arc a in time interval *k*.) If this is the case, then Z_s can then be computed as follows:

$$Z_s = \sum_a \sum_{p \in P_{ak}} f_{pk} c_{ak} \tag{1-5}$$

Variants on this objective, with different coefficients but the same structural form, are used to focus on emissions, energy consumption, risk exposure, or other similar performance metrics.

The most common objective is to find the "user equilibrium" assignment. It is motivated by Wardrop's (1952) second principle. In this case, the costs, c_{pk}, on the chosen paths for a given market *m* in time interval *k* must all be minimal and the same:

$$c_{pk} = c^*_{mk} \ \forall \ p,k \ such\ that \ f_{pk} > 0, \quad else \ f_{pk} = 0 \tag{1-6}$$

and:

$$c^*_{mk} = \min_{p \in P_{mk}} (c_{pk}) \tag{1-7}$$

That is, if path *p* is used in time interval *k*, then the cost for using that path for market *m* in time interval *k* must be the same as the cost for all other paths employed; and that path (those paths) must have the minimum cost among all paths. For example, if four paths exist for a specific market and their costs are 15, 22, 13, and 18, only the third path will be used. It has the minimum cost. Alternately, if the traffic is redistributed among the paths, then the costs might be changed to 14, 22, 14, and 18. Then, the first and third paths will both be used.

This stipulation is often written as a variational inequality:

$$f_{pk} * (c_{pk} - c^*_{mk}) = 0 \quad \forall p \in P_{mk} \quad \forall mk \tag{1-8}$$

where c^*_{mk} is defined, as before, by equation (1-7). Equation (1-8) stipulates that if f_{pk} is non-zero, then $(c_{pk} - c^*_{mk})$, must be zero; c_{pk} must match the minimum value among all paths employed. If c_{pk} is greater than the minimum, then the path cannot be used, and the flow on path p in time interval k, f_{pk}, must be zero.

Beckmann *et al*. (1956) identified an ingenious way to write the objective function for the user equilibrium problem. They showed that the following could be used:

$$Z_{ue} = \min \sum_k \sum_a \int_0^{f_{ak}} G_{ak}(x_{ak}) dx_{ak} \tag{1-9}$$

where x_{ak} is a dummy variable used to compute the integral and $G_{ak}(x_{ak})$ is the non-linear function that computes the cost c_{ak} caused by flow rate x_{ak}.

This formulation is not intuitive until the partial derivatives are taken. Those equations establish the KTT equations for the problem, and they happen to define user equilibrium.

There are two problems with the Beckmann formulation. The first is that it cannot accommodate capacity constraints. The second is that it cannot be used to assess deviations from the user optimal solution if equations (1-8) are violated.

A very useful alternative is to employ a gap function:

$$Z_{ue} = \sum_m \sum_k \delta_{mk} (c_{mk} - c^*_{mk})^\gamma \tag{1-10}$$

where δ_{mk} is 1 if c_{mk} is greater than c^*_{mk} and zero otherwise and γ is a positive value (e.g., 2). This is akin to computing a semi-variance. The gap function easily allows assessment of how close a given solution is to matching all the c^*_{mk} values. This means doing tradeoff assessments among objectives is possible.

Another advantage of using (1-10) is that the values of c^*_{mk} can be controlled by the analyst. They can be set to the user equilibrium values, or they can be set to other target travel times that are desired by the analyst. In fact, because (1-10) focuses only on values of c_{mk} that exceed c^*_{mk}, it penalizes poor performance while not punishing "over performance". Unlike a variance calculation, values of c_{mk} that are lower than c^*_{mk} do not contribute to the value of the objective function.

The single sided variance calculation in (1-10) can be restated using three equations:

$$c^*_{mk} \le c_{mk} \quad \forall\, m,k \tag{1-11}$$

$$\Delta_{mk} \ge c_{mk} - c^*_{mk} \quad \forall\, m,k \tag{1-12}$$

$$Z_{ue} = \sum_m \sum_k \Delta_{mk}{}^{\gamma} \tag{1-13}$$

In this form, (1-10) can be incorporated into a math programming statement of the problem.

A third objective, which is not often used, but very relevant here, focuses on maximizing network resilience. That network attribute is very important to the trucking community. The objective is to minimize the extent to which volume-to-capacity *v/c* ratios η_{ak} (really, flow-to-capacity ratios), exceed target values η^*_{ak}. The η^*_{ak} are desired upper bounds, set by policy. The η^*_{ak} values may be the same for all arcs and time intervals; but they may also be different. Lower η^*_{ak} values are nominally better, because they provide more reserve capacity for accommodating unforeseen conditions, but values that are too low distribute the flows on lengthier paths that traverse less used arcs.

The objective function is written as:

$$Z_{nr} = \sum_k \sum_a (\eta_{ak} - \eta^*_{ak})^{\theta} \tag{1-14}$$

The same restatement of (1-14) as in (1-11) through (1-13) can be used to implement this objective in a math programming formulation.

There is a version of (1-14) that sets θ to infinity. Then, only the largest deviation matters, and the objective function becomes a minmax instead of a minsum.

In all the above discussion, the purpose of the traffic assignment model is to identify values for the flow rates $f_{pk}\ \forall p, k$ that produce the optimal solution.

Additional constraints ensure that the flows are consistent with the overall demand for trips that exists for a given market and time interval, F_{mk}:

$$\sum_{p \in P_{mk}} f_{pk} = F_{mk} \quad \forall\ m, k \tag{1-15}$$

Since it is assumed that the time intervals are all the same duration, the fact that flow rates are used instead of volumes is not a problem. If the time intervals are not all the same duration, then (1-15) must be based on volumes (total trips) instead.

There is a question about whether the demands, F_{mk}, can shift in time. That is, can F_{mk} increase in time interval *m* while it decreases, to compensate, in time interval *n*. Most models assume this cannot happen; but some allow it. There is also a question about whether the demands are congestion responsive. For example, if the network is heavily congested, and the *v/c* ratios are high, do the demands decrease? Or, if they are low, does demand increase? Most traffic assignment models assume this does not happen. But, in real life, it appears that it does happen, especially when congestion decreases because of capacity investments. The total traffic grows.

In summary, a typical traffic assignment formulation involves minimizing (1-10), or its equivalent, (1-9), subject to (1-15), (1-4), and (1-1) or its equivalent. This produces the user equilibrium solution. If (1-3) is used instead of (1-10), then the system optimal solution is obtained. And, if these two objectives are considered simultaneously, then tradeoffs between the two objectives can be considered. If additional objectives are of interest, like (1-14), then they can be included as well.

As (1-6) or (1-8) implicitly suggests, more than one path from P_{mk} may be used for a given optimal solution. In fact, most analyses assume multiple paths will be used. Of course, to make that happen, the path portfolio must be identified. More than one path must be available. To make this happen, multiple paths must be identified. A typical way to do this is to use Dial's (1971) algorithm. It not only finds multiple paths but also suggests a distribution of the flow across these paths. Effectively, the path portfolio identified in this instance is like a fuzzy set. Each of the paths in the portfolio has a likelihood of being used. Another way to find the paths is to use a *k*-shortest path procedure (see, for example, Yen, 1971). Then, not only is the shortest path identified, but a portfolio of paths up to the k^{th} shortest one. The "problem" with either one of these options is that they are computationally more intensive than simply finding the shortest path to each destination, as is done by using Dijkstra's (1959) algorithm.

It is also important to note that, as (1-1) indicates and the highway capacity assessments suggest (see, Transportation Research Board, 2016), arc travel times are affected by their flow rates. That is, they increase as their flow rates increase. Moreover, the relationship is non-linear ($\beta > 1$). In fact, when $\beta = 4$, the travel time remains nearly t_{a0} for low-to-medium v/c ratios, and then it increases rapidly as capacity is approached. (This also means that the optimal solution to the traffic assignment problem is very sensitive to small changes in the v/c ratios.)

That equation, restated with the subscripts included is as follows:

$$t_{ak} = t_{a0}(1+\alpha * (f_{ak} / C_a)^{\beta} \qquad (1\text{-}16)$$

These travel times have historically been equivocated with "cost" of traversing the arc. In that sense, they are the manifestations of function $G_{ak}(x_{ak})$ included in (1-9). As before, t_{ak} is the travel time on arc a in time interval k, t_{a0} is the free-flow (no-flow) travel time, f_{ak} is the flow rate on arc a in time interval k, and C_a is the capacity of arc a (which may be time varying, but that specificity is not shown in the equation).

The implication of this dependence is that there is a circular dependence between identifying the path portfolio and the traffic flow assignments. Changing the traffic assignment changes the travel times which in turn change the path portfolios. So, using Dial's (1971) algorithm or a *k*-shortest path algorithm produces path portfolios whose members are dependent upon the travel times on the arcs. And those travel times are, in turn, dependent upon the path portfolios employed.

Most commercial planning packages use an iterative technique to generate the path portfolios. An initial seed vector of travel times is assumed. These may be the free-flow travel times. Shortest path trees are then built for every origin (to every destination). In math programming terminology, this iterative approach is a "column building" procedure that creates new options (choice variables) to include in the path portfolios for the markets. Traffic is assigned to the network using these trees and then the travel times on the arcs are updated. This process repeats iteratively in that new trees are built and traffic is reassigned based on the portfolio of paths that have been created. Paths no longer in use (from prior trees) are set aside and replaced with paths from newer trees. The process stops when the tree building step fails to suggest that there are new, yet unused paths to add to the path portfolios (for each market) that would allow the travel times on the used paths to be better equilibrated.

This study presents a traffic assignment method which is explicitly data-driven and multi-stage. The objectives presented in (1-2), (1-10), and (1-14) are all employed. The trucks are routed first based on the arc attributes that are important in their multi-objective decision-making. Then the network status is updated to reflect the impacts of those path choices and the remaining traffic (autos) is incrementally added (superimposed) on that solution.

1.2.4 Braess' Paradox

Braess' (1969) paradox is a phenomenon in which adding a new link (arc) to a network, along with its capacity, degrades the network's performance. It is relevant here because pricing can both confound network performance and improve it.

Braess (1969) said: "For each point of a road network, let there be given the number of cars starting from it, and the destination of the cars. Under these conditions, one wishes to estimate the distribution of traffic flow. Whether one street is preferable to another depends not only on the quality of the road, but also on the density of the flow. If every driver takes the path that looks most favorable to him, the resultant running times need not be minimal. Furthermore, it is indicated by an example that an extension of the road network may cause a redistribution of the traffic that result in longer individual running times."

This seemingly impossible situation can arise if the network has a specific structure. The network shown in Figure 1.3, for example, has this property. If the link shown with a dashed line between R and S is created (opened for use), under the case where achieving user equilibrium is the objective, the travelers will choose path ORSD even though the costs of using paths ORD and OSD would be lower, and equal.

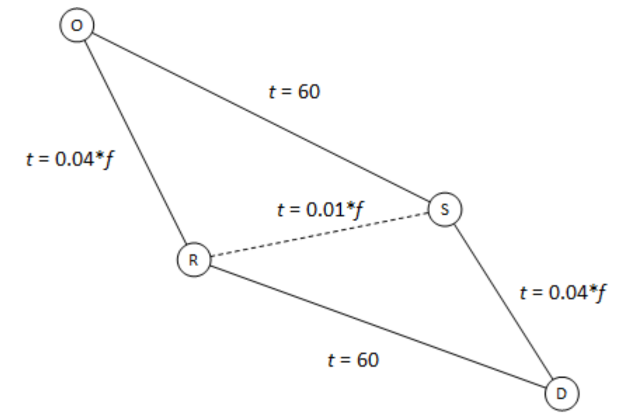


Figure 1.4: A setting where Braess' paradox arises

Braess (1969) focused on the maximum travel time among all the paths with nonzero traffic volumes. For a single O-D pair, calculating the maximum travel time is relatively easy. It is sufficient to compare the travel times on the paths that have nonzero travel volumes and select the largest value. A more widely used measure is the average travel time (same as the total). He stated that both measures are fine and either can be used to detect the paradox. The reason why the average is more popular is because it is related to the system optimal condition. Braess' paradox is usually used to draw a contrast between the system optimal and user equilibrium conditions.

Braess (1969) obtained the steady state traffic assignments with and without the connector under the user equilibrium condition. He assumed that every traveler in the network was selfish and thus individually sought the best route with the smallest travel time. This selfish behavior leads to a Nash equilibrium and that equilibrium is not necessarily system optimal. This is the reason why the paradox happens. He did not apply capacity constraints to the network. Any path could have any amount of traffic on it. This assumption is not realistic; but it makes the problem simple. He first obtains the system performance when there is no connector. Then he recalculates the system performance with the connector (without a capacity constraint). Finally, he determines whether paradox happens by comparing the performance for these two conditions. (He does not explore situations where the capacity is limited.) The relevance of this phenomenon, from a traffic assignment perspective, is that sometimes network performance can be improved by “removing” links from the network (or not adding them; or imposing tolls to control their use. That is, if an appropriate toll is employed on arc RS, the negative impacts of the additional capacity can be mitigated; and, potentially, the network performance can be improved despite the possibility of poorer performance.

Nearly all the literature focusing on Braess’ (1969) paradox assumes that the network is operating under user equilibrium, which means that all of the travelers in the network are selfish and focus on minimizing their own travel times. This assumption makes sense since the paradox is caused by this selfish behavior. But in this study, it is assumed that the travelers are both selfish and altruistic. This assumption is not only more realistic; but it is also valuable when the introduction of route guidance devices is considered. For vehicles equipped with such devices, it will be possible to centrally distribute guidance information and make them more cooperative. This will result in a mixture of system optimal and user optimal drivers. By allowing for this condition, it is possible to study the effects of the percentage of altruistic drivers in the traffic stream. The third group of parameters is about this penetration level.

1.2.5 Real-Time System Management

Management of transportation system performance in real-time is always challenging. Capitalizing on the increasingly rich spectrum of data from system monitors is critical to achieve desired goals. Section 6 describes ways to do this for a transit system The relevance is that scheduled carriers are as challenged, or even more so, than freight service providers in ensuring that their vehicles operate according to a desired plan (timetable). One challenge is to identify an appropriate representation of the “state” of the system so that meaningful, defensible actions can be taken. Section 6 provides guidance about how this can be done.

1.3 REPORT OVERVIEW

The remainder of this report is organized as follows. Section 2 reviews pertinent literature. Section 3 presents the new, data-driven path choice algorithms that have been developed. There are two, one predicated on *k*-shortest path ideas; the other on probabilistic path choice principles. Section 4 presents the new traffic assignment formulation that puts an emphasis on routing trucks first. It incorporates the new path choice algorithms from Section 3 and makes truck-focused alterations to how the traffic assignment problem is solved. Section 5 presents a cameo on the challenging setting where Braess’ paradox is operative. In such settings, the addition of a new network link, and its associated capacity, makes the network’s performance worse, not better. Seemingly counter-intuitive, these situations, while perhaps pathological, can arise. Pricing is a very important tool in redirecting the traffic assignment solutions toward more useful, better solutions. Section 6 takes an in-depth look at the use of real-time information about the status of a network to determine how it should best be operated. Section 7 summarizes the effort and identifies opportunities for future work.

2.0 PERTINENT RESEARCH

This section reviews prior research efforts that have focused on topics that are the same as or like the one addressed here. A prior project report, see List *et al*. (2015), provided a more comprehensive review of the literature focused on path choice.

2.1 SENSOR PLACEMENT AND DATA VALUE

Much of the study is predicated on the hypothesis that data from network sensors can enhance and better inform the decision-making that takes place. This pertains to both the path selection process and traffic assignment.

Zhou and List (2010) have demonstrated that careful sensor placement can enhance the estimation of Origin-Destination (OD) matrices. Ma, Smith, and Zhou (2015) have shown that an agent-based approach can enhance individual user real-time decision-making. Lei, List and Taylor (2015) have illustrated how probe data can be used to characterize corridor-level travel time distributions. Mahmoudi and Zhou (2015) have shown that real-time data can improve freight-based vehicle routing choices. Chen, Zhou, and List (2011) have demonstrated that time-varying tolls can be used to improve truck arrival patterns at port facilities. Cetin, List and Zhou (2005) have explored the number of probes required to develop credible network travel time estimates. And Eisenman and List (2004) explored the ways in which probe data can be used to enhance trip matrix estimation.

2.2 FREIGHT-FOCUSED PATH CHOICE

Chen *et al.* (2011) used the travel time budget as the metric for determining reliability-based user equilibrium (RUE) rather than the typical metric of expected travel time. This allows for inclusion of differing degrees of risk-aversion for distinct user classes evaluated in the system. A column generation technique was used to enumerate over all OD and user class pairs followed by solving the associated restricted subproblems by iteratively shifting flows from the costliest path with maximum travel time budgets to the cheapest path having minimum budget. The cost itself is established using a mean budget (M-B) dominance condition described within the paper.

The simplest routing algorithms find the route from an origin *O* to a destination *D* which minimizes a single measure. The measure must be additive across the links. While many generalizations of these algorithms are available, (e.g., for single objectives see Shier, 1976; for multicriteria problems see Henig, 1985 and Mirchandani and Wiecek, 1989; and for capacitated problems see Handler and Zang, 1980), the adaptation to truck-focused routing problems has been rudimentary. For example, in the context of HazMat shipments, Brogan and Cashwell (1985) use a shortest path algorithm based on a single metric, which must be user specified, to assign routes for specific OD pairs.

Practitioners selecting routes for truck shipments of HazMat are encouraged by the Federal Highway Administration to use the methodology developed by Barber and Hilderbrand (1980). It calls for selecting minimum risk routes based on the following process:

- "eliminate routes with physical mandatory factors" (i.e., physical or legal restraints on use);
- "consider legal and political implications of trying to change legal mandatory factors and exclude or reserve judgement accordingly";

- "select route(s) with smallest risk values" (incident probability on a link times the expected consequence of an incident summed over all links in the route);
- "apply subjective factors (such as proximity to hospitals, nursing homes), if unable to differentiate on risk." (Barber and Hildebrand, pp.14-15).

The methodology does not explicitly consider the trade-off between risks and cost. Neither does it consider the distribution of risk resulting from such all-or-nothing traffic assignment; that is, risk equity is ignored.

Applications of this methodology follow it closely, except for the evaluation of risk. Studies such as Glickman and Rosenfield (1984), Ivancie (1984), and Kessler (1986) use population exposure as a substitute for the more detailed risk evaluation outlined by Barber and Hildebrand (1980). Risk for an individual link is typically defined as the adjacent population (within some prescribed bandwidth) multiplied by the link length. These link-based "risks" are summed to determine the overall "risk" for a route, and the route with the lowest "risk" is selected. While this approach may be simple and straightforward, perhaps a reflection of data limitations and/or the cost of acquiring missing pieces of information, it has conceptual faults. First, differences in accident rates, by link or facility type, are ignored. Second, the entire population within the band is assumed to be equally at risk. Third, the basis for adding the link risks to obtain route risks is not clarified and may not be appropriate when two or more links affect the same population zone. Lastly, the differences in risk due to differences in material properties (e.g., incident likelihoods and material release probabilities) are ignored.

Batta and Chiu (1988) present two single-objective shortest path formulations for multi-objective truck routing. The first formulation does not consider different accident probabilities for the network links, while the second recognizes this fact. In both formulations the size of the population potentially impacted by an accidental release of hazardous materials is also part of the criterion. In addition, both formulations recognize that the network nodes are higher risk points than the links, and incorporate this concept by assigning penalties to the nodes of the transportation network.

Shobrys (1981) is among the earliest efforts to deal explicitly with multiple objectives. Shobrys considers two objectives: (1) minimize ton-miles traveled and (2) minimize population exposure-tons. He points out that the optimal decisions must come from the Pareto-optimal solution set. By using various weights to combine the two objectives, Shobrys was able to use a hybrid distance-population cost for each link, and hence, use a shortest path algorithm to obtain several Pareto-optimal solutions.

Abkowitz and Cheng (1988) present a bi-objective routing model. Two types of damages are considered: direct and indirect. The former occurs at the incident scene while the latter occur in the surrounding vicinity. The area of exposure for the indirect damages is determined using a formula which accounts for dispersion, horizontally and vertically, as well as wind speed and direction. Unit weights combine the fatalities, injuries and property damages into a single overall measure of risk. This is then traded off against transportation cost to identify Pareto-optimal routes for individual OD pairs.

As should be expected, the minimum cost strategy favors expedience at the expense of safety. It is essentially a deregulated solution. Safety may be enhanced, over randomly chosen routes, in that the distance or time during which accidents can occur may be lower, but this is not an assured outcome. Minimization of accident likelihood increases safety by restricting shipments to routes where accident rates are lower. Minimization of risk proceeds one step further by considering the consequences of accidents as well.

These multi-objective models are based (explicitly or implicitly) on identifying some weighting scheme to make the multiple objectives commensurable. A problem with this approach is that it only identifies a subset of the efficient paths, so at best it provides only a partial set of solutions. The reasons not all paths are found is illustrated in Figure 1.1, using a bi-objective example. Figure 1.1 shows non-dominated solutions to a problem with objective functions X and Y. However, a method using a weighted sum of objectives (i.e. minimize $uX+(1-u)Y$ for values of u, $0 \leq u \leq 1$) would only identify the solutions touched by the line. Other four points that are still non-dominated, called "gap points", are missed. These points may also be of interest; and they should be identified. Algorithmic procedures related to the solution of multicriteria shortest path problems can be found in many places including Cox (1984), and Henig (1985). Direct application of such procedures to multi-criteria settings, such as HazMat, were begun by Cox (1984), who developed a node-labeling method for solving multi-objective routing problems; and began the exploration of relationships between routing and scheduling decisions.

A set of problems is defined, one for each potential departure time, and within these, sub-problems are established, each focusing on a particular "copy" of the network, its attribute values having been sampled from the underlying distributions. A multi-objective shortest path algorithm is then used for each of these to find Pareto-optimal routes (i.e., for each sample network and departure time). These routing/scheduling options are then reviewed by the decision-maker to select the "best" combination.

Zografos and Davis (1989) developed a multi-criteria routing model for HazMat which considers the following criteria: 1) general population at risk, 2) risk of special population categories, (groups of persons having evacuation difficulties), 3) property damage, and 4) travel time as a surrogate for the truck operating cost. According to this formulation, each link of the network is assigned a "cost" vector that characterizes the link in terms of the established routing criteria. The proposed model is formulated as a multicriteria shortest path problem and preemptive goal programming is used to obtain the solution to the problem.

The empirical results obtained by Saccomano and Chan (1985) illustrate the importance of random events such as weather on overall risks along a specified route. Turnquist's (1987) work represents one attempt to deal with this issue, in the context of probability distributions for link attributes. Mirchandani and Soroush (1985) have also studied probabilistic networks. They model each link as having an additive attribute, for example travel time, which is randomly distributed. Such distributions can be convolved to determine the travel time distribution from a given origin to any other node in the network. Measures of "expected cost" for such partial paths can also be developed from these distributions using a cost function, which may be nonlinear. Ultimately, for any OD pair, the optimal path is one which has the lowest total "expected cost." When the cost function is linear, expected costs on the links can be used and the optimal path minimizes the sum of these link expected costs. When the cost function is nonlinear, route cost is no longer the sum of these link costs. For such situations, a terminology is presented that helps indicate the dominance of one path over another. A path between two nodes is "absolutely permanently preferred" over another if the preference order is unaltered for any path segment addition to either of the common end points. "Temporary preference" occurs when some path segment addition reverses the preference order. Finally, a path is "relatively permanently preferred" to another for a given destination node if there is no path segment addition to the given destination node which reverses the preference order. Using these concepts, Mirchandani and Soroush develop a node-labelling algorithm to determine optimal paths when the cost function is exponential or quadratic.

All the models discussed up to this point, both single and multi-criteria, assume that the network links have unlimited capacity. This assumption leaves open the possibility that a small number of network links will carry a large fraction of the truck trips. An immediate result of such a procedure is the assignment of risk to the population residing along these links. Thus, these models fail to capture the objective of the equitable distribution of risk.

Recognizing the importance of the equitable distribution of risk or other population-centric objectives, Zografos and Davis (1989) developed a routing model that considers the equity objective. These models consider the same objectives as previous multi-criteria shortest path formulations and incorporate the equitable distribution of risk by imposing capacity constraints on the network links. The resultant formulation is equivalent to a capacitated assignment problem. Preemptive goal programming is the solution methodology.

Along any route chosen, if public authorities enact curfews (time-of-day restrictions), there exists a scheduling problem to be solved. Cox and Turnquist (1986) developed a model to find the departure time for any given route which minimizes curfew delay en-route. The model assumes that the link travel time is a random variable and a recursive algorithm is presented. An application dealing with the highway transport of spent nuclear fuel is discussed. Conclusions from that example include: 1) the solution with deterministic travel times underestimates the expected delay, 2) the relative advantage of precise dispatching decreases as the uncertainty in link travel time increases, and 3) the optimal departure time becomes earlier as uncertainty in travel time increases.

Chen *et al.* (2013) solved the reliable shortest path problem (RSPP) for α - reliable path wherein the goal is to minimize the travel time budget while ensuring an α level of on-time arrival probability. Link travel times follow normal distributions to allow for an analytical reliability measure when evaluating routes. A stochastic-based dominance condition is described to effectively extend Bellman's Principle of Optimality to the stochastic case. Two algorithms are used to solve the problem, notably a multi-criteria label setting algorithm like Dijkstra's algorithm and the multi-criteria A* algorithm. The A* algorithm inherently favor nodes likely to be on α - reliable paths by assigning higher priority in the search space for these nodes, maintaining a set of eligible non-dominated path sets ordered in descending likelihood of appearing on the α - reliable paths. A case study of Hong Kong shows the computational time advantage of the A* algorithm over the more accurate yet more computationally demanding multi-criterial label setting algorithm.

Huang and Gao (2012) investigated stochastic time-dependent (STD) networks with temporal and spatial correlation among links, using a minimum expected disutility (MED) to evaluate routes. Due to the stochasticity, the problem violates Bellman's Principle of Optimality and the Algorithm CD-Path is designed to find only "pure paths" whose sub-paths are non-dominated. This algorithm iteratively trims the search space of dominated paths until the final solution set only contains non-dominated paths.

Ji, Kim, and Chen (2011) described a simulation-based multi-objective genetic algorithm (SMOGA) that is used to find non-dominant (Pareto optimal) paths. This algorithm consists of a Monte Carlo simulation of correlated travel times, a genetic algorithm to explore the combinatorial solution space, and a Pareto solution filter to maximize the diversity of the Pareto solutions. SMOGA solves the chance constrained multi-objective programming (CCMOP) model for optimal path finding while simultaneously minimizing the travel time budget and satisfying travel time reliability constraints. Numerical experiments on the Chicago Sketch network show feasibility and diversity in explored solution space, while further showing that correlation among link travel times create significant discrepancies in travel time budgets. Without correlation among links, Pareto paths resulting have significant travel time budget bias and provide sub-optimal paths.

Srinivasan *et al.* (2014) solved the most reliable path problem with the added feature of shifted log-normal link travel times (MRP-SLN), a constrained nonlinear integer programming problem. The MRP-SLN algorithm uses lower and upper bounds on path reliability measures to force convergence. A sufficient condition is devised to guarantee that the most reliable path is present in a set of least expected travel time paths. A case study of Chennai city in India was examined, and the generated set of paths contains the true optimum in more than 98% of tested OD pairs with an average relative gap between proposed and true optimum paths less than 0.06%. These sets are generated in under an average of 25 seconds. However, using approximations for normal and lognormal times at link and path level lead to sub-optimal solutions in 14% and 12% of cases respectively, with reliability decreases up to 9%.

Xing and Zhou (2011) sought to answer the most reliable path problem under varying spatial correlation assumptions, with total path travel time variability represented by standard deviation. Lagrangian substitution is used to estimate the lower bound of the most reliable path by solving a sequence of shortest path problems, followed by a subgradient descent to iteratively reduce the optimality condition between primal and dual solutions until a termination condition occurs. Further, when spatial correlation exists among link travel times, a sampling-based solution algorithm is embedded in the above Lagrangian technique. A case study of the Bayshore Freeway between Mountain View and San Jose, California was examined. These experiments showed that utilizing these reformulated models on a large-scale network allows for 10-20 iterations of standard shortest path algorithms to reach duality gaps of about 2-6% for uncorrelated travel times.

Methods for multiobjective routing of hazardous materials shipments have been developed by several authors, including Cox (1984), Turnquist (1987), Abkowitz and Cheng (1988), Zografos and Davis (1989), Current, et al. (1990), and McCord and Leu (1995). Some of these methods use specialized techniques limited to two objectives (e.g., Abkowitz and Cheng, (1988) and Current et al. (1990)), while others are more general.

2.3 TRAFFIC ASSIGNMENT

The single-objective, user equilibrium traffic assignment model has been applied extensively in traffic network design and planning. The objective is a demand-oriented cost function, mainly, equilibrating the travel times for all markets.

Managing the network’s performance involves additional objectives such as congestion, air pollution, noise pollution, and reliability. The use of demand-oriented traffic assignment models has its limitations. From an economic point of view, a traffic network involves supply and demand issues. The traveler's interest represents the demand side of the network and the network's capacity represents the supply side. The goal is to maximize the supply capacity usage while minimizing traveler cost. The lowest cost for the travelers is obtained by solving the user-equilibrium (UE) assignment and the best supply capacity usage can be obtained by minimizing arcs' flow/capacity ratio within the network. These objectives are, in fact, mutually interactive and the traffic network's managerial complexity requires the best trade-offs among these objectives. A good way to deal with these trade-offs is to adopt the multicriteria decision-making theory. The fields of traffic assignment and multicriteria decision-making research are well established independently, but the way they can be effectively integrated in mathematical models for traffic network management is at present the subject of much on-going research.

Dial (1996) and Leurent (1996) both discussed the need to use multiple objectives when solving equilibrium traffic assignment problems. From a demand side perspective, they seek to find traffic assignments which minimize a multi-dimensional measure of generalized user cost. the idea of introducing a manager-optimal objective is an extension of this thought. It points to scenarios in which the best interests of the system's drivers and managers are addressed, as well as society.

This idea of seeking optimal solutions based on a user cost is debatable, however, especially if there are negative side effects insofar as network performance is concerned. For example, user-equilibrium solutions have limitations. They tend to produce high congestion levels on arcs that are part of many paths even under low traffic conditions. No incentive exists to either manage or minimize volume-to-capacity (*v/c*) ratios. When a system-optimal objective is pursued, *v/c* ratios tend to be lower, but there is still no emphasis on managing those ratios, so the quality of flow from arc to arc varies widely. Hence, there is a value in introducing a manager-focused objective, such as minimizing the maximum v/c ratio, as was described in Wu *et al.* (1999).

Also, since there is no guarantee that user equilibrium will be obtained under conditions where that is not the only objective, it is important to consider a fourth objective. This one measures the maximum amount of deviation in travel time among all OD pairs in the network. Ideally, this deviation should always be zero for all OD pairs, but where user equilibrium is not stressed, trips can be assigned to paths with greater-than-minimal cost and a deviation in travel times can result. To address equity issues among the users these deviations should be kept to a minimum.

Based on the above, it seems promising to consider network management decision-making in conditions where four objectives are explicitly involved. These are as follows:

- for an individual driver: to minimize individual trip travel time (i.e., to seek the user-equilibrium solution);
- for all drivers: to minimize the differences in travel times among drivers moving between the same origin-destination pairs (to minimize the discord in what might be a near-optimal user-equilibrium solution);
- from a social/economic perspective: to minimize total vehicle-hours spent in travel (i.e., to seek the system-optimal solution); and
- from a network management perspective: to minimize the maximum arc congestion level (i.e., to seek a solution with the lowest possible v/c ratio).

As stated earlier, commonly applied objectives in traffic assignment are user equilibrium and system optimality. These notions were formally proposed by Wardrop (1952).He stated two principles that formalized these equilibrium concepts. His first principle states that each user non-cooperatively seeks to minimize his or her cost of transportation. Network models that apply this principle are usually referred to as "user equilibrium" (UE) models. Wardrop's second principle says that each user behaves cooperatively in choosing its route to ensure that total system travel time is minimized. Network models that apply this principle are generally termed "system optimal" (SO) models. However, there are many network equilibrium situations that do not fit either the UE or SE framework. Haurie and Marcotte (1985) model a more general situation where the travelers in the network are divided into multiple groups. The travelers within each group all exhibit the same behavior and the different groups compete with each other. Each group of users is called a Cournot–Nash (CN) group of users and the corresponding equilibrium for each group is a Cournot–Nash (CN) equilibrium. A Cournot-Nash user group is altruistic if the group's objective is to minimize the total travel time of all its users. Altruistic groups can aim to minimize the total travel time of only their group members. Multi-equilibrium or mixed equilibrium solutions are those where more than one type of traveler group is present.

It is generally acknowledged that best way to solve traffic assignment problems is to formulate an equivalent mathematical program that has a unique solution. Formulation of the UE as a mathematic programming model was pioneered by Beckmann *et al*. (1956). They showed the conditions under which the user equilibrium condition is obtained using a set of equations and inequalities. By applying the KKT conditions, they give the conditions that ensure existence and uniqueness of the solution. Dafermos and Sparrow (1969) give a complete formulation of both the UE and SO problem and the necessary and sufficient conditions for the existence, uniqueness and stability of the solutions. They also present an algorithm for a solving the two problems.

After the concept of a mix equilibrium or multi-equilibrium was proposed by Haurie and Marcotte (1985), people began studying the math formulation for mixed network equilibrium with different types of users.

Harker (1988) presents a model in which the travelers for each O-D pair obey either the UE or SO behavior principle. Link cost functions are no longer symmetric and the optimization formulation is not obtainable under this assumption. Harker uses variational inequality to formulate and solve the problem. He includes maximum link flow constraints in the model. Van Vuren *et al*. (1989) and Bennett (1993) further investigate the existence of equivalent mathematical programs for deterministic, mixed UE and SO with fixed O-D demand. Bennett (1993) investigates the existence of the math formulation for two mix equilibrium problems. In the first, the network is mixed with user optimal users and Cournot altruistic users and the second network is mixed with user optimal users and Cournot private users. For the first problem Bennett shows that an equilibrium formulation can be found if the link travel time function is of the BPR-type with $n > 1$ but there is no equilibrium formulation for the second problem. Zhang, Yang and Huang (2007) extend the study of mixed equilibria to consider a multi-class multi-criteria mixed equilibrium. They develop a variational inequality model to characterize the multi-class multi-criteria UE–CN mixed equilibrium behavior and they also establish the existence of uniform link tolls supporting such mixed equilibrium as a system optimum.

The capacitated assignment problem for Braess' (1969) paradox has been studied only to a limited extent. The main reason is that the solutions cannot be characterized by the classic Wardrop (1952) equilibrium conditions. They may, however, be characterized in terms of generalized travel cost.

Including upper bounds on the link flows can be done in one of two ways. The first is by using asymptotic travel time functions, i.e. functions that indicate a link's travel time goes to infinity when its flow approaches its upper bound. The second is to introduce side constraints.

The first approach has been shown to be problematic. Boyce et al. (1981) empirically find that asymptotic travel time functions yield extremely high travel times, which result in unrealistic and circuitous trips. In addition, Larsson and Patriksson (1995) point out that this approach is inherently ill conditioned [29].

Since the uncapacitated assignment problem can be solved very efficiently, a natural solution strategy is to transform it into a sequence of uncapacitated problems, tending to one which is equivalent to the original, capacitated, problem. The second method is to solve the corresponding uncapacitated problem with travel time functions adjusted by the corresponding optimal shadow prices. See, for example, Jorgensen (1963), Hearn (1980), and Inouye (1987). Larsson and Patriksson (1995) extend this to the general, convexly constrained traffic equilibrium assignment model [7, 11, 15, and 29].

2.4 BRAESS' PARADOX

Braess' paradox motivates two types of network design problems. The first is the discrete network design problem. It ascertains which arcs to add to the network to improve the value of a specified objective function. The second asks, given a network, which edge should be moved to get the best performance in the Nash equilibrium? Or equivalently, given a network, which sub-network will exhibit the best performance when the users are selfish?

Roughgarden (2004) provides several near-optimal results and approximate algorithms for different network types. He concludes that it is impossible to detect Braess' paradox efficiently [33]. Lin, Roughgarden, Tardos and Walkoverwe (2006) construct an infinite family of two-commodity networks, related to the Fibonacci numbers. They use these networks to establish nearly matching upper and lower bounds for both the price of anarchy with respect to the maximum latency and the worst-possible severity of Braess' Paradox. They also prove that there is no polynomial-time algorithm for specific network design problems [11]. LeBlanc (2004) uses a mix integer programming formulation to implement a branch and bound algorithm [3]. Particular attention is paid to the computational aspect of large scale problems. Poorzahedy and Turnquist (1982) provide two algorithms which save computational work by approximating the original problem with a new formulation which is easier to solve. The first algorithm proposed solves this approximate problem exactly, while the second is more efficient, but provides only a near-optimal solution to the approximate problem [24].

These network design problems motivate a similar thought, which is the capacity allocation problem. IN this case, a network designer wants to allocate capacity to optimize the network performance according to some certain system criteria. This problem is usually solved by a bottleneck analysis, but such analyses are only valid for centrally controlled systems. For decentralized systems and non-cooperative systems, this problem becomes complex and potentially counterintuitive, like Braess' paradox.

Korilis, Lazar and Orda (1995) discuss strategies to improve the performance in non-cooperative networks [18].

1) Network design during the provisioning phase. The problem is formulated as one of allocating additional capacity to an existing network. It is shown that the problem exhibits paradoxical behavior. The authors prove that for a system of parallel links, adding capacity

always improves the network's performance. They give the sufficient conditions for Braess paradox not to occur.

2) Improving the performance during the actual operation stage. This problem is a Stackelberg game. The authors discuss the thresholds of the amount of network flow that is controlled by leader in order to get system optimal solution. Methodologies for upgrading general networks while avoiding the Braess' paradox are also investigated. The results show that the capacity can be expanded throughout the whole network.

Korilis, Lazar and Orda (1997, 1999) discuss the optimal strategy for adding capacity to a system of parallel links. Two strategies are considered. The first is the addition of capacity to a system and the second is the transfer of capacity toward the link with the originally highest capacity. The authors prove that in systems with parallel links, both these strategies improve the performance, which means Braess' paradox will not happen in these cases. The authors use the example of the original Braess' network to show that under non-cooperative conditions, adding capacity to the network, even if it is infinite, can make the performance worse [16,17].

Altman, Azouzi and Pourtallier give some guidance on avoiding the Braess' paradox when upgrading the network. The paper uses the original Braess network but is asymmetric. Three cases are considered:

1) Add capacity to the link in the middle, this will lead to the performance degrading, that is, Braess paradox happens in under this case.
2) Add capacity to all the links.
3) Add a new link to connect the origin and destination. Both the second and third method will not cause the Braess paradox.

The authors mentioned above discuss the network design problem in computer networking, however the results can be applied to other network where Braess' paradox is observed.

The results shown in Blumsack and Ilić (2006) mirror those of Korilis, Lazar and Orda (1997, 1999) in electric power systems:

- If the boundary links represent the bottleneck constraint, then the optimal policies are to either reduce demand or expand capacity on both links.
- If the Wheatstone bridge is viewed as the active system constraint, then the optimal policy is to remove the bridge entirely; or, equip the bridge with fast relays or regulate the power to allow use of the bridge only during contingencies [25].

Masuda and Whang (1999) model the network as a queuing system and investigate the capacity expansion problem for a decentralized system with a general network topology. The authors discuss the problem for both the short run and long run condition. In the short run, they prove that extreme pricing can solve the joint problem of demand and routing. In the long run, whether the capacity expansion strategy is valid depends on the short run optimality [34].

Abbad, Azouzi and Kamili (2006) extend Korilis (1997) by considering the case of elastic traffic. Agents are assumed to have a utility function that is concave in the amount of flow that they route, and their price is defined to be the total cost of passing their traffic over the network minus their utility. The authors prove that prices only decrease when capacity is increased (in networks of parallel links) and that users do have an interest in improving their requests. The authors did not give results for when the objective is to minimize user costs, which is considered in Korilis (1997).

2.5 SUMMARY

This section has reviewed prior research efforts that focused on efforts that are the same as or very similar to the ones addressed in this report. A prior project report, see List *et al*. (2015), provides a more comprehensive review of much of the literature focused on path choice.

Clearly, there is a lot of work that examines routing. Some of it pertains to trucks. Some pertains to the use of real-time information. This study examines the nexus of these two aspects, which makes it somewhat unique.

A lot of studies also focus on traffic assignment. It may be the most-explored topic in all transportation research literature. But, the number of papers that explicitly focus on putting freight (trucks) first is very limited. And, the number that focus on multi-commodity assignment is also limited. Moreover, the number that incorporate capacity constraints into the scope of the problem is also limited. Those are among the important aspects of the problem considered here. So, this work is a contribution to the state-of-the-art in that regard.

Also limited at the number of traffic assignment studies that have focused on network resilience; specifically, managing the *v/c* ratios among the arcs in the network. This idea is very uncommon. But, it is critical, the study team suggests, in the context of freight(truck) routing. It is not so much that the v/c ratios need to be kept low for the truck assignments to be either feasible or optimal. But, rather, that keeping the *v/c* ratios low helps ensure that the network can deal with unforeseen situations that arise, from accidents, incidents, weather, or other situations. Hence, a min max objective is incorporated that endeavors to keep the *v/c* ratios for the arcs below target values.

Also unusual is the use of a gap function to measure the achievement of the user optimal solution. Beckmann’s formulation is used far more often, with an emphasis on satisfying the KTT conditions that their objective function creates. Instead, in this study a gap concept is used where the travel times on the path are compared with a target path time that is desirable. The target might be the travel time (or cost) associated with the user equilibrium solution. Or, it could be some other travel time that is a policy objective from the perspective of the network operator. An advantage to adopting this perspective is that the user optimal objective function value obtained in any solution can be compared against the value of 0 that would be obtained were the user optimal solution actually achieved. The Beckmann objective function cannot be used for this purpose.

The section has set the stage for the work that follows. It shows what has been done and how that body of work relates to the ideas and results presented here.

3.0 DATA-DRIVEN PATH CHOICE

This section addresses the topic of path choice as informed by real-time data. It is especially focused on trucks, which have a complex, multi-objective, long-term perspective on the selection of paths.

The focus on real-time data implies that path choice is informed by evolving network conditions, caused by incidents, weather, work zones and influences. List *et al.* (2017) describe ways to assess the travel time distributions associated with specific operating conditions. Here, it is assumed that suitable travel times, costs, and risks have been established for the links (arcs) in the network and that those values are sensitive to (reflective of) the operating conditions extant at the time the trips would take place.

3.1 PROBABILISTIC PATH CHOICE

One option for dealing with multi-objective routing is to use a probabilistic path choice algorithm. This idea was originally put forward by Dial (1971). It is commonly referred to as Dial's probabilistic assignment algorithm. List (1993) and more recently List (2016) used this procedure to develop path choices for truck flows on urban networks.

3.1.1 Algorithm Description

The algorithm identifies the likelihood that paths might be employed between origin O and destination D based on multiple objectives. If o_1 and o_2 are two objectives of interest, then $z = w_1\, o_1 + w_2\, o_2$ can be a weighted combination of those objectives, and z_p can be the value of z for path p. Also, if P_{od} is the set of paths for OD, then the likelihood that path p will be employed can be based on a multinomial probabilistic choice model:

$$P_p = \frac{e^{-\beta c_p}}{\sum_{k \in P_{od}} e^{-\beta c_k}} \quad \forall\ p \in P_{od} \tag{3-1}$$

The value of β can vary widely. In the case study presented here, a value of $\beta = 1$ has been used, but it should be calibrated based on observed path choices using real-time data. Higher values of β drive the solution toward all-or-nothing assignment where only one path sees use. Lower values of β encourage more paths and distribute the probabilities more broadly. In the limit, as β approaches 0, the path probabilities become the same (because the cost does not matter; any path is OK to choose).

From a practical standpoint, after the algorithm is applied, a lower bound can be imposed on the smallest probability that a path is retained in the path set for a given OD pair. Paths with low probabilities can be removed. The probabilities for the remaining paths are upward adjusted proportional to their initial probability of use so the sum once again totals 100%.

Since the probabilities relate to paths, and those paths are sequences of arcs, the algorithm also indicates the likelihood that specific arcs will be used among the paths for each OD. That is:

$$P_{aod} = \sum_{p \in P_{od} \cap P_a} P_p \quad \forall a \text{ and } \forall\, od \tag{3-2}$$

The intersection between $p \in P_{od}$ and $p \in P_a$ captures the paths that both belong to *OD* and traverse arc *a*. For example, if two paths for a given *OD* use arc *a* (but before and after they use different arc sequences) and their probabilities are 5% and 18%, then the total probability of flow from *O* to *D* using arc *a* is 23%.

The values of P_{aod} are sometimes written as α_{aod} and called arc utilization coefficients, or more simply, arc utilizations. They capture the percentage of flow for *OD* that use arc *a*. These values are used extensively in network assignment algorithms. For example, if f_p is the flow on path *p* and f_a is the flow on arc *a*, then f_a can be computed as:

$$f_a = \sum_{od} \sum_{p \in P_{od}} \alpha_{aod}\, f_p \tag{3-3}$$

The details of the algorithm can be found in Dial *et al*. (xxx). Simplistically, a forward pass is made through the network from the origin to develop likelihoods that a single unit of flow from the origin would traverse any given arc. Then a backward pass is made from each destination back to the origin and the probabilities of arc use are identified. Implicitly, paths are identified; with the likelihood of path use being ascertained based on a proportional sharing of total inbound percentage of flow at each node (from all entering arcs) onto the outbound arcs.

One significant drawback to the procedure is that the objectives are not considered individually. They are treated as terms in a composite objective that uses weights to combine the objective function values. Ex post facto, it is possible to determine the values that pertain to the individual objectives, but those values are produced by the combination of weights employed.

It is possible to "overcome" this deficiency, at least in part, by parametrically varying the values for the weights employed. For example, if the weight for cost is made large and all the other weights are small, then the path that minimizes cost will be identified. If the weights are made equal between cost and risk, or two other objectives, then the path that minimizes that weighted combination of the objectives will be identified. In Figure 1.1, selection of these weights results in a slope to line for the blended objective function. This makes it possible to identify the "convex hull" of the non-dominated surface, but the "gap points" that are still non-dominated but lie within this convex hull are missed.

3.1.2 Case Study Example

An example helps illustrate the results this algorithm produces. In a prior research project, a simplified network for the Albany area of New York was created (List and Turnquist, 1991). It is shown in Figure 3.1. There are 86 nodes and124 links (248 arcs). Each arc (link) has a functional class designation ranging from limited access freeway to local street. (The exact interpretation of the classes is not important, but 1 is for limited access freeways, 2 is for high-quality arterials, etc., and the values range from 1 to 9). Each arc (link) also has a length (in miles), a cost (in nominal dollars) and risk (population exposed within a certain distance). The network was used in the previous research project to identify non-dominated paths for hazardous materials (HazMat) shipments between selected *OD* pairs. The purpose then was to identify a) how many paths existed, b) what were their costs and risks, and c) which ones represented reasonable tradeoffs between cost and risk, from the perspective of society (e.g., the government) and the carrier. (These perspectives might have been different.)

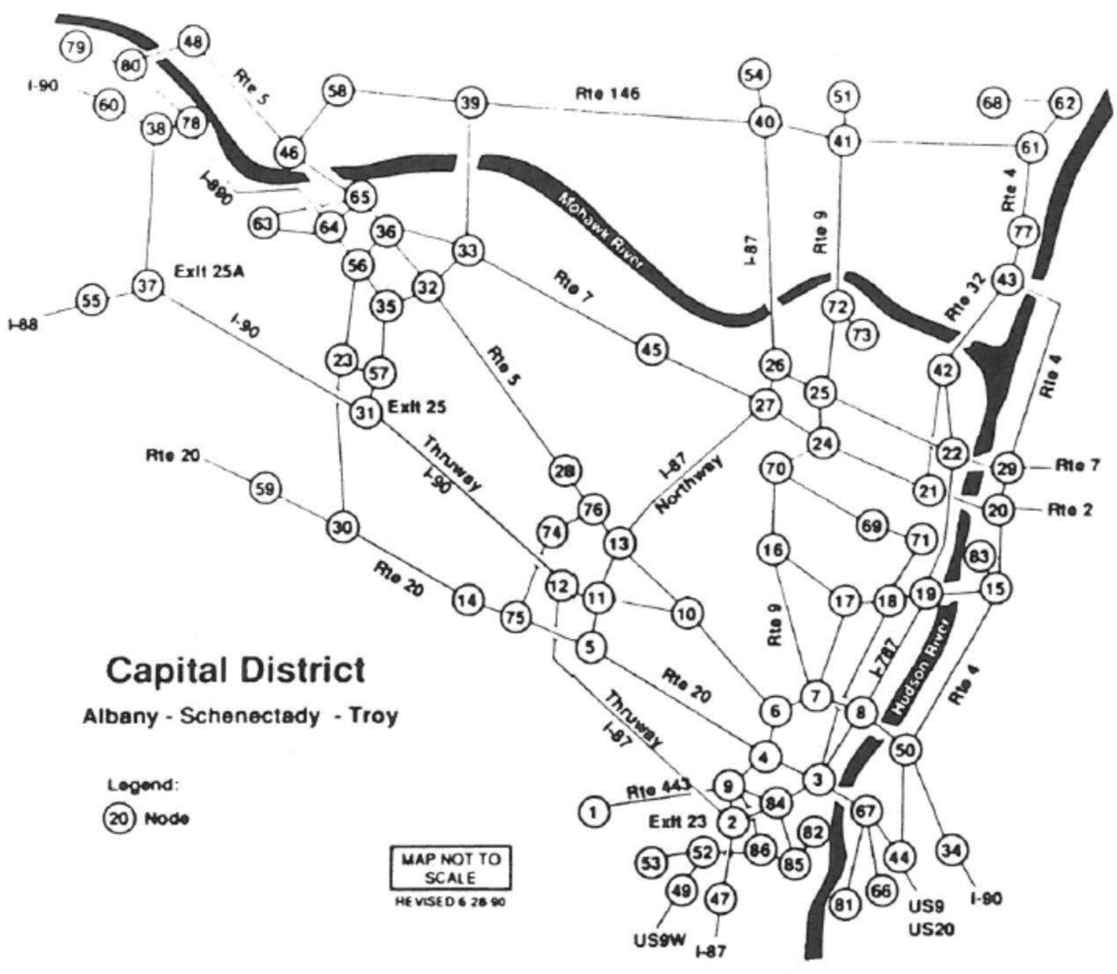


Figure 3.1: Case Study Network. Source: List and Turnquist (1991).

In this instance, the purpose of the analysis is to identify the likelihood that specific arcs will be used for specific *OD* pairs based on 1) a cost equation and 2) a value for β. Three examples are described. In the first, only the cost related to time and distance is considered, as might pertain for auto trip making where there are no tolls on the network. In the second, cost and risk are both considered, as might pertain to a truck where, again, there are no tolls in the network. In the third, there are tolls on the network, and these tolls are considered along with cost and risk, in selecting paths.

Table 3.1 presents the results for the situation where and the weights for cost, risk and tolls are 1.0, 0.0, and 0.0 respectively. The columns are, respectively, the origin, the destination, the arc, a flag indicating whether trucks can use the arc (1) or not (0), the functional class for the arc, and the arc utilization coefficient. A value of uArc = 100 means that 100% of the *OD* flow uses the arc. For example, in the case of *OD* pair (1,13), 100% of the traffic uses arc 1 while only 24% of the traffic uses arc 3; 77% of the traffic uses arcs 12, 17, 27, and 137, while 24% uses arcs 3, 29, 26, and 152. Since these numbers sum to 101 (effectively 100%), the implication is that traffic is split between two paths, about 76.5% on one path and 23.5% on the other.

Table 3.2 shows the results for these same *OD* pairs when the weights are equal for cost and risk. (Tolls are ignored.) The arc utilizations are clearly different from those shown in Table 3.1. The number of arcs employed has changed and the arcs are now exclusively either from class 1 or 6. The use of arcs in class 8 (local streets) has disappeared.

orig	dest	arc	class	tFlag	uArc
1	10	1	6	1	100
1	10	12	8	1	100
1	10	17	6	1	100
1	10	137	8	1	100
1	11	1	6	1	100
1	11	3	1	1	100
1	11	126	1	1	100
1	11	152	1	1	100
1	12	1	6	1	100
1	12	3	1	1	100
1	12	126	1	1	100
1	13	1	6	1	100
1	13	3	1	1	24
1	13	12	8	1	77
1	13	17	6	1	77
1	13	27	6	1	77
1	13	29	1	1	24
1	13	126	1	1	24
1	13	137	8	1	77
1	13	152	1	1	24

orig	dest	arc	class	tFlag	uArc
1	14	1	6	1	100
1	14	11	6	1	100
1	14	15	6	1	100
1	14	137	8	1	100
1	14	159	6	1	100
1	15	1	6	1	100
1	15	7	1	1	51
1	15	8	6	1	50
1	15	22	1	1	51
1	15	24	1	1	100
1	15	43	6	1	50
1	15	134	1	1	100
1	15	160	6	1	100
1	16	1	6	1	100
1	16	12	8	1	100
1	16	16	8	1	100
1	16	20	8	1	100
1	16	137	8	1	100

Table 3.1: Illustrative arc utilizations when only cost is considered

orig	dest	arc	class	tFlag	uArc
1	10	1	6	1	100
1	10	5	1	1	20
1	10	7	1	1	100
1	10	19	1	1	100
1	10	24	1	1	81
1	10	126	1	1	20
1	10	134	1	1	100
1	10	142	1	1	100
1	11	1	6	1	100
1	11	3	1	1	100
1	11	126	1	1	100
1	11	152	1	1	100
1	12	1	6	1	100
1	12	3	1	1	100
1	12	126	1	1	100
1	13	1	6	1	100
1	13	3	1	1	100
1	13	29	1	1	100
1	13	126	1	1	100
1	13	152	1	1	100

orig	dest	arc	class	tFlag	uArc
1	14	1	6	1	100
1	14	3	1	1	100
1	14	15	6	1	100
1	14	126	1	1	100
1	14	138	1	1	100
1	14	152	1	1	100
1	14	159	6	1	100
1	15	1	6	1	100
1	15	5	1	1	20
1	15	7	1	1	100
1	15	22	1	1	100
1	15	24	1	1	81
1	15	126	1	1	20
1	15	134	1	1	100
1	15	160	6	1	100
1	16	1	6	1	100
1	16	5	1	1	20
1	16	7	1	1	100
1	16	20	8	1	100
1	16	24	1	1	81
1	16	126	1	1	20
1	16	134	1	1	100
1	16	142	1	1	100

Table 3.2: Illustrative arc utilizations when cost and risk have equal weights

Table 3.3 shows the results when tolls are included. That is, there are equal weights among cost, risk and tolls. Counter-intuitively, there are tolls on the local streets to keep the vehicles from using them, as might be done to discourage truck traffic. The results have again changed. The values and lists in this table are different from both Table 3.1 and 3.2. The arc utilizations, arc lists, and class choices have changed. With regard to the latter, there are more 6's and one 8.

orig	dest	arc	class	tFlag	uArc
1	10	1	6	1	100
1	10	5	1	1	20
1	10	7	1	1	100
1	10	19	1	1	100
1	10	24	1	1	81
1	10	126	1	1	20
1	10	134	1	1	100
1	10	142	1	1	100
1	11	1	6	1	100
1	11	3	1	1	100
1	11	126	1	1	100
1	11	152	1	1	100

orig	dest	arc	class	tFlag	uArc
1	12	1	6	1	100
1	12	3	1	1	100
1	12	126	1	1	100
1	13	1	6	1	100
1	13	3	1	1	100
1	13	29	1	1	100
1	13	126	1	1	100
1	13	152	1	1	100
1	14	1	6	1	100
1	14	3	1	1	100
1	14	15	6	1	100
1	14	126	1	1	100
1	14	138	1	1	100
1	14	152	1	1	100
1	14	159	6	1	100

orig	dest	arc	class	tFlag	uArc
1	15	1	6	1	100
1	15	5	1	1	20
1	15	7	1	1	100
1	15	22	1	1	100
1	15	24	1	1	81
1	15	126	1	1	20
1	15	134	1	1	100
1	15	160	6	1	100
1	16	1	6	1	100
1	16	5	1	1	20
1	16	7	1	1	100
1	16	20	8	1	100
1	16	24	1	1	81
1	16	126	1	1	20
1	16	134	1	1	100
1	16	142	1	1	100

Table 3.3: Illustrative link utilizations when cost, risk and tolls have equal weights

These three examples show how the weights for the objectives can change the results obtained. The algorithm responds to the combinations of cost, risk and tolls that pertain to the links based on the weights employed.

From a real-time data perspective, this algorithm would be sensitive to the changing travel times, since those affect the costs, and the tolls. If the tolls were time-of-day dependent or dynamic in response to traffic conditions (or incidents or some other consideration), the algorithm would take those changes into account and provide different suggestions for arc utilizations (and, implicitly, classes).

3.2 K-SHORTEST PATHS

A second option for dealing with multi-objective routing is to use a k-shortest path algorithm. In this case, multiple paths are identified for each *OD* pair based on a weighted composite objective function and how many paths to be identified, *K*. (The weighted composite objective function is the same as before. If o_1 and o_2 are two objectives of interest, then a weighted composite objective, z, can be computed as $z = w_1\, o_1 + w_2\, o_2$, and z_p can be the value of z for path p). P_{od} is the set of K paths for *OD* pair *od*, and the set is in ascending order from the one with the best (lowest) value of z_p to the worst (highest).

3.2.1 Algorithm Description

The procedure is an extension of Dijkstra's (1959) algorithm, and it is as follows:

1) For each *OD* pair in the network.
    a. Let n_e be the current node n from which partial paths are being extended. Let r be an alternate label for n_e. Initialize $n_e = r = O$.
    b. Let *aNode* designate the "from" node for any arc. For the arcs emanating from a given node r, this means r is the *aNode*. Let *bNode* be the "to" node and let s be an alternate label for the *bNode* of the arc. Let a be a designation for a specific arc. Let z_a be the composite objective function value for arc a.
    c. Let $P(n_e)$ be the sorted list of paths that are candidates for extension from n_e. Let K be the number of paths in that set (hence K-shortest paths) and let $I(n_e)$ be the index on that set. Let k be the k^{th} member of that indexed set. Set $P(n) = \Phi$, the null set for all nodes n. Let $I(n) = \Phi$ as well, for all n.
    d. Let k_e be the k^{th} shortest path to n_e and the one for which paths are being extended. Set $k_e = 1$. ($n_e = r = O$ was set in 1a above). (There will be no other k values for the origin.)
    e. Let $A(n_e)$ be the arcs for which the *aNode* is n_e. Set $A(n_e) = A(O) = A(r)$. Set $P_{nd} = \Phi$.
2) Create new partial paths that are one-arc extensions of k_e from r. That is, for arcs a that are members of $A(r)$, create new paths that extend to the nodes reachable via $A(r)$ from r.
    a. See if the partial path would lead backwards (back to the node that created the current partial path). If it would, reject it.
    b. Otherwise, compute a composite objective function value for the new partial path. Let it be z_e, being the sum of z_a values leading to s from O for this partial path.
    c. Check z_e against the existing composite objective function values z_k that already exist for s. If z_e is smaller than z_1, then place it first and bump the others up one value of k. Discard

the last one. If it is not first, then identify the k, $k + 1$ pair between which z_e fits. Once this pair is found, including being last, bump the ones from $k + 1$ to $K – 1$ to one-higher value of k. Again, discard the last one.

3) For all the partial paths $P(n)$ that exist among all nodes n
   a. Select the one that has the smallest composite objective function value. Make it the new path for extension.
   b. Set $r = s$, the *bNode* for the arc that produced the partial path and update $A(r)$.
4) If a new path to extend has been identified, then go to step 2), otherwise terminate.

3.2.2 Case Study Example

The algorithm has been applied to the same network shown in Figure 3.1. Various weight combinations have been explored and insights developed.

Figure 3.2 shows the findings for equal weights among cost and risk for *OD* pair (1,3). Since *K* was set to 4, four paths are shown. The numbers in the ovals indicate the sequence. The line from (0,115) to (115,0) shows the locus of points in the bi-objective space for which the composite objective function value is the same. Moving that line outward from the origin (without changing its slope) shows the locus of bi-objective points for which the composite objective function value is the same. Hence, the line touches path 1 first, then 2, 3, and 4. It is admittedly difficult to tell whether the line touches 1 or 2 first, and the same for 3 and 4, but examination of the numerical values for the composite objective function shows that this sequence is correct.

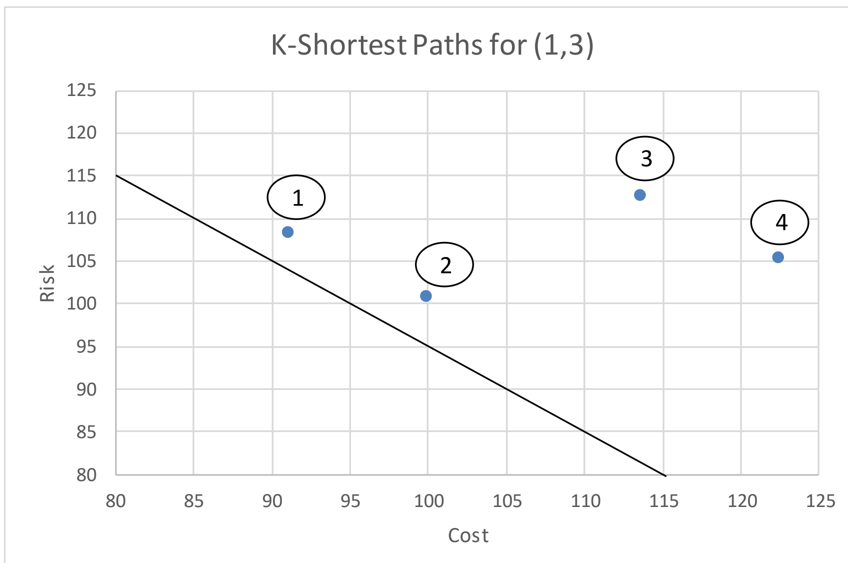


Figure 3.2: K-shortest paths for *OD* pair (1,3) where equal weight is given to cost and risk.

3.3 MULTI-OBJECTIVE SHORTEST PATH

A third option for dealing with multi-objective routing is to use an explicit multi-objective shortest path algorithm. This is somewhat computationally intensive, and it requires care. This section describes a procedure for creating multi-objective shortest paths. It is based on List and Turnquist (1991) and Turnquist (1987).

Addressing the multi-objective shortest path problem explicitly means finding the paths that have non-dominated combinations of the objective function values. This has been illustrated in Figure 1.1

using a bi-objective example. In that figure, paths *A* through *H* lie on the non-dominated surface. No other paths (down and/or to the left) dominate these. But, *J* and *K* are dominated. For them, there are other paths that have better combinations of the objectives. For example, *D* dominates both *J* and *K*. Drawing lines up from D and to the right will show this. *J* will lie within the space created by those lines. *E* also dominates *K*, but not *J*. In general, for paths *A* and *B*, *A* dominates *B* if, for every objective *i*, *A* has a better value than does *B*. If the objectives are all to be minimized, and z_i is the objective function value for objective *i*, then, if $z_i(A) \leq z_i(B)\ \forall\ I$, then *A* dominates *B*. It is useful to think of this as a backwards test that checks every pair of paths. That is, path *B* is dominated if there is *any other path* that dominates it. Hence, every path must be checked against every other path to see if it is dominated or not. In the case of Figure 1.1, it is possible to do the line creations described earlier and see, graphically, that, for paths *A* through *H*, no other path dominates these paths. There are no paths downward and to the left that "shadow" them. For path D, for example, there is no path closer to the origin that has a combination of the two objectives where both are better than the values for *D*. This is true despite the fact that paths *A*, *B*, and *C* have better values for objective 1, and paths *E, F, G*, and *H* have better values for objective 2. But, none of those paths have a combination of objective function values where both $z_1(*) \leq z_1(D)$ and $z_2(*) \leq z_2(D)$ where $*$ is any one of the other paths. Even path *C*, which at first glance might appear to be dominated, is not. If the vertical and horizontal lines are created upward and to the right from paths *B* or *D*, the "shadows", open boxes, that are upward-and-to-the right of these paths do not contain path *C*. Hence, *C* is not dominated by either of these paths (or by any other).

3.3.1 Solution Algorithm

An extension of Dijkstra's algorithm can be created to obtain the non-dominated set of paths from *O* to *D*. The extension has three elements. First, the objective values for each path must be placed in a specified order that is always followed. The order is not important, it can be hierarchical if desired, with the value for the "most important" objective being placed first and the least, last, but that is not necessary. Just, the same order must always be used. To illustrate, if there are objectives z_1, z_2, z_3, and z_4, and the values for each can be from "a" to "z" with "a" being first and "z" last (for reasons that will be apparent momentarily), then the multi-objective "word" for each path is formed by placing the objective function values in a specific sequence [z_1, z_2, z_3, z_4], or $z_1\&z_2\&z_3\&z_4$ where *&* is the concatenation operator. For example, if path *A* has the values a, c, m, and d for objectives 1 through 4 respectively, then the "word" for this path is "acmd".

The second element is to use a lexicographic (word)-sorted list of the paths, just like a dictionary does with words. That is, if there are two paths *A* and *B*, and the multi-objective word for *A* is "acmd" and for *B* it is "bqef", then *A* comes before *B* in the sorted list. By extension, if there are four paths, *A*, *B*, *C*, and *D*, and their multi-objective words are acmd, bqef, acme, and rtuv, then their order is *A*, *C*, *B*, *D*, or acmd, acme, bqef, and rtuv. Moreover, if the objectives are measured numerically instead (which, clearly, is far more common), and paths *A*, *B*, *C*, and *D* have multi-objective words of [25, 15, 23, 79], [23, 10, 15, 18], [25, 15, 23, 78], and [25, 15, 20, 73], then the their lexicographic order is *B*, *D*, *C*, *A*, or [23, 10, 15, 18], [25, 15, 20, 73], [25, 15, 23, 78], and [25, 15, 23, 79].

The third element is to extend the partial paths based on their lexicographic sequence. This is the same as it is for the original Dijkstra's algorithm, where the partial paths are kept in objective function value order, from lowest to highest (or best to worst), and the one with the lowest value is always the next one selected for extension. If the four partial paths above were the candidates for extension, regardless of where they end (what node), then *B* would be selected. Moreover, if that partial path ended at node *G*, then the new, partial paths would be created by extending path *B* to the nodes that can be reached by using the arcs that start at *G*. (Those new paths would then be added to the lexicographic order and placed in their proper location.)

An algorithm that implements this logic has been created using VBA inside an Excel workbook. It finds non-dominated paths for all *OD* pairs within a network. The pseudocode describing the algorithm is as follows:

5) For each *OD* pair in the network.
   a. Let n_e be the current node from which partial paths are being extended. Let r be an alternate label for n_e. Initialize $n_e = r = O$.
   b. Let *aNode* designate the "from" node for any arc. For the arcs emanating from a given node r, this means r is the *aNode*. Let *bNode* be the "to" node and let s be an alternate label for the *bNode* of the arc. Let a be a designation for a specific arc.
   c. Let P_e be the (lexicographic sorted) list of non-permanent paths that are candidates for extension and let p_e be the one selected for extension. Let N_p be the number of partial paths in that set and let I_p be the lexicographic index on that set. Let k be the k^{th} member of that set. Set $P_e = \Phi$, the null set, $I_e = \Phi$, p_e to φ, "null" and $k = 0$.
   d. Let $A(n_e)$ be the set of arcs for which the *aNode* is n_e. Set $A(n_e) = A(O) = A(r)$. Set $P_{nd} = \Phi$.
   e. Let P_{nd} be the set of non-dominated paths (to all nodes in the network starting from O).
6) Create new partial paths that are one-arc extensions from r. That is, for arcs a that are members of $A(r)$, create new paths that extend to the nodes reachable via $A(r)$ from r.
   a. Check to see if the partial path would lead backwards (back to the node that created the current partial path). If it would, reject it.
   b. Otherwise, add the new partial path to P_e.
   c. Place it in its correct spot in the lexicographic order
7) For the partial paths in P_e
   a. Select the one that is first in the sorted order. Make it the new candidate path for extension.
   b. Remove the candidate path from P_e.
   c. Check to see if it is non-dominated
      i. If it leads to a node that has never been visited, treat it as though it was non-dominated and keep it. (It may, in fact, be dominated.)
      ii. Otherwise, check to see if it is dominated by any other path to s that has previously been saved to the list of non-dominated paths, P_{nd}.
      iii. If it is not dominated, then add the new partial path to P_{nd}. (It has already been removed from P_e, so it no longer exists in that list either.)
      iv. If it is dominated, then "discard" it. (It will never be used as the basis for path extension because it is dominated.)
   d. If the partial path is not dominated, and its *bNode* is not D, then select the candidate path as the new one to be extended and set $r = s$, the *bNode* for the arc that produced the partial path and update $A(r)$.
8) If a new path to extend has been identified, then go to step 2), otherwise terminate.

3.3.2 Case Study Example

In the current analysis, the non-dominated paths for cost and risk between every pair of nodes has been identified. Since the number of non-dominated paths can become countably infinite (until all possible non-looping paths are identified), an upper bound was placed on the extent to which the cost and risk values, separately, could be larger than the average values identified in the Dial-based algorithm. For the results presented here, that upper bound was set to 50% higher than the average.

Since there are 86 nodes, there are 7,310 *OD* pairs (86*85) excluding the instances where *O*=*D*. Among those, 16,020 non-dominated paths were identified, or slightly more than 2 for each *OD* pair. The maximum number of non-dominated paths for a specific *OD* pair is 15 for *OD* pair (46,18). This means there are some *OD* pairs for which only one path exists. A total of 128,877 partial paths were saved among all *OD* pairs, or an average of between 17-18 for each *OD* pair. The maximum number saved was 171 for *OD* pair (46,83).

An example non-dominated surface is shown for an *OD* pair (1,13) in Figure 3.2. The cost ranges from 170 to 230 and the risk 150 to 350. (The units are unimportant.) It is "obvious" that paths 1, 2, and 4 are non-dominated. But, paths 3 and 5 are also non-dominated. Careful examination of their numerical values for cost and risk show that neither paths 2 or 4 dominate 3; and path 4 does not dominate path 5.

Similar results could be displayed for all other *OD* pairs, but the figures would not contain new revelations. The same trend would be seen. The paths stretch in a rough crescent from top left to bottom right, and there are some paths, like 3 and 5 that appear to be dominated; but, in fact, the numerical values would show they are not.

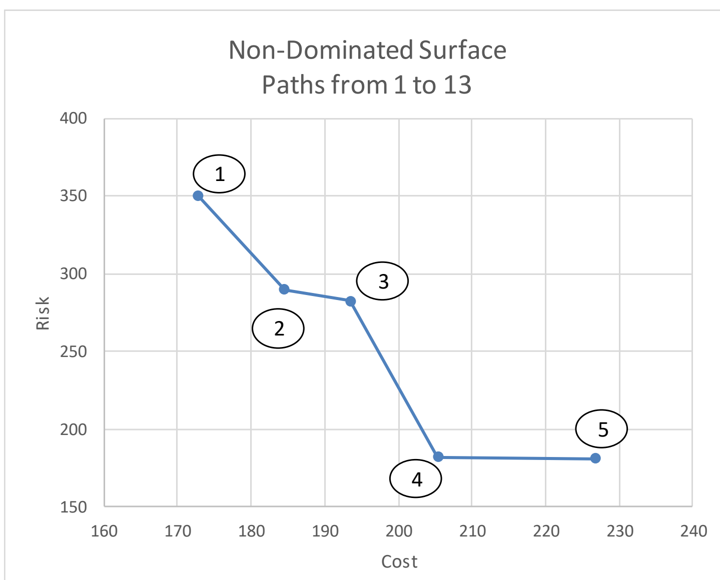


Figure 3.3: Non-dominated paths for *OD* pair (1,13)

3.4 SUMMARY

This section has focused on data-driven path choices, especially for trucks (freight). As a topic, path choice has been explored extensively. It may be the most comprehensively explored topic in the transportation research literature.

But, much of the research work focuses on path choice in the abstract, simply assuming that there is an origin, a destination, a network, and paths to be identified. The notion that the path choice problem should emphasize trucks, especially, or path choice factors that are of special concern to trucks is uncommon. One exception is the HazMat literature, that explicitly focuses on path choice for trucks. Admittedly, the commodity is special, and there are special concerns, but the findings are generalizable to trucks more generally. This is especially true in two senses. First, there are only some links (arcs) in the network that are available for use. Many urban areas designate a truck network that can (must) be used except for local pick-ups and deliveries. Second, multiple objectives are considered. In the case of HazMat, the post common ones are cost and risk. But, there can be others, such as exposure to accidents or challenging geometric conditions.

This study effort simply suggests that multiple objectives are typically important. And, it uses cost, risk, and tolls to motivate that thought. In fact, the tolls are seen as a control mechanism that network managers can use to encourage specific tendencies in the selection of truck routes. The idea is somewhat counter-intuitive since the "objective" is to keep the trucks off local streets and encourage them to use high0type facilities. This is important, because the more common purpose in introducing tolls is to rationalize the use of capacity-challenged facilities, like bridges and tunnels, that have limited capacity but high demand. Those tolls tend to discourage traffic from using facilities that are "good" from the perspective of truck paths. Hence, this motivates the idea that two toll structures might be useful. One would focus on auto trips, and discourage highly peaked demands, and highly-peaked flows, and the use of capacity-challenged facilities where other (lower class facility) options are available. In contrast, the pricing structure for the trucks should encourage the use of high-type facilities (e.g., freeways) and discourage the use of low-type facilities (like local streets). It is important to see that this difference in perspective is important and to utilize this insight to design the network pricing structure.

The section presents three ways to develop truck-focused paths. The first follows the paradigm created by Dial *et al.* (xxx) that emphasizes the distribution of flows among paths based on their relative "impedances" or costs. That is, paths with better (lower) costs would see a higher percentage of the flows and those with worse (higher) costs would see less. A drawback is that the procedure does not examine the objectives individually, it combines them through a generalized cost function and uses that function to compute the "cost" of each route. Also, the method is "weak" in that it does not identify paths explicitly. Rather, it ascertains the percentage of each OD flow that will use specific arcs. These values are commonly called arc utilizations and are employed heavily in traffic assignment procedures. In a truck context, the main nuance is that the generalized cost functions are different for the flow classes. Autos are focused on cost (based on travel time and distance) and to some degree tolls. But, the trucks are far more focused on tolls; and, to some degree risks. So, in identifying paths by class, it is important that these different generalized cost functions be employed.

The second methodology identifies the K-shortest paths for each OD pair. As might be obvious, these paths range from the "shortest" (lowest cost) to the "longest" (highest cost). In the sense that a (linear) generalized cost function is employed to identify the paths, all K of the paths reflect the weights employed by that function. The non-dominated paths that reflect trade offs among those objectives are not identified. In fact, as Figure 3.2 shows, the relative combination of the objectives can vary widely from one of the K shortest paths to another. This means the decision maker needs to be comfortable with seeing the ratios between the objectives vary considerably as the paths progress from k = 1 to k = K.

The second methodology explicitly identifies the non-dominated set of paths for each OD pair. If there are two objectives, then the result is akin to Figure 1.3, as is shown by Figure 3.3, where a set of four non-dominated paths were found for a specific OD pair. An advantage to this methodology is that it identifies the set of paths that has the "optimal" tradeoffs among the objectives. That is, if switching from path A to path B allows one objective's value to be made better at the expense of making another objective's value "worse", the least damage to the second objective is done by selecting path B. Also, unlike the first path choice algorithm, the paths are identified explicitly. The two significant drawbacks are 1) the algorithm, unlike Dial's (1971) algorithm, it provides no guidance about the probabilities that the paths should be chosen. Or, alternately put, the percentage of flow that should use any path. Also, if flow is moved from one path to another, then a shift must be made to a path that proportionally involves a different weighted combination of the objectives, because it lies on the non-dominated surface, and those paths, inherently, span the space between and among the solutions that optimize the objectives one at a time.

It is not that one of these path choice options is "best" or "correct". They have strengths and weaknesses. Any one of them provides path choices for the traffic assignment problem that are useful and valuable. Also, they can all be sensitive to multiple objectives, either through a generalized cost function, or through an explicit identification of the non-dominated multi-objective path options. Which one is best to use depends on the manner in which the traffic assignment problem is approached. In this study, the third one has been carried forward because it provides a set of paths that lie on the non-dominated surface.

All the path choice methods can be made sensitive to real-time information about the evolving network conditions. They can all make use of emerging trends in the network travel times (rates). And, they can be sensitive to the inclusion / exclusion of specific arcs because of use restrictions (no trucks), either permanent, temporary, or condition (time-of-day) dependent; and they can be sensitive to changes in the "impedances" for the arcs in the sense of varying tolls. The one unexpected insight is that the tolls pertaining to trucks might be different from and motivated by a different objective that those that pertain to autos. In fact, the two may have opposite trends. While the auto-focused tolls may be intended to discourage peaking and the use of capacity-challenged arcs, the truck tolls may want to discourage use of local streets and low-type facilities; and encourage trucks to use the high-quality facilities, which may, in fact, be the ones that are capacity-challenged. This means the trucks are (or should be) encouraged to use the "best" facilities so that they do not "rat run" through the network to avoid tolls on the best facilities available. Rather, the tolls should encourage the trucks to stay on the best arcs in the network so that they do not seek paths that use local streets. This is a pricing strategy that has significant, "freight-first" implications.

4.0 TRAFFIC ASSIGNMENT

This section addresses the issue of traffic assignment in the context of a truck-first perspective. As described in Section 1.0, researchers typically think about traffic assignment in the context of urban highway networks, with the "objective" of finding "the best" way for trips to traverse a capacitated network, across space and time, from origins to destinations. The idea of "best" focuses on optimizing one or more measures of network (system) performance.

Section 1.0 provides a detailed description of the math typically involved in formulating traffic assignment problems. This section focuses more narrowly on the algorithm that has been developed to place an emphasis on trucks.

4.1 PROBLEM FORMULATION

There are many ways to formulate the problem. Consistent with Section 1.0, the network is assumed to be comprised of one-way arcs. Every bi-directional link has two one-way arcs. Paths are sequences of arcs. The paths that exist between origin and destination nodes are the most important ones.

The "objective" is to find path flows (between the *OD* pairs) that "optimize" one or more objectives. Here, three are considered: total cost, equity in cost among all *OD* travelers, and network resilience. The first is equivalent to the system optimal objective described in Section 1.0; the second, the user equilibrium "objective"; and the third, the network resilience objective.

Two ways are common for "solving" the problem. One is to formulate it as a mathematical programming problem for which a simultaneous optimal solution is to be found and solve it using a general-purpose problem solver, like LINGO or CPLEX. The other is to treat it algorithmically with a sequential, customized solver, that iterates between path development and traffic assignment. The math programming approach is employed here.

The notation for the model is as follows. Let $\underline{z}$ be a vector of objectives to be considered. Specifically, z_1 is the total cost of all travel; z_2 is the sum of exceedance "gaps" that exist between target travel costs for the *OD* pairs and the costs that arise for the flows; and z_3 is the sum of exceedance "gaps" between target v/c (volume to capacity) targets that exist for the arcs in the network and the v/c ratios that arise. Set N is the set of all nodes, and n is an index on those nodes. Some of the nodes are origin/ destination locations. That is, network flows originate and terminate at those nodes. The others are junctions between network arcs. Set A is the set of all arcs. Each arc a has a set of attributes: its multi-class flow rates (e.g., autos and trucks), va_a and vt_a; the overall flow rate combining the classes, v_a; the target upper bound for that flow rate, va_0; the extent to which that target is exceeded, va_e; the number of lanes nl_a, used to compute the capacity C_a; and the additional capacity dC_a, needed to accommodate the flow (and to ensure that feasible solutions are found). The paths have a flow rate for autos, va_p, and trucks vt_p; a cost for the autos, ca_p, and the trucks ct_p; target values for those costs, $ca0_p$ and $ct0_p$; and extents to which those target costs are exceeded, cae_p and cte_p. The path costs are computed by summing the arc costs, ca_a and ct_a. There are sets that map that paths to the *OD*s, P_{od}, and the arcs, P_a.

The paths are created by the multi-objective path building algorithm described in Section 3.3. The fact that all non-dominated paths are employed means that the objective functions affect the paths that are chosen for any given problem solution. The two objectives considered in the multi-objective path search algorithm were cost and risk. And, tolls were included in the computation of the cost. Separate paths could be identified for autos and trucks, with the tolls being included or not, but this has not yet been done. Moreover, real-time data about the distributions of travel times on the arcs could be used to determine the assignments identified; and, from a feedback perspective, then influence the vector of tolls employed.

The objective is to minimize the weighted sum of the objectives z_i:

$$z = \sum_i w_i z_i \tag{4-1}$$

The first objective is the total cost of all travel:

$$z_1 = \sum_p (ca_p fa_p + ct_p ft_p) \tag{4-2}$$

Summing over the *OD* pairs is implicit because each path belongs to some *OD* pair. This objective can also, alternately, be computed based on the arc flows and costs:

$$z_1 = \sum_a (ca_a fa_a + ct_a ft_a) \tag{4-3}$$

The second objective is the sum of the extent to which the path costs exceed target values:

$$z_2 = \sum_p (cae_p + cte_p) \tag{4-4}$$

where:

$$cae_p \geq ca_p - ca0_p \quad and \quad cte_p \geq ct_p - ct0_p \tag{4-5}$$

The values of cae_p and cte_p are zero if $ca_p \leq ca0_p$ and $ct_p \leq ct0_p$ respectively. This is a surrogate for the user equilibrium objective. If the target values are set to the equilibrium costs, then (4-4) measures the extent to which those values are exceeded among all paths.

The third objective captures the extent to which target *v/c* ratios are exceeded on the arcs:

$$z_2 = \sum_a fe_a \tag{4-6}$$

where:

$$fe_a \geq f_a - f0_a \tag{4-7}$$

and

$$f_a = fa_a + \theta ft_a \tag{4-8}$$

where θ is the passenger car equivalent of a truck (typically 2.0 for urban settings) and the values of fa_a and ft_a are computed as follows:

$$fa_a = \sum_{p \in P_a} fa_p \quad and \quad ft_a = \sum_{p \in P_a} ft_p \tag{4-9}$$

The target values are based on a policy-based upper limit β on the *v/c* ratio that is to be achieved, the nominal capacity of the arc C_a and the capacity, if any, that has been added, dC_a :

$$f0_a = \beta(C_a + dC_a) \tag{4-10}$$

The fourth objective captures the cost of adding capacity to the network:

$$z_4 = \sum_a \lambda * d_a * dC_a \tag{4-11}$$

where λ is the per-unit cost of adding capacity to an arc and d_a is the length of the arc.

The travel times on the arcs a function of the flows on the arc:

$$t_a = t_{0a}(1.0 + \alpha(v_a / C_a)^\beta \tag{4-12}$$

where, the value of α is 0.1 and β is 1.0 (to avoid strong non-linearity). The costs are given by a weighted combination of these travel times and the arc lengths, to capture time and distance-related cost elements. The fact that the times are sensitive to the flow rates means the overall problem is non-linear.

All the flow F_{od} for each *OD* pair must be assigned to one or more of the paths that exist:

$$\sum_{p \in P_{od}} fa_p = Fa_{od} \quad and \quad \sum_{p \in P_{od}} ft_p = Ft_{od} \tag{4-13}$$

4.2 CASE STUDY

To show how the model can be used to obtain traffic assignment solutions, it has been applied to the same Albany network used previously and described in Section 3.1. The problem formulation shown in Section 4.1 was translated into a LINGO problem statement. The actual LINGO model statement is shown below. It may be of limited value to most readers. But, to some, it will be particularly clear about the way in which the problem has been represented:

```
************************************

model:

 sets:
   objs/1 .. 4/: z, w;
   node/1 .. 86/;
```

```
  arc/1 .. 248/: va, vaa, vat, va0, nln, d, ca, ra, csta, cstt, dv, dv0, cap, dcap;
  path/1 .. 224/: vpa, vpt, cpa, cpt, cpea, cpet, cp0a, cp0t, pfr, pto, ptf;
  flow/1 .. 202/: odo, odd, oda, odt, odflow;
  pamap/1 .. 393/: pk, ak;
 endsets

 [obj] min = @sum(objs(i): w(i)*z(i));
 z(1) = @sum(arc(n): csta(n)*vaa(n) + cstt(n)*vat(n));
 z(2) = @sum(path(p): cpea(p) + cpet(p));
 z(3) = @sum(arc(n): dv(n));
 z(4) = @sum(arc(n): d(n)*dcap(n));

 @for (flow(k): [aflw] oda(k) = (1-tpc)*odflow(k)*100);
 @for (flow(k): [tflw] odt(k) = tpc*odflow(k)*100);
 @for (flow(k):[toda]
    @sum(path(p) | pfr(p) #eq# odo(k) #and# pto(p) #eq# odd(k): vpa(p)) = oda(k) );
 @for (flow(k):[todt]
    @sum(path(p) | pfr(p) #eq# odo(k) #and# pto(p) #eq# odd(k): vpt(p)) = odt(k) );
 @for (path(p)| ptf(p) #eq# 0:[ntrk] vpt(p) = 0 );

 @ for (arc(n): [avol] vaa(n) = @sum(pamap(k) | ak(k) #eq# n : vpa(pk(k))) );
 @ for (arc(n): [tvol] vat(n) = @sum(pamap(k) | ak(k) #eq# n : vpt(pk(k))) );
 @ for (arc(n): [tavol] va(n) = vaa(n) + 2.0*vat(n) );
 @ for (arc(n): [capa] va(n) < cap(n) + dcap(n) );
 @ for (arc(n): [dvval] dv(n) > va(n) - 0.7*(cap(n) + dcap(n)) );
 @ for (arc(n): [acsta] csta(n) = ca(n)*(1.0 + 0.2*dv(n)));
 @ for (arc(n): [acstt] cstt(n) = ca(n)*(1.0 + 0.2*dv(n)) + ra(n) );
 @ for (arc(n): [acap] cap(n) = 1700*nln(n) );
 @ for (arc(n): [udcap] dcap(n) < 1700 );

 @for (path(p): [pcsta] cpa(p) = @sum(pamap(k) | pk(k) #eq# p: csta(ak(k))) );
 @for (path(p): [pcstt] cpt(p) = @sum(pamap(k) | pk(k) #eq# p: cstt(ak(k))) );
 @for (path(p): [covra] cpea(p) > cpa(p) - cp0a(p) );
 @for (path(p): [covrt] cpet(p) > cpt(p) - cp0t(p) );
 @for (path(p): [pcst0a] cp0a(p) = @sum(pamap(k) | pk(k) #eq# p: ca(ak(k))) );
 @for (path(p): [pcst0t] cp0t(p) = @sum(pamap(k) | pk(k) #eq# p: ca(ak(k)) + ra(ak(k))) );

data:
 w     = 1.0, 1.0, 1.0, 1.0;
 tpc   = 0.1;
 va0   = @ole('Albany.xlsm','va');
 d     = @ole('Albany.xlsm','dist');
 odo   = @ole('Albany.xlsm','odo');
 odd   = @ole('Albany.xlsm','odd');
 odflow = @ole('Albany.xlsm','odflow');
 ca    = @ole('Albany.xlsm','ca');
 ra    = @ole('Albany.xlsm','ra');
```

```
 nln   = @ole('Albany.xlsm','nln');
 pfr   = @ole('Albany.xlsm','pfr');
 pto    = @ole('Albany.xlsm','pto');
 ptf    = @ole('Albany.xlsm','ptf');
 pk     = @ole('Albany.xlsm','pk');
 ak     = @ole('Albany.xlsm','ak');
 @ole('Albany.xlsm','vpa') = vpa;
 @ole('Albany.xlsm','vpt') = vpt;
 @ole('Albany.xlsm','cpa') = cpa;
 @ole('Albany.xlsm','cpt') = cpt;
 @ole('Albany.xlsm','csta') = csta;
 @ole('Albany.xlsm','cstt') = cstt;
 @ole('Albany.xlsm','vaa') = vaa;
 @ole('Albany.xlsm','vat') = vat;
 @ole('Albany.xlsm','dcap') = dcap;
 @ole('Albany.xlsm','zVal') = z;
 @ole('Albany.xlsm','objAssn') = obj;
enddata

***********************************
```

The "sets" section defines the size of the problem and the choice variables employed. The statements between the "endsets" and "data" lines define the equations in the problem. The @for statements identify the ranges over which the statements pertain; the @sum statements identify the limits over which sums are to be computed; the labels in […] identify names for the constraints; and the qualification statements following the vertical lines "|" qualify the conditions for which the equations are to be created and/or the linkages between choice variables that are in different sets. The @ole statements connect the LINGO formulation to an Excel workbook that contains the data. Where the @ole is on the right-hand side of the assignment, data are being transferred from the Excel workbook into LINGO. Where they are on the left-hand side, results from LINGO are being deposited into the Excel workbook. The named ranges that follow identification of the workbook indicate where the data reside or into which the results should be placed.

o	d	p	tf	vpa	vpt	cpa	cpt
1	9	1	1	1283	143	60	145.3
2	9	2	1	4920	547	3625	3629
2	12	3	1	2206	245	76.8	112.8
2	47	4	1	647	71.9	32.4	40.43
2	84	5	1	1028	114	11.3	13.47
3	4	6	1	1133	126	24	82.27
3	8	7	1	5091	566	8102	8120
3	18	8	1	353	39.2	54	275.1
3	18	9	1	0	0	8143	8177
3	67	10	1	4568	508	2264	2268
3	84	11	1	1643	183	14.2	23.37
4	3	12	1	1133	126	24	82.27
4	5	13	1	212	23.5	54	262.2
5	4	14	1	212	23.5	54	262.2
5	11	15	1	670	74.4	12	12.9
5	75	16	1	1643	183	9	12.13
6	10	17	1	1238	138	30	95.06
7	8	18	1	1138	126	14.4	21.15
7	10	19	1	3940	438	365	373.2
7	17	20	1	1601	178	31.2	71.41
7	17	21	1	0	0	44.4	83.22

Table 4.1: Excerpt of the path-based results from solving the multi-class traffic assignment problem

The results from exercising the model can be presented in a tabular format. Table 4.1 shows an excerpt of the path-based output. Each row shows the origin, destination, path, a truck-use flag (1 means trucks can use the path), the auto flow rate on the path, the truck flow rate on the path, the auto cost on the path and the truck cost for the path. The auto and truck costs are different because the cost equations involve different weights for a) time, distance, and toll-based costs and b) risk.

An excerpt of the results from an arc perspective are shown in Table 4.2. Each row shows the arc, its from and to nodes, its class, truck use flag, and number of lanes; its nominal flow rate based on field observations, its nominal cost and risk, predicated on field observations (especially of the travel times) during the timespan (e.g., AM peak) of interest, and the results of the traffic assignment: the auto and truck flow rates, the auto and truck costs (again, different because of the weights employed), and the capacity, if any, added to the arc to accommodate the flow rates.

Link	Arc	From	To	Class	tFlag	Dist	nLane	Flow	Time	Cost	Risk	vaa	vat	csta	cstt	dcap
1	1	1	9	6	1	5	1	1426	8.6	60	85.3	1283	143	60	145.3	541
2	2	2	9	1	1	1.2	3	5467	1.1	14.4	4.24	4920	547	3625	3629	1700
3	3	2	12	1	1	6.4	3	2451	5.9	76.8	36	2206	245	76.8	112.8	0
4	4	2	47	1	1	2.7	1	719	2.5	32.4	8.03	647	71.9	32.4	40.43	0
5	5	2	84	1	1	0.94	3	1142	0.9	11.3	2.19	1028	114	11.28	13.47	0
6	6	3	4	6	1	2	1	1259	3.4	24	58.3	1133	126	24	82.27	278
7	7	3	8	1	1	2.3	3	5657	2.1	27.6	18.6	5091	566	8102	8120	1700
8	8	3	18	6	1	4.5	1	392	7.7	54	221	353	39.2	54	275.1	0
9	9	3	67	1	1	1.14	3	5075	1.1	13.7	4.43	4568	508	2264	2268	1700
10	10	3	84	1	1	1.18	3	1826	1.1	14.2	9.21	1643	183	14.16	23.37	0
11	11	4	5	6	1	4.5	1	235	7.7	54	208	212	23.5	54	262.2	0
12	12	4	6	8	1	0.5	1	0	1.2	6	10.9	0	0	6	16.91	0
13	13	4	9	8	1	3.6	1	0	8.6	43.2	142	0	0	43.2	184.8	0
14	14	5	11	1	1	1	3	744	0.9	12	0.9	670	74.4	12	12.9	0
15	15	5	75	6	1	0.75	2	1825	1.3	9	3.13	1643	183	9	12.13	0
16	16	6	7	8	1	1.2	1	0	2.9	14.4	28.4	0	0	14.4	42.83	0
17	17	6	10	6	1	2.5	1	1375	4.3	30	65.1	1238	138	30	95.06	461
18	18	7	8	1	1	1.2	3	1264	1.1	14.4	6.75	1138	126	14.4	21.15	0
19	19	7	10	1	1	2.5	3	4378	2.3	30	8.37	3940	438	364.8	373.2	1700
20	20	7	16	8	1	2.7	1	0	6.5	32.4	35	0	0	32.4	67.42	0

Table 4.2: Excerpt of the arc-based results from solving the multi-class traffic assignment problem

4.3 SUMMARY

This section has described a realization of the traffic assignment problem in which trucks are represented separate from autos. The problem is non-linear in that the travel times on the arcs are sensitive to the arc flows. And that, in turn, leads to one of the objectives being quadratic. The terms involve the multiplication of one choice variable times another. Fortunately, for small problems, there are generalized solvers that can deal with such situations. For large-scale problems, like the ones that are faced for most metropolitan areas, such a non-linear formulation is not practical to solve explicitly. But, for illustration purposes, a simplified network for the Albany, NY metropolitan area is used in the case study to show the type of results obtained.

The most important nuance in the model is the fact that the path choices for the trucks are different from those for the non-trucks (autos). The generalized cost function is different and the network over which the trucks can travel is more restrictive. The trucks take into consideration the risk associated with traversing the arcs (as in exposure to accidents) and the tolls are given considerable weight. Also, the toll structure is different for trucks than it is for autos. The tolls for autos discourage use of capacity-challenged facilities like bridges, tolls, and heavily used freeway links, encouraging the vehicles to use other, comparable, but “lower quality” facilities, like the arterials, in their paths. For the trucks, however, the toll structure discourages the use of local streets and other “lower class” facilities so that neighborhoods are not exposed to unnecessary truck traffic. And, they are encouraged, through low tolls, to use the “high type” facilities, such as freeways, except for local pickups and deliveries. This bi-pronged pricing strategy helps to put “freight-first” in the context of the traffic assignment. And, it leads to solutions that the public is “more likely” to accept because it discourages trucks from using local streets and highways.

The mathematical equations that are involved in the traffic assignment model are presented and described as well as the LINGO problem statement which implements them. A case study example is presented, with excerpts of the results, so that the reader can gain a sense of the results obtained. (The complete, machine readable workspaces exist and are available for anyone wishing to use them.)

5.0 BRAESS' PARADOX

This section focuses on an examination of Braess' paradox under stochastic conditions. Much of what is presented is based on the research work of Yi Chen, a doctoral student. The paradox asserts that adding capacity to a network, in the form of a new arc, can degrade overall performance; for example, increasing travel times. For example, as was discussed in Section 1, for the network shown in Figure 5.1, adding arc RS to the network increases the travel times.

If the total flow between *O* and *D* is 1000 trips per hour and arc RS is absent, then the trips will divide equally between path ORD and path OSD. It is easy to see this because the total travel time equations on the two paths are the same: 60 + 0.04**f*. So, if $f_{ORD} = f_{OSD}$, then $t_{ORD} = t_{OSD}$. And, if the total flow is 1000 trips, then 500 trips should use each of the paths because the travel times would both be 80 = 60 + 0.04*500.

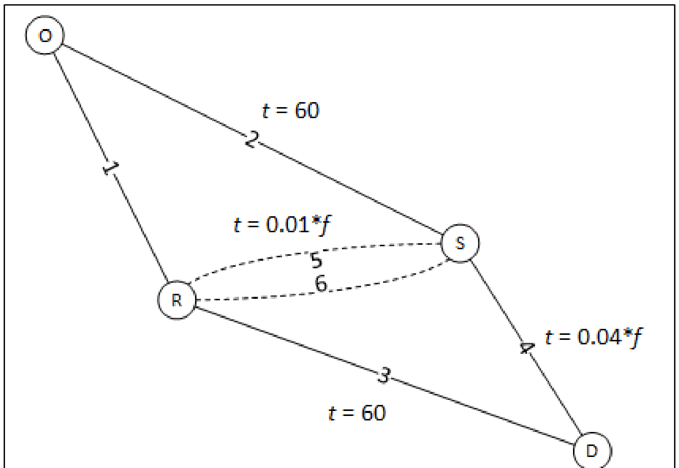


Figure 5.1: A setting where Braess' paradox

If arc RS is introduced, however, from R to S, and nothing else changes, Wardrop's (1952) second principle suggests that users will check to see if ORSD is faster. And if so, they will switch. For the first user to switch, say from ORD to ORSD, the travel time will be t_{ORSD} = 40.05 = 0.04*500 + 0.01*1 + 0.04*501 = 20 + .01 + 20.04. In this case, that switching will continue from both ORD to ORSD and OSD to ORSD until if $f_{ORD} = 0$, $f_{OSD} = 0$, and if $f_{ORSD} = 1000$. At that point, t_{ORSD} = 90 = 40 + 10 + 40; and t_{ORD} = 100 = 40 + 60; and t_{OSD} = 100 = 60 + 40. No user has an incentive to shift from ORSD to one of the other two paths. Yet, unfortunately, this equilibrium condition involves a trip time for everyone which is larger than it was without arc RS, 90 versus 80. But, under the premises of user equilibrium, there is no incentive for any user to leave path ORSD. So, introducing arc RS, even though it produces a much shorter travel time for the first user, 40.05 versus 80, in the limit, once all the traffic shifts, the result is a higher travel time.

5.1 SETTING THE CONTEXT

Milchtaich (2005) concludes that Braess' paradox cannot occur unless the network contains a network (or subnetwork) like the one shown in Figure 5.1 This diamond-shaped structure with a transverse connection is called a Wheatstone (1843) bridge. It is the operation of the two competing, parallel, and nearly identical cost routes that is confounded (compromised) by the new arc.

Most studies of Braess' paradox assume the flows are directional, from *O* to *D*, there are no other flows. Also, the arc cost functions are such that one arc, from R to S in this case, can cause the paradox to occur.

In this study, however, the coefficients of the arc cost functions are random variables, so it is not known a priori, whether the network will exhibit the paradox at all; and if it does, it is not possible to pre-determine whether arc RS will cause the paradox. It might be arc SR instead. Hence, both arcs RS and SR are included in the network. (This is also a more realistic, since, when links are added to the network, in most cases two-way flow is allowed.)

Most studies focusing on Braess' paradox assume that the network is operating under user equilibrium conditions, as has been done in the example above. But, in the study presented here, the users are assumed to be of two types as was assumed by Bennett (1993): those that align with the user equilibrium (UE) objective and those that adhere to the system optimal (SO) objective. The former, it could be argued, could be trucks; since they are profit motivated. The latter could be autos, since citizens, in general, are "expected" to follow public agency edicts about minimizing the societal cost of the transport activity.

Given the six-arc network shown in Figure 5.1, there are four paths (arc sequences): [1, 3], [2, 4], [1, 5, 4], and [2, 6, 3]. Given the way that the network is drawn, some of these paths, like [2, 6, 3] seem like non-starters. But, the reader must keep in mind that the coefficient values on the arcs are being treated as random variables, so the actual travel time-scaled depiction of the network might be very different from the one drawn.

5.2 PROBLEM DEFINITION

The following inputs are involved:

Vu : *demand for the traffic that subject to user equilibrium criteria.*

Vs : *demand for the traffic that subject to system equilibrium criteria.*

V : *total demand*, $V = Vu + Vs$.

ca_i : *capacity for link* i, $i = 1\ldots6$.

a_i : *slope of the link travel time function.*

b_i : *intercept of link travel time function.*

The variables are:

u_i : *volume for the traffic that subject to user equilibrium criteria on* path i, $i \in I$

s_i : *volume for the traffic that subject to system equilibrium criteria on* path i, $i \in I$.

vp_i : *traffic volume on path* i, $i \in I$, $vp_i = u_i + s_i$.

u_j : *volume for the traffic that subject to user equilibrium criteria on* link j, $\mathrm{j} \in \mathrm{J}$.

s_j : *volume for the traffic that subject to system equilibrium criteria on* link j, $\mathrm{j} \in \mathrm{J}$..

va_j : *traffic volume on link* j, $\mathrm{j} \in \mathrm{J}$, $va_j = u_j + \mathrm{s}_j$.

ta_j : *travel time on link* i, *as a linear function of traffic volume on it* $\mathrm{ta}_j = a_j * \mathrm{va}_j^{\,\mathrm{n}} + b_j, \mathrm{j} \in \mathrm{J}$.

tp_i : *travel time on path i*, $tp_i = \sum_{j=1}^{6} \delta_{ij} ta_j \;\; i \in I$

$$\delta_{ij} := \begin{cases} 1, \textit{link } \text{j is on path i} \\ 0, \text{o/ w} \end{cases}$$

I : index set for path, I={1,2,3,4}.

J: index set for link, J={5,6,7,8,9,10}

Feasible flows for the network must satisfy the following constraints:

$$\text{tp}_i = \sum_{j \in J} \zeta(i, j) * ta_j \qquad \text{i} \in \text{I} \tag{5.1}$$

$$u_j = \sum_{i \in I} \delta(i, j) * u_i \qquad \text{j} \in J \tag{5.2}$$

$$s_j = \sum_{i \in I} \delta(i, j) * s_i \qquad \text{j} \in J \tag{5.3}$$

$$\text{va}_j = \text{u}_j + \text{s}_j \qquad \text{j} \in J; \tag{5.4}$$

$$\text{vp}_i = \text{u}_i + \text{s}_i \qquad \text{i} \in I; \tag{5.5}$$

$$\sum_{i \in I} u_i = \text{Vu}; \tag{5.6}$$

$$\sum_{i \in I} s_i = \text{V-Vu}; \tag{5.7}$$

$$\text{ta}_j = \text{b}_j + \text{a}_j * \text{va}_j^{\text{n}} \qquad \text{j} \in J; \tag{5.8}$$

$$\text{va}_j < \text{ca}_j \qquad \text{j} \in J; \tag{5.9}$$

$$\text{Z} = \zeta(i, j) = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{5.10}$$

$$\Delta = \delta(\text{i, j}) = Z^{'} \tag{5.11}$$

The SO problem can be stated as:

$$(P1) \quad \min \sum_{j \in J} \text{ta}_j * \text{va}_j = \sum_{i \in I} tp_i * vp_i$$
$$st: \quad (5.1)\text{-}(5.11)$$

The UE problem can be stated as:

$$(P2) \quad \min \sum_{j \in J} \int_0^{\text{va}_j} ta_j dx$$
$$st: \quad (5.1)\text{-}(5.11)$$

Bennett (1993) shows that the multi-class traffic assignment problem involving both SO and UE-motivated users can be formulated as follows:

$$(P3)\quad \min \sum_{j\in J}\left(\int_0^{va_j} a_j x^n dx + b_j u_j + \frac{1}{n+1} b_j \mathrm{s}_j\right)$$

$$st: \quad (5.1)\text{-}(5.11)$$

Bennett (1993) does not consider capacity constraints. However, Larsson and Patriksson (1995) show that under user equilibrium conditions, capacity can be included as a side constraint. The revised formulation is equivalent to solving the traditional Wradrop (1952) equilibrium problem with a generalized link travel cost function. The augmented objective function becomes the original travel cost function plus a Lagrange multiplier for each capacity constraint. It can be proved that the same equivalence holds under the multi-equilibrium situation and thus the capacity constraints can be included by adding side constraints to the model.

Theorem 5.1. In P3, let $\beta = (\beta_1 \dots \beta_6)^t$ be the vector of Lagrange multipliers for the capacity constraints (5.9) and let $\mu = (\mu_1 \dots \mu_{31})^t$ be the Lagrange multipliers for constraints 1 to 8). Then P3 is equivalent to following traditional multi-equilibrium problem:

$$(P4)\quad \min \sum_{j\in J}\left(\int_0^{va_j} a_j x^n dx + (b_j+\beta_j) u_j + \frac{1}{n+1} (b_j+\beta_j) \mathrm{s}_j\right)$$

st: (5.1)-(5.11)

and

$$\mathrm{ta}_j = \mathrm{b}_j + \beta_j + \mathrm{a}_j * \mathrm{va}_j^{\,\mathrm{n}} \qquad j \in J; \tag{5.12}$$

The proof is as follows: in cases where the travel cost functions are separable, then (P2) is a convex network optimization problem. Thus, for (P3) and (P4) the first part of the objective function can be shown to be convex and the second and third parts of the objective function are linear and convex. Since the sum of convex functions is still convex, the objective function is also convex.

Problem P3 can be written in the following general form:

$$(P5)\quad \min \mathrm{f(x)}$$

$$\text{st: } \mathrm{h(x)=0};$$

$$\mathrm{g(x)} \le 0;$$

This is a convex programming problem.

μ is the vector of Lagrange multipliers for the constraints h(x)=0.

β is vector of Lagrange multipliers for the contraints g(x)=0

The optimal solution for P5 can be found by solving:

$$\nabla L(x) = \nabla f(x) + \sum \mu_i \nabla h_i(x) + \sum \beta_i \nabla g_i(x) = 0 \qquad (5.13)$$

$$\because \nabla g_i(x) = e_i$$

$$\therefore (5.13) \Leftrightarrow \nabla L(x) = \nabla f(x) + \sum \mu_i \nabla h_i(x) + \beta = 0 \qquad (5.14)$$

Problem P4 can be re-written in the following form:

$$(P6) \quad \min \mathrm{f}^1(\mathrm{x})$$
$$\text{st: } \mathrm{h}^1(\mathrm{x})=0;$$

This is a convex programming problem; hence: $\nabla h^1(x) = \nabla h(x)$ and $\nabla \mathrm{f}^1(x) = \nabla f(x) + \beta$. This means the optimal solution for P6 is the solution for:

$$\nabla L^1(x) = \nabla f^1(x) + \sum \mu_i \nabla h_i^1(x) = \nabla L(x) = \nabla f(x) + \sum \mu_i \nabla h_i(x) + \beta = 0.$$

Thus, P5 and P6 have the same optimal solution. And, in as much as they are both convex problems, they are equivalent. Thus, the problem with capacity-related side constraints can be treated as a generalized Wardrop equilibrium problem.

The link travel times can be assumed linearly depend on the traffic volume on the link:

$$ta_j = a_j * va_j + b_j$$

By rearranging the equations and setting the traffic volume on each path as a decision variable, it is possible to write the problem P3 as follows:

$$\min \sum_{j \in J} \{\mathrm{a}_j (\sum_{i \in I} \zeta(\mathrm{i, j})(\mathrm{u}_i + s_i))^2 + b_j \sum_{i \in I} \zeta(\mathrm{i, j})\,\mathrm{u}_i + \frac{1}{2}\zeta(\mathrm{i, j})\,\mathrm{s}_i\}$$

$$st.$$

$$\sum_{i \in I} \zeta(\mathrm{i, j})(\mathrm{u}_i + \mathrm{s}_i) < \mathrm{ca}_j \qquad j \in J \qquad (5.15)$$

$$\sum_{i \in I} u_i = Vu \qquad (5.16)$$

$$\sum_{i \in I} s_i = V - Vu \qquad (5.17)$$

$$u_i \ge 0; \qquad i \in I \qquad (5.18)$$

$$s_i \ge 0; \qquad i \in I \qquad (5.19)$$

5.3 CASE STUDIES

As mentioned previously, the main objective of this study is to examine the relationship between the network parameters and system performance. Although the network shown in Figure 5.1 is simple, it has many parameters that need to be studied. These are 1) the slope and intercept values for the travel time functions on the arcs, 2) the capacity values for each arc, 3) the total traffic volume assigned to the network, and 4) the breakdown of users between altruistic and selfish.

To stress the importance of this experimental design, unlike most studies of the paradox, it is assumed that the users are of two types, altruistic and selfish, with a breakdown between the two. This means the solutions are not just UE or SO. They are a combination. So, the travel times on the paths may or may not all be the same. If the users are all UE and the capacities of the arcs are non-binding, then the travel times will be the same on all paths (because the UE solution will pertain), but otherwise, they may not. Also, the capacity of the new arc can vary. (In Braess' original study it was only either zero or infinite.) It is purposefully varied parametrically from zero to infinity; looking to see if the paradox arises for any two capacity values c_1 and c_2 across the spectrum, where $c_2 > c_1$. It is concluded that it does arise if the average travel time for c_2 is greater than it is for c_1, that is $\bar{t}_2 > \bar{t}_1$, or if the maximum travel time among all paths for c_2 is greater than it is for c_1, that is $\hat{t}_2 > \hat{t}_1$. The notation, here, is that $\bar{t}_\bullet$ is the average travel time for all paths and $\hat{t}_\bullet$ is the maximum travel time among all used paths.

From preliminary analyses, it became apparent that Braess' definition of the paradox would be insufficient to capture the results of the analyses. An expanded set of three paradox conditions was needed (the definition of t can be either the maximum or the average):

1) $t_{\max} > t_0$. This is the original definition of the paradox, if only UE users exist, with the nuance that the travel times may not all be the same since both SO and UE users exist.
2) For some c_1, where $c_0 < c_1 < c_{max}$, $t_1 > t_0$. That is, the paradox arises for at least one value of capacity in-between c_0 and c_{max}.
3) For some c_1 and c_2, where $c_0 < c_1 < c_2 < c_{max}$, $t_2 > t_1$. That is, the paradox arises for some intermediate pair of capacities.

A random number generator is used to create realizations of the network. For each, the performance trends are obtained such as the influence of the capacity of the new arc. That is, how the patterns in the maximum and average travel times are influenced by the capacity. Many realizations are examined so that the percentage occurrence of the paradox can be tracked to the combinations of parameter values. Statistical analyses are conducted to gain a general understanding of the trends.

Illustrations of the three conditions are shown in Figure 5.2 for the averages and maximums. The left-hand graph shows the average travel time trends. The one at right shows the maximum.

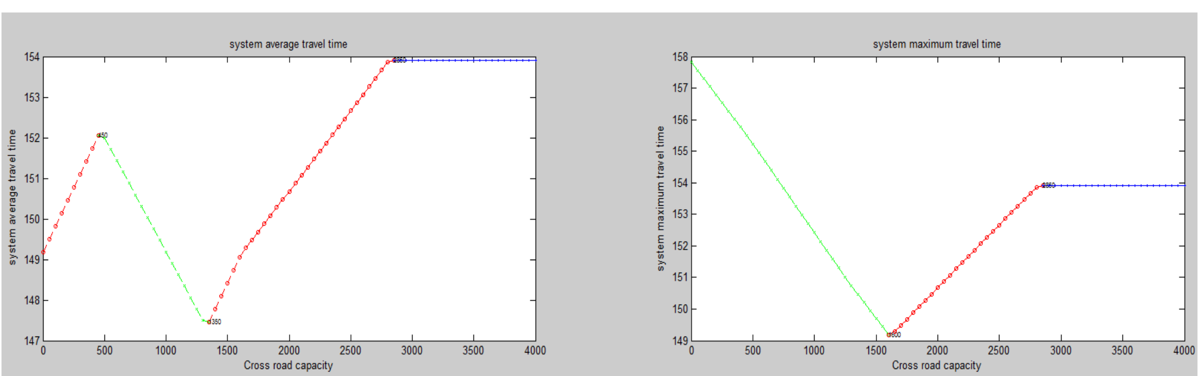


Figure 5.2: Trends in travel times as a function of arc capacity

It is important to note that the trends are piecewise linear. The travel times (average and maximum) do not simply increase as the capacity increases, which is the impression given by Braess' original analysis. Rather, they increase *and* decrease as the capacity changes, showing the impacts on the path choices. For the graph at the left, conditions 1, 2, and 3 are all satisfied. The first increasing trend satisfies condition 2; the second upward trend satisfies condition 3, and the graph, overall, satisfies condition 1. For the maximum travel time trends, condition 3 is satisfied, but not either 1 or 2. That is, there are combinations of c_1 and c_2, where $c_0 < c_1 < c_2 < c_{max}$, such that $\hat{t}_2 > \hat{t}_1$, but there is no c_1 such that $\hat{t}_1 > \hat{t}_0$, nor is it true that $\hat{t}_{\max} > \hat{t}_0$. This does show that the paradox assessment is complex, and care is required in analyzing and presenting the results, but the nuances do show that the paradox can arise under various conditions.

To gain a sense of when the paradox occurs, 30,000 network realizations were examined. Each one was created by randomly sampling values of the intercepts and slopes for the six arcs and the percentage of UE users (the others are SO). The values of a_j, b_j, and V_u, were sampled as follows:

$$a_j \in Unif[0, 0.02], \quad \forall\ j \in J$$
$$b_j \in Unif[0, 90], \quad \forall\ j \in J$$
$$Vu \in Unif[0, 4000]$$

For each realization, the nonlinear network flow optimization problem was solved multiple times for discrete values of capacity on arcs 5 and 6. The capacities of both arcs were increased simultaneously with the expectation that only one of the two arcs would be used for any realization. The main metrics of interest were: 1) the traffic volumes on the four paths 2) the travel times on the four paths, 3) the weighted average path travel time, and 4) the maximum path travel time.

A graphical tool was used to visualize the trends in the average and maximum travel times. By generating a set of plots for each realization it was possible to ascertain:

1) Whether the paradox occurs for the average travel time or the maximum travel time;
2) The number of local minimum/maximum points and their coordinates;
3) The number of increasing/decreasing intervals and how those intervals were ordered;
4) The piecewise linear relationship between the maximum/minimum points and the increasing/decreasing intervals based on the crossroad capacity.

An example realization helps illustrate. The parameter values that pertain are shown in Table 5.1.

a1	a2	a3	a4	a5	a6
0.0121	0.0192	0.0022	0.0069	0.0044	0.0050
b1	b2	b3	b4	b5	b6
14.1635	65.8377	61.7673	0.9927	6.8979	69.3316
Vu					
1983.3699					

Table 5.1: Parameter values for the realization

The trends in the traffic assignments are shown in Figure 5.3. The average travel time is at top left; the maximum travel time at top right; the path flows are shown at bottom left; and the path travel times at bottom right.

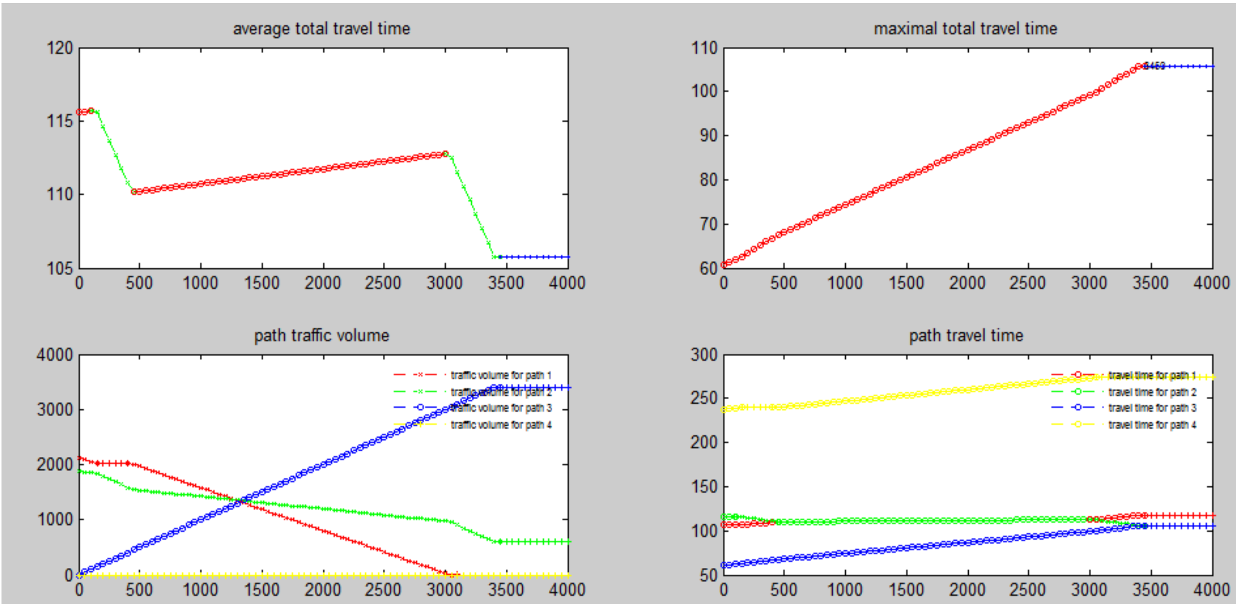


Figure 5.3: Trends in the travel times for a realization

As can be seen, the patterns are complex. The plot at top left plot shows the trends in the average total travel time. The one at top right plot shows the corresponding trend for the maximum travel time. The bottom left shows the trends in the path flows and the bottom right shows the trends in the path travel times. The average time increases at first, then decreases, then increases for a second time, decreases again, and finally remains constant. The maximum travel time increases and then stays constant. Hence, the average time has four intervals: two increasing and two decreasing; while the maximum travel time only has one. In terms of the paradox definitions, the second and third pertain to the average, but not the first. For the maximum trend, all three pertain.

A second example has the parameter values shown in Table 5.2.

a1	a2	a3	a4	a5	a6
0.0100	0.0153	0.0112	0.0190	0.0142	0.0053
b1	b2	b3	b4	b5	b6
30.4518	25.0503	0.0594	66.6014	39.1486	6.3119
	Vu				
	3858.9780				

Table 5.2: Parameter values for the second example realization

The trends in the traffic assignments are shown in Figure 5.4. As before, the average travel time is at top left; the maximum travel time at top right; the path flows are shown at bottom left; and the path travel times at bottom right.

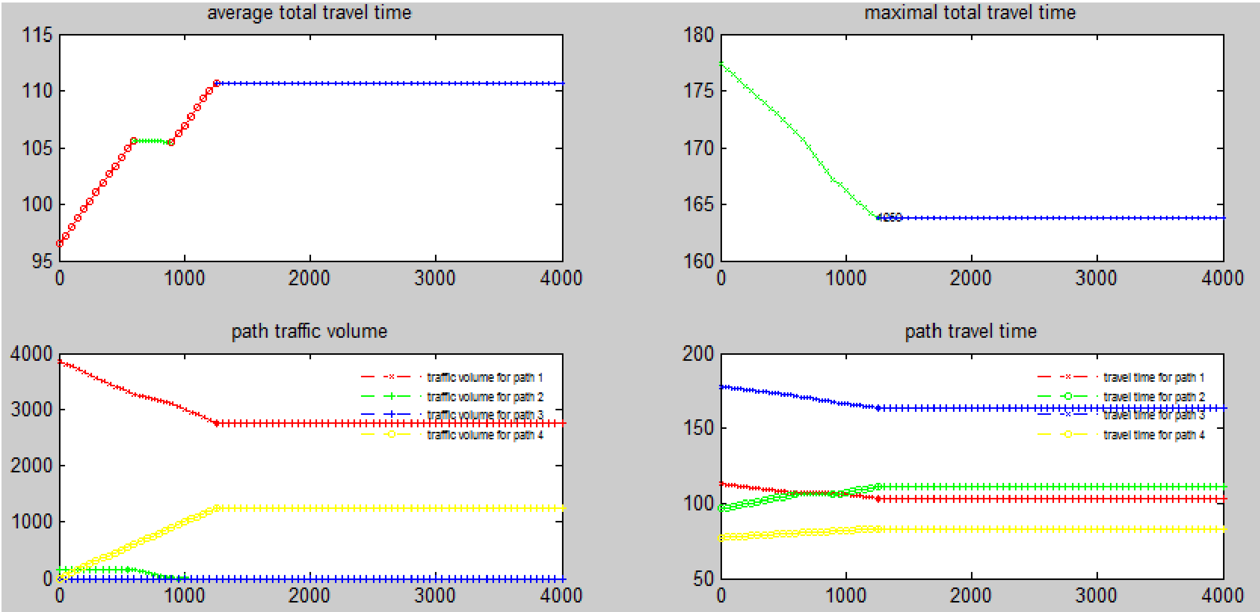


Figure 5.4: Trends in the travel times for a second realization

In this realization, the average time increases, decreases, increases, and then remains constant. It has three intervals. The maximum travel time has only one interval. The paradox only happens for the average travel time, but all three definitions are all satisfied.

A third example has the parameter values shown in Table 5.3.

a1	a2	a3	a4	a5	a6
0.0194	0.0104	0.0006	0.0067	0.0055	0.0029
b1	b2	b3	b4	b5	b6
30.8361	43.6451	76.9689	11.2087	1.6696	42.1350
Vu					
1527.8					

Table 5.3: Parameter values for the third example realization

The trends in the traffic assignments are shown in Figure 5.5. As before, the average travel time is at top left; the maximum travel time at top right; the path flows are shown at bottom left; and the path travel times at bottom right. In this realization, the average time has two intervals: an increase and then a decrease. The maximum time increases monotonically. The maximum time satisfies definitions 4, 5, and 6. The average travel time trends satisfy only the second and third definitions.

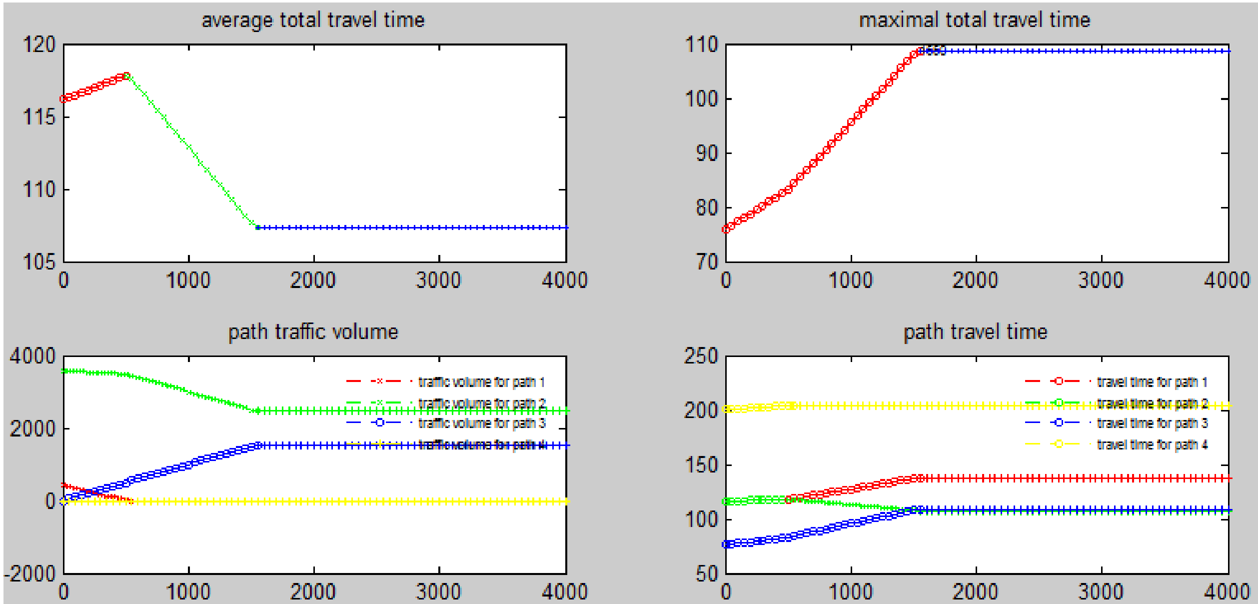


Figure 5.5: Trends in the travel times for a third realization

A fourth example has the parameter values shown in Table 5.4.

a1	a2	a3	a4	a5	a6
0.0046	0.0199	0.0160	0.0056	0.0192	0.0016
b1	b2	b3	b4	b5	b6
52.2448	5.3131	49.3495	77.6966	6.8670	0.6498
Vu					
2810.1					

Table 5.4: Parameter values for the fourth example realization

The trends in the traffic assignments are shown in Figure 5.5. As before, the average travel time is at top left; the maximum travel time at top right; the path flows are shown at bottom left; and the path travel times at bottom right.

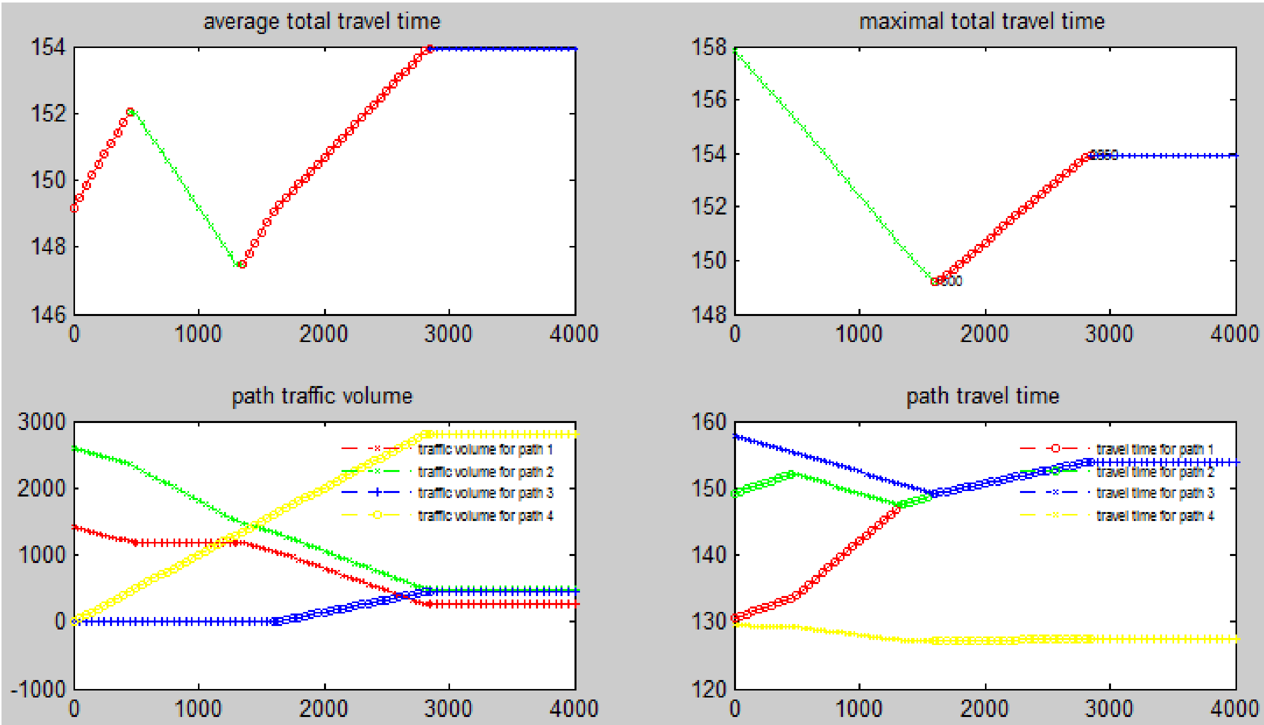


Figure 5.6: Trends in the travel times for a fourth realization

In this case, the trend in the average time has three intervals - increase, decrease, and then increase - while the trend for the maximum has two - decrease then increase. The average satisfies definitions 1, 2, and 3. The maximum only satisfies the third.

To summarize the trends in the four realizations:

1) All the stable intervals are last, which should be true under all circumstances since the changes in travel times are due to the amount of traffic assigned to the connecting arc. And, eventually, that stabilizes.
2) The traffic starts to use the connecting arc as soon as its capacity is non-zero.
3) The average travel time has more complex trend patterns. The maximum number of intervals for the average is four but for the maximum it is only two.
4) For any realization and capacity range, the first definition seems to be the most limiting one. The second is less limiting and the third is the least. Each definition is increasingly more inclusive. That is, if the first definition is met, all three will be met. If the second is met, then the third will be met.

These observations suggest that two pieces of information can be used to categorize the trend patterns: 1) the paradox definition that pertains, and 2) the number and order of intervals that occur.

Based on these pieces of information, the outcomes from the 30,000 realizations were obtained:

1) A 2 x 4 matrix that indicates which paradox (including none) occurred for the average and the maximum travel time

2) A count of the number of increasing and decreasing intervals for both the average and maximum travel time.
3) Two Boolean vectors, variable in length, that indicate the sequence of increases (1) and decreases (0) for the average and maximum travel times. For example, [0, 1, 0] indicates a decrease, then an increase, and then a decrease.
4) The flow rates associated with the inflection points. These are the capacity values at which the local minima and maxima occur.

Of the 30,000 different network realizations, 29,998 proved to be useful for analysis. Two had illogical results. Table 5.5 presents a breakdown of the observations for all three paradox definitions. Each is classified based on whether the paradox was satisfied for neither the average nor maximum travel time, just the average, just the maximum, or both.

Counts of Paradox Combinations (Average and Maximum Travel Times) - 29,998 Total											
Paradox Definition #1				Paradox Definition #2				Paradox Definition #3			
No - Average / No Maximum											
Maximum Intervals -> Average Intervals -v	0	1	2	Maximum Intervals -> Average Intervals -v	0	1	2	Maximum Intervals -> Average Intervals -v	0	1	2
0	22819	0	0	0	22819	0	0	0	22819	0	0
1	0	1910	0	1	0	1910	0	1	0	1910	0
2	0	255	0	2	0	140	0	2	0	0	0
3	0	45	0	3	0	25	0	3	0	0	0
4	0	16	0	4	0	4	0	4	0	0	0
		Total:	25045			Total:	24898			Total:	24729
No - Average / Yes - Maximum											
0	0	0	0	0	0	0	0	0	0	0	0
1	0	1854	0	1	0	1854	0	1	0	1854	0
2	0	226	0	2	0	129	0	2	0	0	0
3	0	66	0	3	0	37	0	3	0	0	0
4	0	12	0	4	0	5	0	4	0	0	0
		Total:	2158			Total:	2025			Total:	1854
Yes - Average / No - Maximum											
0	0	0	0	0	0	0	0	0	0	0	0
1	0	1048	1	1	0	1048	1	1	0	1048	0
2	0	235	0	2	0	350	0	2	0	490	0
3	0	38	1	3	0	58	1	3	0	83	0
4	0	13	0	4	0	25	0	4	0	29	0
		Total:	1336			Total:	1483			Total:	1650
Yes - Average / Yes - Maximum											
0	0	0	0	0	0	0	0	0	0	0	0
1	0	1163	0	1	0	1163	0	1	0	1163	1
2	0	241	0	2	0	338	0	2	0	467	0
3	0	40	0	3	0	69	0	3	0	106	1
4	0	15	0	4	0	22	0	4	0	27	0
		Total:	1459			Total:	1592			Total:	1765

Table 5.5: Interval combinations and definition categorizations for all 29,998 realizations

The three blocks of columns are for the three paradox definitions. The columns within the column blocks indicate how many intervals exist for the maximum travel time. The rows indicate how many intervals exist for the average travel time.

Many trends can be observed:

- The greatest number of intervals for the maximum travel time is two. For the average it is four.
- Under any paradox definition, most do not exhibit the paradox. For definition #1 there are 25,045 realizations where it does not occur; for definition #2, there are 24,898; and for definition #3, there are 24,729. Alternately put, the definition #1 is satisfied for only 4,953 realizations (16.5%); for definition #2, only 5,100 (17%); and for definition #3, only 5,269 (17.6%).
- As the paradox definition becomes more inclusive, the number of realizations exhibiting the paradox increases. Conversely, the number not exhibiting the paradox decreases.
- As the paradox definition becomes more inclusive, the number of realizations exhibiting the paradox for just the maximum and not the average decreases (2,158 -> 2,025 -> 1,854). For the average and not the maximum, it increases (1,336 -> 1,483 -> 1,650) and for realizations where both fulfill a definition, the trend is increasing (1,459 -> 1,592 -> 1,765).
- Of the realizations that do not exhibit the paradox, most (22,819) have no increasing or decreasing interval for either the average or the maximum.
- All the realizations exhibiting the paradox have one or more intervals for the average and one or two for the maximum. Alternately put, there are no instances where the average has no intervals; or the maximum has no intervals.
- Most of the realizations exhibiting the paradox have one interval for the maximum and one interval for the average. For example, in the case of paradox definition #1 for the realizations where the maximum exhibited the paradox, 1,854 of the 2,158 realizations have one interval for the maximum and one for the average.
- In contrast to the above, the realizations that change their paradox classification have two or more intervals for the average, not one.

Table 5.6 shows the incremental changes in Table 5.5 as the paradox definition becomes more inclusive. The first block of columns shows the changes from definition #1 to #2. The second block shows similar information for definition #2 to #3.

As might be expected the cells with the largest changes are those for which the number of intervals for the average is 2-4 and for the maximum is 1. The interval combinations with the most changes are 2 intervals for the average and 1 interval for the maximum. These realizations will be studied in more detail shortly.

Incremental Changes							
Definition #2 - Definition #1				Definition #3 - Definition #2			
No - Average / No Maximum							
Maximum Intervals -> Average Intervals -v	0	1	2	Maximum Intervals -> Average Intervals -v	0	1	2
0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	0
2	0	-115	0	2	0	-140	0
3	0	-20	0	3	0	-25	0
4	0	-12	0	4	0	-4	0
		Total:	-147			Total:	-169
No - Average / Yes - Maximum							
0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	0
2	0	-97	0	2	0	-129	0
3	0	-29	0	3	0	-37	0
4	0	-7	0	4	0	-5	0
		Total:	-133			Total:	-171
Yes - Average / No Maximum							
0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	-1
2	0	115	0	2	0	140	0
3	0	20	0	3	0	25	-1
4	0	12	0	4	0	4	0
		Total:	147			Total:	167
Yes - Average / Yes - Maximum							
0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	1
2	0	97	0	2	0	129	0
3	0	29	0	3	0	37	1
4	0	7	0	4	0	5	0
		Total:	133			Total:	173

Table 5.6: Changes in the distribution of realizations by definition and paradox occurrence

Some important observations from this table are the following:

- There appears to be significant symmetry in the cell values. The decreases in the cells where neither the average nor the maximum exhibit the paradox (e.g., the top block of rows, including -115 for definition #1 to #2 for two average intervals and one maximum) seem to match the increases for the realizations where the average satisfies definition #2. This makes sense because the paradox for definition #2 holds if for some capacity the average travel time exceeds the initial value for capacity = 0.
- Similarly, the decreases for the condition where only the maximum satisfies the paradox (the second block of rows) match with the condition where both the average and the maximum satisfy the paradox (the fourth block of rows). This also makes sense because in the second block the average did not satisfy the definition while in the fourth block it does.

- The changes from row block #1 to #3 and from row block #2 to #4 both appear to be caused by changes related to the average. This makes sense because it is the average and not the maximum that has numerous realizations for two or more intervals.
- The one notable exception to the above observations is for the difference between definitions #2 and #3 where there are two "-1" values for "Yes – Average / No – Maximum" and corresponding "+1" values for "Yes – Average / Yes – Maximum".
- The cells with the most significant changes involve two intervals for the average and one for the maximum (e.g., -115 for "No – Average / No – Maximum". As indicated before, these realizations will be studied in more detail next.

Table 5.7 shows the realizations involving two intervals for the average and one for the maximum broken down based on whether the intervals for the average were down/up or up/down and whether the interval for the maximum was down or up.

Counts of Paradox Combinations (Average and Maximum Travel Times) - 29,998 Total								
Paradox Definition #1			Paradox Definition #2			Paradox Definition #3		
Maximum Interval -> Average Intervals -v	Down	Up	Maximum Interval -> Average Intervals -v	Down	Up	Maximum Interval -> Average Intervals -v	Down	Up
No - Average / No Maximum								
Down Up	115	0	Down Up	0	0	Down Up	0	0
Up Down	140	0	Up Down	140	0	Up Down	0	0
	Total:	255		Total:	140		Total:	0
No - Average / Yes - Maximum								
Down Up	0	97	Down Up	0	0	Down Up	0	0
Up Down	0	129	Up Down	0	129	Up Down	0	0
	Total:	226		Total:	129		Total:	0
Yes - Avergage / No - Maximum								
Down Up	52	0	Down Up	167	0	Down Up	167	0
Up Down	183	0	Up Down	183	0	Up Down	323	0
	Total:	235		Total:	350		Total:	490
Yes - Average / Yes - Maximum								
Down Up	0	69	Down Up	0	166	Down Up	0	166
Up Down	0	172	Up Down	0	172	Up Down	0	301
	Total:	241		Total:	338		Total:	467

Table 5.7: Breakdown of the realizations involving two intervals for the average and one for the maximum

Several observations can be made based on this table.

- For the realizations that do not satisfy the paradox for either the average or the maximum, All the maximum intervals are "down" and all of the average intervals are either down/up or up/down.
- The same holds true for the realizations where the average satisfies the paradigm, but not the maximum.

- For the instances where the maximum satisfies the paradox, the maximum interval is "up". This makes sense, the maximum must increase to satisfy the paradox.
- The same holds true for the realizations where both the average and the maximum satisfy the paradox. This makes sense, again, because the maximum must increase to satisfy the paradox.
- As the definitions become more inclusive, realizations for which only the maximum satisfied the paradox migrate to realizations where both the average and the maximum satisfy the paradox.

Table 5.8 makes the definition-to-definition changes clear.

Incremental Changes					
Definition #2 - Definition #1			Definition #3 - Definition #2		
Maximum Interval -> Average Intervals -v	Down	Up	Maximum Interval -> Average Intervals -v	Down	Up
No - Average / No Maximum					
Down Up	-115	0	Down Up	0	0
Up Down	0	0	Up Down	-140	0
	Total:	-115		Total:	-140
No - Average / Yes - Maximum					
Down Up	0	-97	Down Up	0	0
Up Down	0	0	Up Down	0	-129
	Total:	-97		Total:	-129
Yes - Average / No - Maximum					
Down Up	115	0	Down Up	0	0
Up Down	0	0	Up Down	140	0
	Total:	115		Total:	140
Yes - Average / Yes - Maximum					
Down Up	0	97	Down Up	0	0
Up Down	0	0	Up Down	0	129
	Total:	97		Total:	129

Table 5.8: Changes in realization classifications as the paradox definition changes

As was the case for Table 5.6, there are similar observations about the information in Table 5.8:

- There is symmetry between the incremental changes (decreases) for the "No – Average / No – Maximum" block and the "Yes – Average / No – Maximum" block.
- In the case of definition #1 to #2, there is similar symmetry for the "No – Average / No – Maximum" block and the "Yes – Average / No – Maximum" block. For example, the "-115" in the "No – Average / No – Maximum" block for "down/up" and "down" is matched by the "+115" in the "Yes – Average / No – Maximum" block.
- Similar symmetry exists for the "No – Average / Yes – Maximum" block and the "Yes – Average / Yes – Maximum" block. For example, the "-97" in the "down/up" and "up" cell in

the “No – Average / Yes – Maximum” block is matched by the “+97” in the “Yes – Average / Yes – Maximum” block.

- In the case of definition #2 to #3, there is symmetry between the same pairs of blocks. For example, the “-140” in the “No – Average / No – Maximum” block for “up/down” and “down” is matched by the “+140” in the “Yes – Average / No – Maximum” block. And the “-129” in the “up/down” and “up” cell in the “No – Average / Yes – Maximum” block is matched by the “+129” in the “Yes – Average / Yes – Maximum” block.
- This again seems to be evidence that it is the classification of the paradox for the average travel time (yes or no) that is causing the shifts.

Table 5.9 looks at the trends for the average travel time, ignoring the trends for the maximum. The table organization is different from the earlier tables. The first two column blocks show classifications of the realizations based on whether the paradox was satisfied (on the right) or not (on the left). The individual columns show the number of descending intervals and the rows show the number of ascending intervals. For example, if the number of descending intervals is 2 and the number of ascending intervals is 2, then the total number of intervals is 4 (2+2). The right-most column block shows the incremental changes in the realization classifications as the paradox definition changes from #1 to #2 and then #3.

Average Travel Time - Counts of Realizations (29,998 total)													
No - Paradox Definition #1				Yes - Paradox Definition #1						Initial Increment			
Decending intervals --> Ascending intervals -v	0	1	2	Decending intervals --> Ascending intervals -v	0	1	2			Decending intervals --> Ascending intervals -v	0	1	2
0	22819	3764	0	0	0	0	0			0	0	0	0
1	0	481	86	1	2212	476	30			1	2212	476	30
2	0	25	28	2	0	49	28			2	0	49	28
		Total	27203			Total	2795					Total	2795
No - Paradox Definition #2				Yes - Paradox Definition #2						Definition #2 - Definition #1			
0	22819	3764	0	0	0	0	0			0	0	0	0
1	0	269	62	1	2212	688	54			1	0	212	24
2	0	0	9	2	0	74	47			2	0	25	19
		Total	26923			Total	3075					Total	280
No - Paradox Definition #3				Yes - Paradox Definition #3						Definition #3 - Definition #2			
0	22819	3764	0	0	0	0	0			0	0	0	0
1	0	0	0	1	2212	957	116			1	0	269	62
2	0	0	0	2	0	74	56			2	0	0	9
		Total	26583			Total	3415					Total	340

Table 5.9: Trends in the classification of realizations for the average travel time alone

Several observations are useful for this table:

- There are 2795 realizations that satisfy definition #1, 3,075 for definition #2, and 3,415 for definition #3.
- The 2212 realizations for 1 ascending interval and no descending intervals satisfy definition #1, as they should, and they carry forward to satisfying definitions #2 and #3.
- For definition #1 there are 583 realizations in four other cells that satisfy the paradox.
- For definition #2, an additional 280 realizations in those same four cells satisfy the paradox.
- For definition #3, an additional 340 realizations in those same four cells satisfy the paradox.

Table 5.10 looks at trends for the maximum travel time, ignoring the average. The table organization is the same as in Table 5.9.

Maximum Travel Time - Counts of Realizations (29,998 total)													
No - Paradox Definition #1				Yes - Paradox Definition #1						Initial Increment			
Decending intervals --> Ascending intervals -v	0	1	2	Decending intervals --> Ascending intervals -v	0	1	2			Decending intervals --> Ascending intervals -v	0	1	2
0	22819	3560	0	0	0	0	0			0	0	0	0
1	0	2	0	1	3617	0	0			1	3617	0	0
2	0	0	0	2	0	0	0			2	0	0	0
		Total	26381			Total	3617					Total	3617
No - Paradox Definition #2				Yes - Paradox Definition #2						Definition #2 - Definition #1			
0	22819	3560	0	0	0	0	0			0	0	0	0
1	0	2	0	1	3617	0	0			1	0	0	0
2	0	0	0	2	0	0	0			2	0	0	0
		Total	26381			Total	3617					Total	0
No - Paradox Definition #3				Yes - Paradox Definition #3						Definition #3 - Definition #2			
0	22819	3560	0	0	0	0	0			0	0	0	0
1	0	0	0	1	3617	2	0			1	0	2	0
2	0	0	0	2	0	0	0			2	0	0	0
		Total	26379			Total	3619					Total	2

Table 5.10: Trends in the classification of realizations for the maximum travel time alone

The observations for this table are quite different from those for Table 5.8:

- There are 3617 realizations that satisfy definition #1, the same for definition #2, and 3,619 for definition #3. Hence, only two realizations are added in going from definition #2 to #3 and none are added going from #1 to #2.
- These 3,617 realizations arise for conditions where there is one ascending interval and no descending interval.
- Two additional realizations satisfy definition #3 that did not satisfy either #1 or #2. They arise for the condition where there is one ascending interval and one descending interval.

Observations based on Tables 5.9 and 5.10 combined are as follows:

- The percentage of realizations for which the maximum travel time satisfies the paradox is larger than that of average travel time, for all three paradox definitions.
- For all definitions of paradox, if the paradox occurs, there is at least one increasing interval.
- The maximum travel time has simpler trend patterns. Only two of the 29,998 realizations have more than one interval.

5.4 SUMMARY

This section has presented a study of Braess' paradox in the context of multi-class traffic assignment. The material is relevant for two reasons. First, the combined auto-truck traffic assignment problem is multi-class. And, this section presents a novel way to address that situation where one class pursues the system optimal solution for the traffic assignment (the autos) and the other pursues the user-optimal solution (the trucks). Second, the illustration, and condition, are relevant because pricing is an effective way to combat the occurrence of the paradox. And, simple rules about how to price the use of the arcs can lead to "poor" traffic assignment solutions. Ones that could be improved, in terms of one or both objectives, through a more thoughtful pricing structure.

Ultimately, as connected and automated vehicles become more prevalent, and the route guidance devices in the vehicles become more dependent on and informed by real-time information about the network's current and future conditions, these pricing structures will be used by the network operators to guide the arc loading in a desirable direction. So, examining these pricing issues in a setting where the network can perform "badly" if the prices are not set "right", provides useful insights about how prices should be set, and the network should be controlled, so that its performance is "the best possible".

6.0 SCHEDULED SYSTEM MANAGEMENT

The currently rapid innovations and developments of transportation system intelligence in multi-source sensing and information sharing continuously generates huge volumes of various data and information for planners and managers to better observe time-varying traffic conditions and accordingly propose adaptive travel demand management and supply (capacity) control strategies. However, the data sparsity problem still exists, because it is impossible to install fixed sensors on each link or to cover all links by point-to-point moving sensors. As a result, it surely requires new methodologies to recover the system-wide transportation conditions based on the limited observations (Zheng et al., 2014). It leads to a fundamental question; that is, how well the time-dependent transportation system states can be estimated or observed based on currently available heterogeneous source data. In other word, the research motivation is how to quantify the transportation system observability for further optimal control and policy making.

Specifically, as shown in Figure 6.1, the continuous multi-source sensor data that include fixed sensor data (loop detector, video imaging processing data), mobile sensor data (GPS trajectory data, mobile phone data, AVI, AVL, AFC) and social media with useful travel information, provide a heterogeneous information flow to transportation systems for observing different level of system states (flow, density, travel time), covering from link level, route level, OD pair level to the whole network level, which are usually evaluated by transportation system managers under different goals for further operation, planning and policy making as the fundamental inputs. To reach the desired transportation system performances, it requires to actively manage the demand and supply sides in corresponding level of requests as follows. Travel cost (pricing/incentive), route choice (route guidance, travel reservation), departure time choice (transit station entry flow control, travel reservation), and total demand control (rationing, license plate lottery) are usually considered in the demand management side. On the other hand, the available transportation supply resources are optimized or increased, for example, adaptive signal control, dynamic lane use control, dynamic speed limit control, transit vehicle update and rescheduling, new infrastructure constructions. Meanwhile, the feedback loop between system states and optimal control keeps moving forward along the time horizon with any external new disturbances, such as, weather, incidence, special events, new land use, population changes, etc.

As pointed out by Daganzo (2007), the most recent computer transportation models can theoretically predict almost anything but not in practice, because it is difficult to obtain accurate dynamic origin-destination matrices, capture travelers' unpredictable gaming behaviors, and catch the hypersensitive response of congested networks to any mini inputs. However, the increasingly available multi-source big data and powerful computation capability are creating great opportunities to continuously provide accurate model inputs, capture individually complicated travel behaviors, and efficiently calculate the outputs. For example, the AFC data, mobile phone data and travel request apps greatly improve the accuracy of observed OD travel demand, and the GPS-enable devices enable us to record each passenger/vehicle's high-resolution trajectory in real time. In the meantime, it should be not ignored that the overwhelming volume of data are also incurring new challenges on data use and transportation modeling.

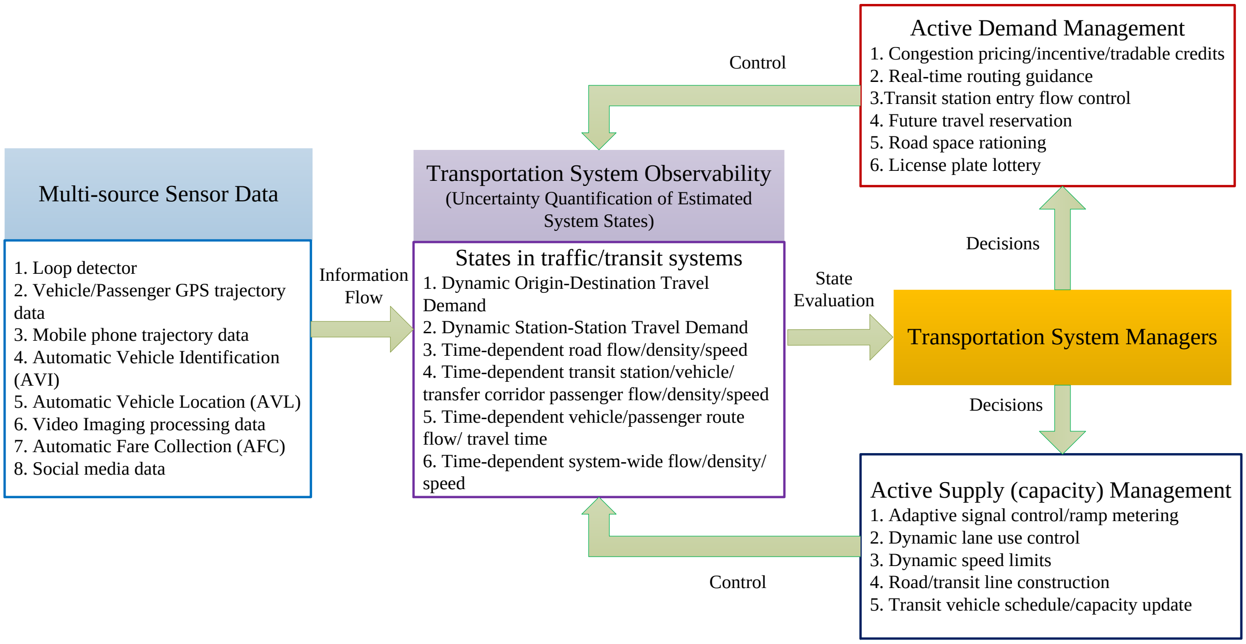


Figure 6.1: Transportation system management under multi-source sensor data

In the data use side: (i) Is the big data useful enough and what is the value of the data? (ii) Under what goals, one kind of data is more useful that others? (iii) How to fuse multi-source data to keep observation result consistency? Meanwhile, in the transportation modeling side: (i) how to mathematically represent the available multi-source information so that different system states can be estimated in a unified modeling framework; (ii) how to model the exact inner relation between the information and some interested system states, (iii) how to quantify the system observability (uncertainty of estimated states) for further optimal control and management, and (iv) how to design efficient and scalable algorithms for solving those models. Motivated by the general challenges above, this research aims to explore the theoretical relation among sensor data, system states, and system observability in transportation systems by proposing insightful analysis and theoretically sound linear programming models, especially taking the urban rail transit system as a starting point.

6.1 STATE ESTIMATION AND SENSOR NETWORK DESIGN

Observability is a concept introduced by Kalman (1959) for linear dynamic systems in control theory. It is a measure for how well internal states of a system can be inferred by knowledge of its external outputs. In other words, it aims to quantify or measure the uncertainty of estimated internal states based on the available external observations under a given sensor environment with sampling errors, sensor error and model errors. A comprehensive literature review can be found at the paper by Castillo et al. (2015). As for evaluating the estimation uncertainty or accuracy, origin-destination (OD) trip matrix estimation is a widely studied classical problem due to its under-determination attribute, which means that there is an infinite number of OD trips that can generate link flows consistent with the observations. Yang et al. (1991) first introduced the concept of Maximum Possible Relative Error (MPRE) to theoretically investigate the estimation uncertainty and reliability of the OD estimated trips obtained by the entropy model. Bianco et al. (2001) further explored the accuracy of estimated OD matrix bound under different sensor location strategies. In addition, Bierlaire (2002) proposed the novel concept of total demand scale as a new measure to examine the quality of estimated OD trip tables from link counts, by maximizing/minimizing the total travel demand satisfying all observations. In the general transportation observability problems, many studies (Castillo et al., 2007; Castillo et al., 2008; Gentili and Mirchandani, 2012) modeled the problems as a system of linear equations and/or inequalities and then determine whether the system or one unknown variable is observable or not by analyzing the properties of its coefficient matrix.

Meanwhile, in the system of linear inequalities, a general bound of unknown variables can be derived through the dual cone approach. In general, the observability problem more cares about the list of variables to be observed rather than the specific system states uncertainty ranges.

To increase the estimation quality, the integration of ubiquitous sensor network design and state estimation has received growing attentions in the past 2 decades. Yang and Zhou (1998) proposed integer linear programming models and four sensor location rules to determine the optimal number and location of point sensors for origin-destination matrix estimation. Based on the trace of the a posteriori covariance matrix produced in a Kalman filtering model, Zhou and List (2006) offered an information-theoretic framework for locating fixed sensors in the traffic OD demand estimation problem. In addition, based on the observability problem definition, the optimal count number and location of active sensors (Gentili and Mirchandani, 2005) and counting and scanning sensors (Castillo et al., 2012) for estimating path/link flows are studied by analyzing a set of linear equations. Hu et al. (2009) proposed one “basis link” method to find the smallest subset of links in a network to locate sensors so that the traffic flow of all links can be accurately estimated under steady-state conditions. Further, Ng (2012) introduced a new solution approach (‘‘synergistic sensor location’’) to avoid possible path enumeration under the assumed steady-state conditions. Xu et al. (2016) proposed a robust network sensor design to completely observe link flows whiling accounting for the accumulation of observation errors.

For other traffic states, link travel time estimation errors are commonly selected as the optimization criterion for point sensor location problems (Ban et al., 2009; Danczyk and Liu, 2011), and a reliable sensor location method is proposed to consider probabilistic sensor failures (Li and Ouyang, 2011). Herrera et al. (2010) developed a real-time traffic monitoring system by using vehicles with GPS-enable mobile phones and suggested that 2-3% penetration rate of mobile phones can provide accurate traffic speed estimations. Based on a Kalman filtering structure, Xing et al. (2013) developed measurement and uncertainty quantification models to explicitly consider several important sources of errors in path travel time estimation/prediction. In the real-time traffic situations, Eisenman et al. (2006) conducted a sensitivity analysis of estimation and prediction accuracy under different sensor locations and coverage scenarios based on a real-time dynamic simulation system, DYNASMART-X. Boyles and Waller (2011) studied the optimal location selection for providing the real-time traffic information to drivers with the adaptive travel behavior by proposing heuristic algorithms, and Ban et al. (2011) studied the real-time queue length estimation at signalized intersections by focusing on queuing delay patterns and queue length changes based on travel times from mobile traffic sensors.

Under the framework of urban computing, Zheng et al (2014) reviewed the related date-mining and machine-learning approaches used in transportation system state estimation and prediction. Thiagarajan et al. (2009) applied the WiFi signals to estimate route travel time and identify delay-prone segments by using a hidden Markov model (HMM)-based map matching scheme. Focusing on the GPS trajectory data, the normal traffic patterns are mined for real-time city-wide travel time estimation and prediction by building landmark graphs (Yuan et al., 2011) and for estimating/detecting the traffic anomalies based on the representative terms from twitter-like social media data (Pan et al., 2013). Zhang et al. (2015) proposed a context-aware tensor factorization model to estimate the time spent of each GPS-equipped taxicab on gas stations with consideration of gas price, brand, and weather conditions. Tang et al. (2016) proposed a novel time-dependent graph model to estimate the most likely space-time paths of vehicles with point-to-point data and then implemented a dynamic programming algorithm for the offline and online map-matching applications. Using crowdsourced data from location-based service apps, Zhao and Zhang (2017) examined the individual dynamic choices of activity chains by proposing a data-driven Markov chain approach in activity-travel space-time-state network. Recently, Zhu et al (2018) developed a data-driven link-based network sensor location model to maximize the travel time information gain with accounting for the uncertainty in the prior travel time distribution. Wu et al. (2018) proposed a novel deep-learning-based framework to simultaneously estimate static OD demand and road traffic conditions based on multi-source sensor information by developing a transportation computational graph tool.

6.2 TRANSIT SYSTEM CONSIDERATIONS

In urban transit systems, usually, the automatic fare collection system (AFC) or smart card usually records both the time and station for entry and exit for each passenger in rail transit systems, but only the boarding time and stop and route number normally can be reported in bus transit systems. A comprehensive literature review about smart card data use can be found in the papers (Pelletier et al., 2011; Ma et al., 2013). Obviously, the unknown destination information greatly increase the state uncertainty of bus transit system. Trépanier et al. (2007) estimate the alighting point for each passenger based on the smallest distance to the boarding stop of his/her next route from individually continuous riding records in smart card. Seaborn et al. (2009) proposed maximum elapsed time thresholds to identify transfers for bus-to-underground, underground-to-bus, and bus-to-bus to identify and assess multi-modal trips in London. Meanwhile, Munizaga and Palma (2012) estimated a multimodal transport OD matrix from smartcard and GPS data whiling consider unobserved trips by expansion factors in Santiago, Chile. Yuan et al. (2013) proposed a space alignment approach by aligning the monetary space and geospatial space with the temporal space to infer each passenger’s trajectory and the results improve the detection of uses’ home and work places. Nassir et al. (2015) applied the smart card data to detect activity and identify transfers to estimate the true origins and destinations. Nunes et al. (2016) further proposed four endogenous spatial validation rules to enhance the accuracy of estimated passenger destination choice. Alsger et al. (2016) evaluated and improved existing OD estimation method according to available OD information and assessed the previous last destination assumptions in bus transit systems. Under the situation that passenger’s boarding stop information is not record in smart cards, Ma et al. (2012) developed a Markov chain based Bayesian decision tree algorithm to estimate the sequential stops on the bus route and then match those stops with the recorded boarding time to infer passengers’ origin. Further, Ma et al. (2015) improved their previous algorithms to increase the estimation accuracy and computation efficiency. In addition, in case that buses don’t have Automatic Vehicle Location data, Zimmerman et al. (2011) developed a system named Tiramisu that can estimate and predict the real-time bus arrival time by applying the crowd-sourcing data from commuters sharing their GPS-enabled mobile phones.

Depending on the available OD travel information from smart card in urban rail transit systems, many studies focus on the route choices and transfer patterns, which can be viewed as different system states required for estimation. Kusakabe et al. (2010) focused on the passengers’ train choice behavior by assuming that each passenger aims to minimize the total waiting time at the departure station, loss time at the arrival station, and the transfer frequency. Zhao et al. (2007) chose the logit discrete choice model, but the tight side constraints (e.g. strict vehicle capacity constraint) are still hard to include. Ceapa et al. (2012) mined the regular spatial-temporal trip relations from AFC data to estimate and predict the crowding level for providing more accurate personalized trip planning services. Sun and Xu (2012) estimated the path choice based on the observed overall probability density of journey time and the derived distribution of individual path travel time from the rail transit smart card. In addition, Zhou and Xu (2012) used a matching degree function value to assign the trip to the most likely path based on derived boarding plan of path. Remarked that the verification of those assigned path flows above has not been soundly performed due to the limit of observed data and complicated route choice behaviors, such as passenger's travel preference. Kusakabe and Asakura (2014) proposed a data fusion methodology to consider both the smart card data and person trip survey data by Bayes probabilistic model to estimate behavioral attributes of trips in the smart card data. Based on passenger OD matrix information and vehicle stop time and location data, Zhu et al. (2017a; 2017b) proposed probabilistic models to estimate the individual train loads, left behind probabilities, time-dependent crowding levels at stations under tight vehicle capacity considerations. In addition, Nair et al. (2013) focused on a large-scale bicycle sharing system and analyze the connection between bicycle usage and public transit systems.

6.3 CONCEPTUAL ILLUSTRATION

Through applying some concepts from game theory and control theory into the problem (LaValle, 2012), a state space is defined as the all possible internal system state based on the external physical transportation world, and an information space is a place where the internal states live when available information is involved. A state is specifically defined and can be associated with the available information. As shown in Figure 6.2, the information space is formed by the available information, and the states are tightly connected by different projection functions, which mathematically define the states according to the managers' needs. The bound among all possible states represents the state uncertainty under current available information.

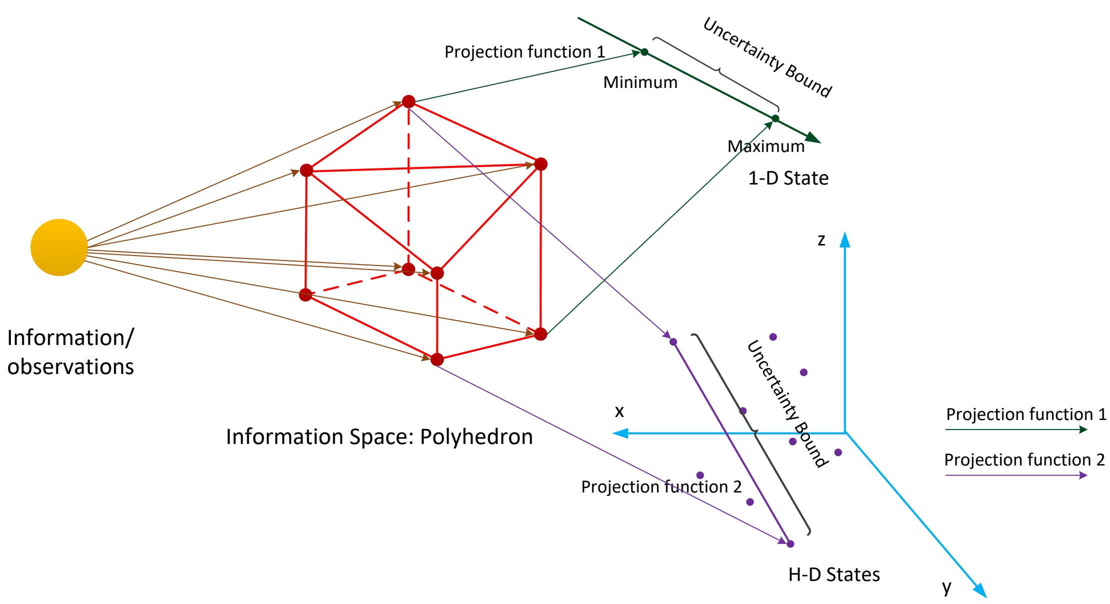


Figure 6.2: Relation among Information, Information Space, and Flexible States

When the information space is generated as one single point, it can state that the system is observable; otherwise, a non-empty set or space leads to unobservable or partially observable system. In this section, the connection between the internal states and the external states (observation or information) is built by information space as a bridge or communication channel, and further quantify the corresponding uncertainty of states defined by users. How to design sensor network configurations to alter information space and further increase the state estimation accuracy will be addressed in the future research.

For illustrative purposes, Figure 6.3(a) depicts a simple transportation network with four nodes and five links. The link travel time and capacity are also provided as physical network attributes. Let x_1, x_2 and x_3 represent the path flow on paths 1, 2 and 3. Based on the tight capacity constraints, the following relation can be obtained: $0 \le x_1 \le 2$, $0 \le x_2 \le 3$, and $0 \le x_3 \le 1$, which defines the system state space shown as a blue cuboid in Figure 3.3(b).

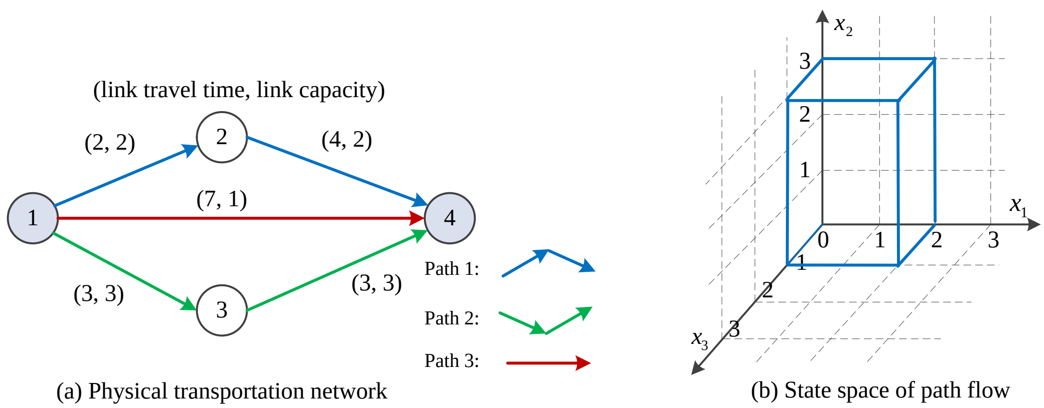


Figure 6.3: An Illustrative Transportation Network and Its State Space

Assume that the OD information is available through survey that there are four vehicles departing from node 1 to node 4. Then, one corresponding constraint will be $x_1 + x_2 + x_3 = 4$. Figure 6.4(a) displays the information space as the intersection of the red triangle and the blue cuboid based on the available OD information. Two scenarios are designed as follows to analyze the relation between system states and available information.

- **Scenario 1**: Assume that there is one flow count detector on link $(2,4)$ and its link count is 1. The relation gets updated as follows: $1 + x_2 + x_3 = 4$, $0 \le x_2 \le 3$, and $0 \le x_3 \le 1$, so the corresponding information space is reduced to be the intersection of the red triangle and the green rectangle shown in Figure 6.4(b).
- **Scenario 2**: Suppose that the automatic vehicle identification (AVI) detectors are available at nodes 1 and 4. One vehicle's travel time is observed as 7min. Since only path 3's travel time is 7 and its capacity is just 1, it implies that path flow $x_3 = 1$. As a result, the relation changes as follows: $x_1 + x_2 + 1 = 4$, $0 \le x_1 \le 2$, and $0 \le x_2 \le 3$. The corresponding information space becomes the intersection of the purple rectangle and the red triangle displayed in Figure 6.4(c).

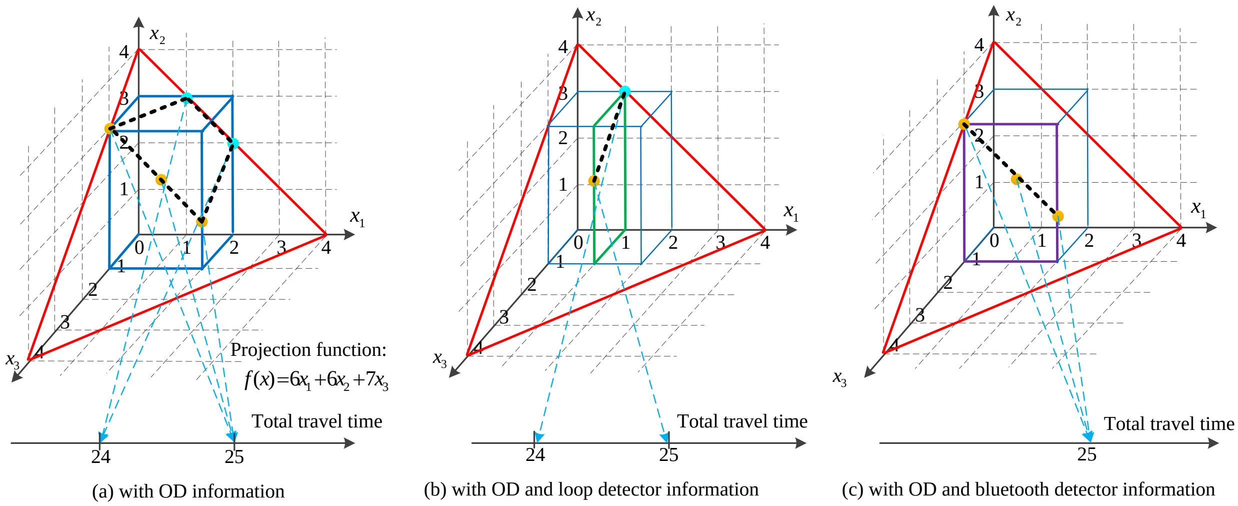


Figure 6.4: Information Spaces and Its Projected Bound based on Available Information

As shown in Figure 6.4, the information spaces are generated as polyhedrons based on different available information. A projection function is defined to map the information space into one dimension state (total travel time). It should be remarked that, the projection functions could be defined based on different goals interested by the system managers, such as, the total system travel time, the number of vehicles in one area, etc. In Figure 6.4(a), one projection function is defined as $f(\boldsymbol{x}) = 6x_1 + 6x_2 + 7x_3$, which means that the total system travel time is the analysis goal. Accordingly, different optimization models are solved by maximizing and minimizing $f(\boldsymbol{x})$ subject to different information spaces.

As demonstrated in Figure 6.4(a), there are five feasible integer solutions in the information space and the bound of total travel time formed by projection is [24,25]. When the link count data are added in Figure 6.4(b), the information space by integer solutions is reduced, but the projected bound is still the same. It indicates that the new information from link count doesn't contribute to reduce the uncertainty of this state estimation, even though a smaller information space is generated in Figure 6.4(b). It reminds us that the value of "big data" is determined by not only "big volume" but also the usefulness of information to the system.

In addition, the point-to-point Bluetooth data (end-to-end passenger id detector data) in Figure 6.4(c) makes the projected bound converged to be one unique value, which implies that the point-to-point data is more powerful than the point detector for increasing the observability of system travel time in this case. Therefore, evaluating the values of different information should be based on which states the system manager really cares. One information that seems worthless for one goal may be much useful for other state estimation applications. Moreover, the information space in Figure 6.4(c) is bigger than that in Figure 6.4(b), but the uncertainty of state estimate in Figure 6.4(c) is 0 and less than that in Figure 6.4(b). Thus, the volume of information space might not be the best criteria to judge the bound of state estimate uncertainty or system observability.

Except for those information from physical sensors above, the previous travel experiences or currently published traffic information from transportation agencies could also take important roles in quantifying the state uncertainty for managers' further actions. For example, if everyone has perfect information over the network attributes based on their experiences and each one aims to find the best route for his/her trip, which is usually entitled Wardrop's first principle, the information space will be redefined as, $x_1 + x_2 = 4, 0 \leq x_1 \leq 2, 0 \leq x_2 \leq 3$. Compared with the two scenarios above, the information space could be further reduced by this assumed and (potentially questionable) travel behavior. Therefore, one accurate travel behavior can also provide much rich information to determine system states, and this could help us understand why many studies focus on travel behavior estimation to better understand the system, from the information space perspective.

As a remark, the information space concept in game theory and control theory for storing the amounts of ambiguity in state also may have a connection with Shannon's information theory using entropy-based constructs. For example, the entropy-maximization approach is usually used to represent the most likely traffic state, a kind of stable equilibrium state, for origin-destination travel demand estimation problems. However, in information space theory, the system state is unknown and gradually constructed based on available information. It is possible that the system state derived from the information space is finally the same as the equilibrium state based on maximum entropy.

6.4 PROBLEM STATEMENT

Table 6.1 lists the general indices, sets, parameters and variables in the proposed models.

Table 6.1: Indices, sets, parameters and variables

Indices	Definition
i, j	Index of nodes, $i, j \in N$
(i, j)	Index of physical link between two adjacent nodes, $(i, j) \in L$
a	Index of passenger group, $a \in A$
$o(a)$	Index of origin node of group a
$d(a)$	Index of destination node of group a
t, s	Index of time intervals in the space-time network
τ	Index of time period for the observed passenger flow
p	Index of paths, $p \in P$
r	Index of transit companies
Sets	
N	Set of nodes in the physical transit network
L	Set of links in the physical transit network
A	Set of passenger groups
V	Set of vertices in the space-time network
E	Set of edges/arcs in the space-time network
G	Set of time periods for the observed passenger flows
$S_{p,a}$	Set of paths p of group a
$G(i, j, \tau)$	Set of arcs on observed link (i, j) at time period τ
Parameters	
β_1, β_2	The weights on target passengers' trip time and observed link/arc flows, respectively
μ_a	The observed aggregated trip time of group a from smart card data
$\mu_{i,j,\tau}$	The observed aggregated passenger count on link (i, j) during time period τ
w_p	The travel time of path p
c_p^r	The earning collected on path p of transit company r

$Cap_{i,j,t,s}$	Capacity of traveling arc (i,j,t,s) in the space-time network
DT^a	The departure time of group a
AT^a	The assumed arrival time of group a
D_a	The number of passengers in group a
$c_{i,j,t,s}$	Travel cost of traveling arc (i,j,t,s) in the space-time network
T	The time horizon in the space-time network
$\delta^{p,a}_{(i,j,t,s)}$	Path-link incidence index of route p of group a on arc (i,j,t,s)
w^p	The path travel time of path p
Variables	
$x^a_{i,j,t,s}$	The number of passengers in group a is assigned on traveling/waiting arc (i,j,t,s) in the space-time network
$\theta_a, \theta_{i,j,\tau}$	Continuous positive deviation variables for group a's trip time and link (i,j) during time period τ, respectively
x^p_a	The number of passengers of group a choosing their feasible path p
μ^*_a	The preprocessed aggregated trip time of group a from smart card data
$\mu^*_{i,j,\tau}$	The preprocessed aggregated passenger count on link (i,j) during time period τ

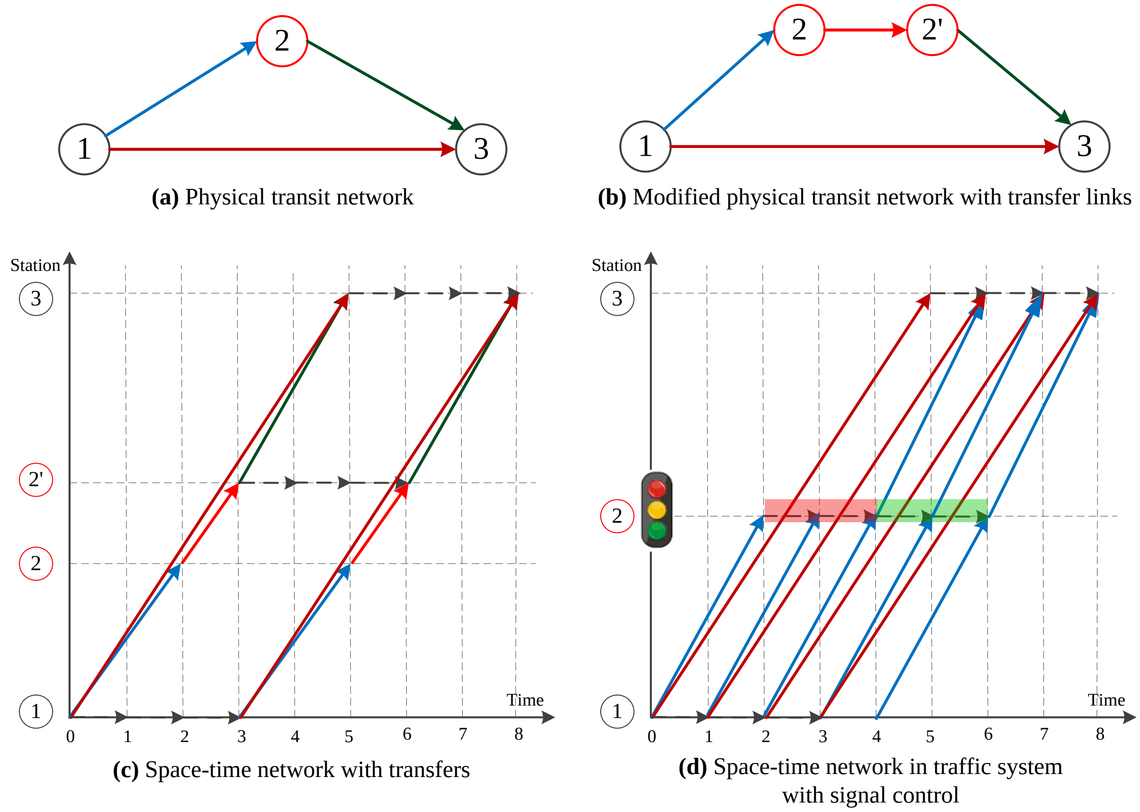
6.4.1 Space-time Network Construction

To properly account for the evolution of system dynamics over time, Ford and Fulkerson (1958) first introduced dynamic network flow models to solve the dynamic maximum flow problem in time extended networks. The space-time network flow models are then widely used in dynamic transportation systems, such as, dynamic system optimal with a point queue model (Zawack and Thompson, 1987), dynamic user equilibrium with a spatial queue model (Drissi-Kaıtouni and Hameda-Benchekroun, 1992), dynamic system optimal with departure time, route choice and congestion toll (Yang and Meng, 1997), dynamic user equilibrium with link travel time functions (Chen and Hsueh, 1998), activity-based dynamic user equilibrium (Lam and Yin, 2001). Recently, in order to maximize network accessibility, Tong et al. (2015) proposed a space-time network flow model with binary decision variables, which actually derives a number of agent-based models in space-time networks later.

There are a number of studies providing how to construct specific time-expanded networks for different transportation systems, such as, freeway network (Lu et al., 2016), road network with signal settings (Li et al., 2016), urban transit network (Liu and Zhou, 2016), bike-sharing network (Lu, 2016), road network with activity requests (Liu et al., 2017), and vehicle trajectory network (Wei et al., 2017). This section considers a physical urban rail transit network with a set of nodes (stops/stations) N and a set of links L as a starting point. Each link can be denoted as a directed link (i,j) from upstream node i to downstream node j. A deterministic transit schedule is supposed to be obtained from Automatic Vehicle Location (AVL) data from vehicle tracking systems. A space-time network is then constructed, where V is the set of vertices and E is the set of arcs. Node i is extended to a set of vertices (i,t) at each time interval t in the study horizon, $t = 1,2,\dots,T$, where T is the length of the optimization horizon. The transit schedule from node i to node j from time t to time s can be represented by a travelling arc (i,j,t,s) where $(s-t)$ is the exact scheduled/running link travel time and should be integer multipliers of one time interval. The capacity of travelling arcs can be viewed as the transit vehicle's carrying capacity. In addition, a waiting arc is built from (i, t) to $(i, t+1)$ at node i with waiting time of one time unit and its capacity is defined as the station/platform storage capacity.

In urban rail transit systems, individual passengers should have a trip record with origin, departure time, destination and arrival time from the smart card. However, transit agencies may just provide aggregated trip data for groups of passengers. Each group a with D_a passengers has a departure time DT^a at origin node $o(a)$ to its destination node $d(a)$. At each destination node, there is one assumed large arrival time T for all groups so that the following proposed model will be one-origin-one destination problem in the space-time network. It should be noted that the travel cost of waiting arcs on the destination node is 0, which means that once the passengers in a group arrive at the destination, the waiting cost to the super-destination (at larger arrival time T) is 0. Finally, the estimated trip time in the model should be equal to the observed trip time, which will be presented in the following sections.

In addition, one transfer node can be divided as multiple nodes, depending on how many transit lines intersect at this node. One illustrative example is shown in Figure 6.5(a) where two lines intersect at node 2 and make it as a transfer station. Then node 2 is split to node $2'$ and node $2"$ and the modified physical network is drawn in Figure 6.5(b). The travel time of transfer links could be the actual walking time, and its capacity is the maximum passenger throughput at transfer corridors. As a remark, based on the maximum transfer distance accepted by passengers, it is possible to connect different stops by transfer links or extended to multimodal networks. Figure 6.5(c) shows the transfer process where all transfer time is assumed to be one time unit. In addition, it is also feasible to consider the uncertainty of walking time on transfer links or from station entry to the platform through constructing more service/travelling arcs with different arc travel times. Furthermore, in traffic networks, the road can provide its service at each time interval with a specific arc capacity and the signal timing rules whether those service arcs are open or closed, which is represented in a space-time network in Figure 6.5(d) to show the unified modeling framework.



(a) Physical transit network

(b) Modified physical transit network with transfer links

(c) Space-time network with transfers

(d) Space-time network in traffic system with signal control

Figure 6.5: Information space-time network construction

6.4.2 Information space generation based on multi-source sensor data

Information space (LaValle, 2012) works as a communication channel to connect the external physical world and the internal system states. The external physical world is sensed by heterogeneous sensors in terms of different observations or information, which finally forms a corresponding information space. Meanwhile, the internal system states are reflected through the information space based on the specific state definitions.

In addition to the physical transit lines and schedules as useful information, the possible observations in the urban transit systems are summarized in **Error! Reference source not found.**.

Table 6.2: Available trip information in urban transit systems

Bus transit system	Rail transit system
(1) origin, boarding time, destination, alighting time for each person in individual bus	(1) origin, entry time, destination, exit time for each person or aggregated for each passenger group in transit network
(2) origin and boarding time for each person in individual bus	(2) origin and boarding time for each person or aggregated for each passenger group in transit network
(3) historical time-dependent OD information for transit networks	

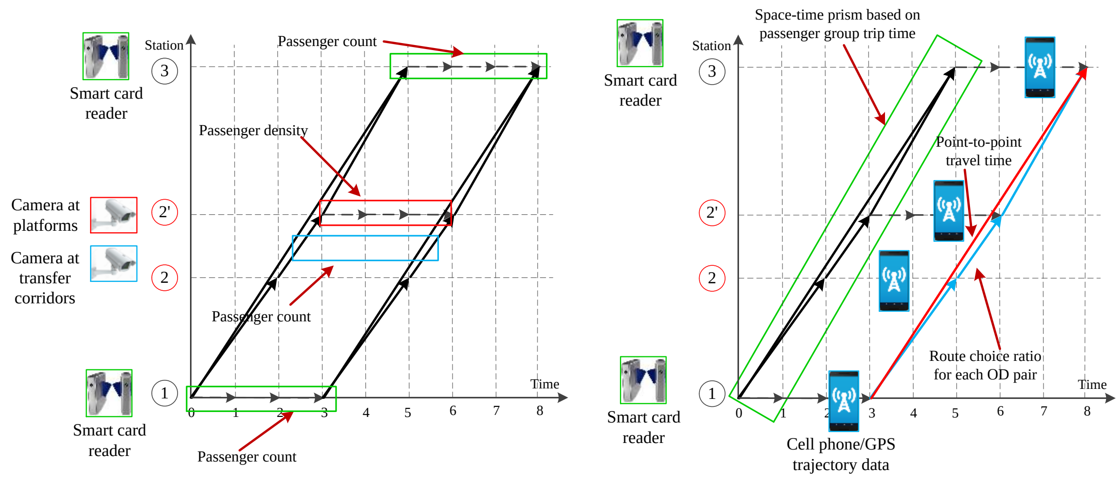
In bus transit systems, (i) when the origin, boarding time, destination, alighting time for each person in individual buses are available from smart card data, the state in each vehicle can be completely observed when the transit schedule or transit vehicle trajectory is available from AVL data, and that accurate trip information can provide great values for the operational transit planning. (ii) If only the origin, and boarding time can be recorded, algorithms are needed to estimate the individual destination. Usually it is estimated as the nearest stop to the boarding stop of traveler's next route based on the continuous riding records, so the corresponding alighting time will be also available. However, if travelers just have a single transit trip, it is still difficult to estimate the destination, which causes a large uncertainty for state estimation. (iii) If the market penetration rate of smart card is very low or the goal is for operational transit planning, the historical time-dependent OD information has to be used to perform transit network assignment with assumed travel behaviors (Szeto and Jiang, 2014; Jiang and Szeto, 2016; Cats et al., 2016; Liu and Zhou, 2016; Codina et al., 2017), which could create a larger uncertainty in the system and needs to be carefully calibrated and validated by real-world survey and observations.

In rail transit systems, (i) the origin, entry time, destination, exit time usually are available for each passenger or group, but the path/vehicle/transfer selection in the network level still has a large uncertainty. (ii) When the destination and exit time to stations are not recorded, the system uncertainty will be increased more.

In addition, with the development of sense technologies, more available sensor information from big data applications can be used in the transit systems.

i. Video data processed to gain the aggregated passenger flow at key points during specific time periods, such as, transfer corridors, the entry and exit of stations, or the stop/platforms.
ii. Cellphone/GPS based point-to-point trajectory data. A general path choice ratio bound is available when the penetration/sample rate of cell phone used as sensors is big enough. The granularity of the trajectory points is highly depended on the cell tower locations. Also, Bluetooth data can provide a point-to-point travel time and general path choice ratio.
iii. General travel behavior data (e.g. preference) through survey. It can provide the path choice of some specific passengers, so the path choice uncertainty of all passengers can be reduced, to some extent.
iv. Social media: it can share the location information, users' interests, and events happening around them, which could be used to mine the human mobility pattern and predict the travel demand related to those events.

Taking the transit network as an example, the possible observations are illustrated in Figure 6.6(a) and (b) for fixed sensors and mobile sensors, respectively.



(a) Observations from fixed sensors **(b)** Observations from mobile sensors

Figure 6.6: Observations from multi-source sensors

The specific modeling on generating information space based on those available sensor data is developed as follows. It should be remarked that, the first-in-first-out (FIFO) rule is not incorporated in the proposed space-time network, because it can be violated in transit networks and the multi-source information can better present travelers' decision by reducing the feasible space of information space. Also, readers who are interested in FIFO in space-time networks can refer to the details in the paper (Shang et al., 2018).

Taking the rail transit system as the modeling example, following formulation can be developed. According to the physical network, transit schedule, and dynamic *OD* information from smart card data, the standard flow balance constraint can be given as

$$\sum_{i,t:(i,j,t,s)\in E} x^a_{i,j,t,s} - \sum_{i,t:(j,i,s,t)\in E} x^a_{j,i,s,t} = \begin{cases} -D_a & \forall a, j = o(a), s = DT^a \\ D_a & \forall a, j = d(a), s = T \\ 0 & otherwise \end{cases} \tag{6-1}$$

Strict vehicle and station platform capacity constraint

$$\sum_a x^a_{i,j,t,s} \le Cap_{i,j,t,s},\ \forall (i,j,t,s) \in A \tag{6-2}$$

As stated previously, the estimated trip time of each group in the model should be consistent with the observation (average trip time of each group) from smart card.

$$\sum_{(i,j,t,s)} (x^a_{i,j,t,s} \times c_{i,j,t,s}) = D_a \times \mu_a, \forall\ a \tag{6-3}$$

Estimated aggregated passenger flow count on link (i,j) during time period τ is expected to be the observation from video data or counting by person.

$$\sum_a \sum_{t\in\tau} x^a_{i,j,t,s} = \mu_{i,j,\tau}, \forall (i,j,\tau) \tag{6-4}$$

(v) Non-negative arc flow variables

$$x_{i,j,t,s}^{a} \geq 0 \quad (6\text{-}5)$$

Note that if a passenger is viewed as a group of passengers and $x_{i,j,t,s}^{a}$ will be a binary variable, the above modeling is still available and an agent-based model arises. Usually, agent-based trajectory data can provide more point-to-point travel time information rather than just path choice.

The above presents an arc-based formulation for constructing the information space. In comparison, a path-based formulation can be offered in the following based on the feasible path generation.

Flow balance constraint:

$$\sum_{p} x_{a}^{p} = D_{a}, \forall a \quad (6\text{-}6)$$

Capacity constraint:

$$\sum_{(p,a) \in S_{(p,a)}} (\delta_{(i,j,t,s)}^{p,a} \times x_{a}^{p}) \leq Cap_{i,j,t,s}, \forall (i,j,t,s) \in A \quad (6\text{-}7)$$

Trip time constraint:

$$\sum_{p} x_{a}^{p} * w^{p} = D_{a} \times \mu_{a}, \forall a \quad (6\text{-}8)$$

Aggregated passenger flow count constraint:

$$\sum_{a} \sum_{t \in \tau} (\delta_{(i,j,t,s)}^{p,a} \times x_{a}^{p}) = \mu_{i,j,\tau},\ \forall (i,j,\tau) \quad (6\text{-}9)$$

Non-negative path flow:

$$x_{a}^{p} \geq 0 \quad (6\text{-}10)$$

Given the time-expanded space-time network constructed in subsection 3.1, the general feasible path set for a passenger group with specific OD pair and a departure time can be generated by a forward label correcting algorithm from the vertex (origin and departure time) to its destination node based on the observed trip time of that group as a prism.

It should be remarked that when considering the bus transit systems, the smart card data usually only have the origin and departure time without passengers' destination and arrival time information. To model this condition that dynamic OD trips are unknown, D_{a} will be a variable in equation (1) and the summation of D_{a} with same origin and departure time should be equal to the recorded total trip production at this origin and departure time from smart card data. If the structure of each OD pair with departure time is given based on the historical OD information, the number of unknown OD variables will be greatly reduced. In addition, from the data mining perspective, interesting readers can refer to the paper (Ma et al., 2013) to find studies related to estimate destination probabilities.

6.4.3 Dantzig-Wolfe decomposition for special flow-balance blocks

As shown in the arc-based and path-based formulation in section 3.2, the flow balance constraint is a special block that can be solved by classical shortest path algorithms and further be incorporated by Dantzig–Wolfe decomposition. Actually, this method has been adopted for static traffic assignment (Larsson and Patriksson, 1992), side constrained traffic equilibrium (Larrson et al., 2004), time constrained shortest path problem (Desrosiers and Lubbecke, 2005), etc. The advantage of this decomposition allows to solve the special blocks in parallel via independent computation threads to address large-scale networks, especially when the computer hardware has a rapid development in current days. It also has the re-optimization capability if the travel demand, arc performance function or network topology has any changes in future (Larsson and Patriksson, 1992). Specifically, Dantzig–Wolfe decomposition is originally proposed by Dantzig and Wolfe (1960) for solving linear programming problems with special structure. A general primal linear program can be represented as: min $c^T\boldsymbol{x}$, subject to, $A\boldsymbol{x} \le \boldsymbol{b}$, $D\boldsymbol{x} \le \boldsymbol{d}$, and $\boldsymbol{x} \ge 0$. According to Minkowski-Weyl's Theorem, given the convex set $X = \{x \in \mathbb{R}^n | Ax \le b\}$ where $Ax \le b$ is a **special block**, X can be represented by the extreme points and extreme rays of X: $X = \{x = \sum_i \lambda_i x^i + \sum_j \mu_i y^j | \sum_i \lambda_i = 1, \lambda_i \ge 0, \mu_j \ge 0\}$. When X is a bounded polyhedron, X can be represented by the extreme points, $X = \{x = \sum_i \lambda_i x^i | \sum_i \lambda_i = 1\,, \lambda_i \ge 0\}$.

Substituting the expression above to the original model leads to the following Master Problem:

$$\min \sum_i c^T \lambda_i \boldsymbol{x}^i \ (11)$$
$$\text{Subject to, } \sum_i D\lambda_i \boldsymbol{x}^i \le \boldsymbol{d}, \sum_i \lambda_i = 1 \text{ and } \lambda_i \ge 0 \qquad (6\text{-}11)$$

Suppose that a subset of extreme points P is available. The Restricted Master Problem (**RMP**) can be obtained: min $\sum_{i \in P} c^T \lambda_i \boldsymbol{x}^i$, subject to, $\sum_{i \in P} D\lambda_i \boldsymbol{x}^i \le \boldsymbol{d}, \sum_{i \in P} \lambda_i = 1$ and $\lambda_{i \in P} \ge 0$. Assume that λ^* and (π, ω) is the optimal and dual solutions to the RMP, respectively. The reduced cost is defined as $\gamma(\boldsymbol{x}) = c^T\boldsymbol{x} - \pi^T A\boldsymbol{x} - \omega$. Solve the subproblem: min $c^T\boldsymbol{x} - \pi^T A\boldsymbol{x} - \omega$, subject to $A\boldsymbol{x} \le b$ and $\boldsymbol{x} \ge 0$. If the reduced cost is non-negative, the solution is optimal; otherwise, the solution can be viewed as a new extreme point and added to the RMP until the reduced cost is non-negative.

With different objective functions related to different estimated states, the proposed models in section 4.1 based the generated information space in Section 3.2, will be solved under the framework of Dantzig–Wolfe decomposition in section 4.2. Specifically, based on the flow-balance constraint, the flow on a particular path (or path flow for a passenger group a) can represent one extreme point. A path flow uniquely corresponds to its path, so a particular path implicitly indicates a specific extreme point. This enables us to express the arc flow of group a on arc (i,j,t,s) as $x^a_{i,j,t,s} = \sum_h (\delta^{p(h),a}_{(i,j,t,s)} \times x^{p(h)}_a \times \lambda_{(a,h)})$, where $x^{p(h)}_a = D_a$ for each generated extreme point h, and $\sum_{h \epsilon H(a)} \lambda_{(a,h)} = 1$. Since variable $x^a_{i,j,t,s}$ is continuous rather discrete, it should be a continuous combination of extreme points and $\lambda_{(h,a)} \ge 0$ according to Minkowski-Weyl's Theorem. On the other hand, the link flow vector of each group $\boldsymbol{x}^a_{i,j,t,s}$ can also be seen as one extreme point, as it is the result of one specific path flow vector.

The system state uncertainty reflected by observability mainly arises from two sources: one is no useful information, which results in the many-to-one mapping between the many possible system states and one partial observation, and the other is the possible measurement error due to the noise and disturbance in sensing systems. This section will focus on quantifying the uncertainty of state estimates based on available limited useful observations. How to address the measurement error issues will also be discussed later.

As illustrated previously, the generated information space can work as a channel to connect available observations with different states. This section will propose different projection functions as the mapping between the feasible information space and specific transit system states. Table 6.3 introduces the focused states, which are also displayed in Figure 6.7. Those important states in traffic systems will not be focused here, but the modeling approach proposed in this paper can be well extended those systems.

Table 6.3: Focused states and motivations

Focused states	Motivations
(1) **Arc flow/density state**: passenger density on station platforms, in vehicles, and transfer corridors	(i) identify possible dangerous spots for safety; (ii) make decisions on vehicle updates, line/timetable changes and stop location adjustment
(2) **Path flow state**: the number of passengers taking one specific line segment	(i) clear the total ticket fare to each company based on the service they provide; (ii) evaluate the current liquidation policy and quantify the unreasonable income bound for each company
(3) **Path flow state**: the path flow range of each time-dependent OD pair	(i) compare or verify the traditional logit route choice model for better understanding travel behavior
(4) **Network-level arc flow/density state**: network-level time-dependent passenger flow/density states on serval key stations/vehicles	(i) distribute the network-level transit condition and intelligent passenger trip guidance (ii) evaluate network-level control and policy

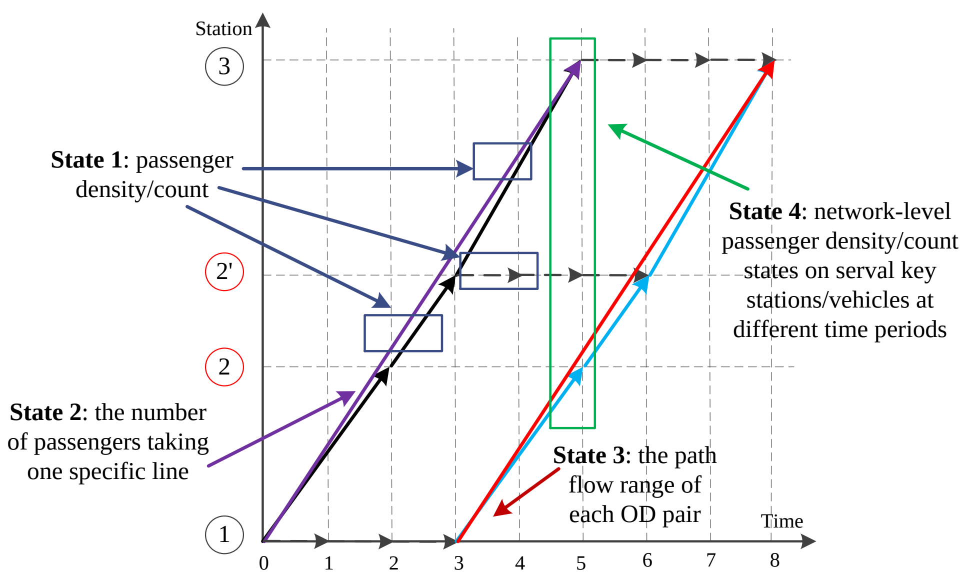


Figure 6.7: States illustration in a space-time network

The focused states and its uncertainty quantification are listed as follows.

- **Projection function 1 for arc flow state**: the number of passengers on one specific arc (i, j, t, s) (station platform, vehicle, transfer corridor) in the space-time network is represented as $\sum_a x^a_{i,j,t,s}$, so the arc flow uncertainty can be quantified by maximizing and minimizing $\sum_a x^a_{i,j,t,s}$, subject to constraints (1) to (5).

- **Projection function 2 for path flow state 1**: the earnings that one transit company r can obtain is represented as $\sum_a \sum_p (x^p_a \times c^r_p)$, where c^r_p is the income of using the segment in company r's operation area of path p. It can be calculated as a parameter in advance based on the ticket price and segment and path distance. Therefore, the earning bound is estimated by maximizing and minimizing$\sum_a \sum_p (x^p_a \times c^r_p)$ subject to constraints (6) to (10).
- **Projection function 3 for path flow state 2**: the flow rate on path p is $\sum_a x^p_a$, so the uncertainty bound of path flow is measured by maximizing and minimizing $\sum_a x^p_a$, subject to constraints (6) to (10).
- **Projection function 4 for network-level arc flow state**: the passenger flow (density) states on key station platforms at one time index (e.g., at 7:30am) is a high-dimensional vector $\{\boldsymbol{q}(\boldsymbol{i},\boldsymbol{t})\}$ where i is one of key stations. For one specific station i, $q(i,t) = \sum_a x^a_{i,j,t,s}$ is the number of passengers at station i at time t. Since the state is not one dimension anymore, the concept of the Maximal Possible Relative Error (MPRE) first introduced by Yang et al. (1991) is adopted to quantify the state uncertainty of high-dimensional variables. As shown in Figure 3.8(a), the state solution (vector $\{\boldsymbol{x}^{\boldsymbol{a}}_{\boldsymbol{i,j,t,s}}\}$) based on different projection functions for one-dimensional state above is one feasible solution in the information space, so each solution (vector $\{\boldsymbol{x}^{\boldsymbol{a}}_{\boldsymbol{i,j,t,s}}\}$) can be mapped to high-dimensional states to generate new state points (vector $\boldsymbol{q}(\boldsymbol{i},\boldsymbol{t})$)illustrated in Figure 6.8(b), which are used as sample points to approximately obtain the MPRE. Specifically, the average relative error between any two points needs to be calculated, and find the maximal one as the MPRE.

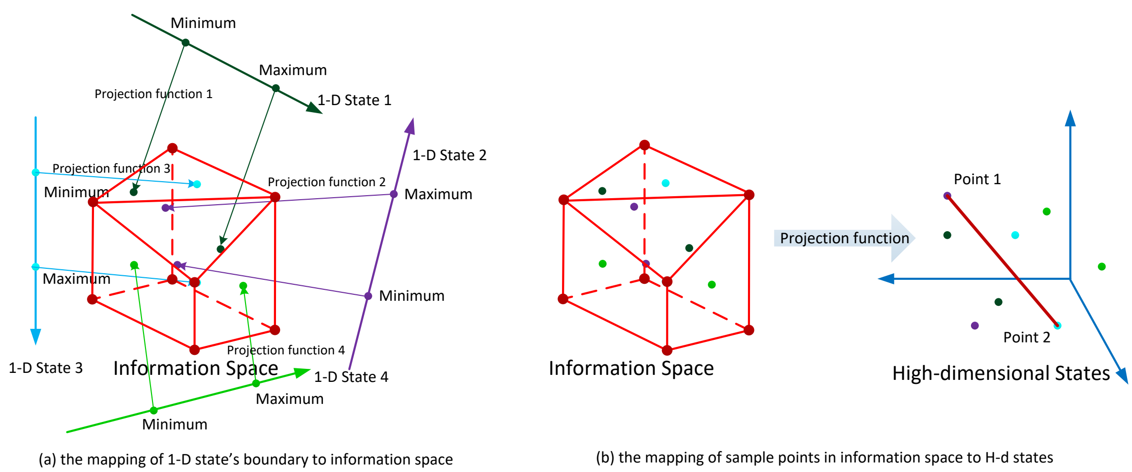


Figure 6.8: Relation of information space and different types of states

For example, the average relative error between point 1 and point 2 is calculated as follows (Yang et al., 1991), where $\boldsymbol{q_1}(\boldsymbol{i},\boldsymbol{t})$ and $\boldsymbol{q_2}(\boldsymbol{i},\boldsymbol{t})$ are a m-dimensional vector recording m stations' passenger flow at time t. The relative deviation is $\lambda_{(1,2,i,t)} = \frac{q_1(i,t) - q_2(i,t)}{q_1(i,t)}$ and the average relative deviation $AV\left(\lambda_{(1,2,t)}\right) = \sqrt{\frac{\phi(\lambda_{(1,2,t)})}{m}}$ (6-12)

where:

e$\phi(\lambda_{(1,2,t)}) = \sum_{i=1}^{m} \lambda_{(1,2,i,t)}{}^2$and $\lambda_{(1,2,t)} = \{\lambda_{(1,2,1,t)}, \lambda_{(1,2,2,t)}, \dots, \lambda_{(1,2,m,t)}\}$.

In addition, Yang et al. (1991) defined the concept of Estimation Reliability as a measure about the state uncertainty; that is, $Re = \frac{1}{1+AV(\lambda)}$, which shows that when the $AV(\lambda)$ is 0, the reliability of the estimated state is 1. In contrast, when $AV(\lambda)$ tends to infinity, there is almost no reliability guarantee. This result is just based on some sample points, so it is still an approximation approach.

From the perspective of Dantzig-Wolfe decomposition, based on the master problem in formulas (11) and (12), $\boldsymbol{x}^i$ as one extreme point can be replaced by variable vector $\boldsymbol{x}_{i,j,t,s}^{a}$ and variable $\boldsymbol{x}_{a}^{p}$ for arc-based and path-based models above, respectively, and c^T is the corresponding cost on each variable.

Taking minimizing the flow on arc (i, j, t, s) as an example, the general procedure of the algorithm is described as follows:

- **Step 1**: Initialization. Find one feasible passenger arc flow vector $\{\boldsymbol{x}_{i,j,t,s}^{a,k}\}$ on the shortest path as the k^{th} extreme point for each passenger group a. It indicates that $(k-1)$ extreme points have been generated before finding one feasible solution, which will be explained after Step 3 as a remark.
- **Step 2**: Solve the restricted master problem to obtain the duals of side constraints.
    - Min $\sum_k \sum_a (x_{i,j,t,s}^{a,k} \times \lambda_{a,k})$ (6-13)
- Subject to,
    - $\sum_k \sum_a (x_{i,j,t,s}^{a,k} \times \lambda_{a,1}) \le Cap_{i,j,t,s},\ \forall (i,j,t,s) \in A$ (6-14)
    - $\sum_k \sum_{(i,j,t,s)} [(x_{i,j,t,s}^{a,k} \times \lambda_{a,1}) \times c_{i,j,t,s}] = D_a \times \mu_a, \forall\, a$ (6-15)
    - $\sum_k \sum_a \sum_{t \in \tau} (x_{i,j,t,s}^{a,k} \times \lambda_{a,1}) = \mu_{i,j,\tau}, \forall (i,j,\tau)$ (6-16)
    - $\sum_k \lambda_{a,k} = 1, \forall\, a$ (6-17)
    - $\lambda_{a,k} \ge 0$ (6-18)
    -

  $\pi_{i,j,t,s}$, π_a, $\pi_{i,j,\tau}$ and ω_a are the duals of side constraints (14)-(17), respectively.
- **Step 3**: Solve each sub-problem as a time-dependent shortest path problem to calculate its reduced cost for each passenger group, which can be implemented by parallel computing techniques, such as, Multi Process Interface(MPI). The sub-problem for each passenger group a is:

$$\text{Min}\ \left(c_{i,j,t,s}^{a}\ 〖\times x〗\_(i,j,t,s)\text{^}(a,k+1)\right) - \sum_{(i,j,t,s)} (\pi_{i,j,t,s} \times x_{i,j,t,s}^{a,k+1}) - \pi_a \times \sum_{(i,j,t,s)} \left(c_{i,j,t,s} \times x_{i,j,t,s}^{a,k+1}\right) - \sum_{(i,j,\tau)} (\pi_{i,j,\tau} \times \sum_{t \in \tau} x_{i,j,t,s}^{a,k+1}) - \omega_a \quad (6\text{-}19)$$

  Subject to:

$$\sum_{i,t:(i,j,t,s) \in E} x_{i,j,t,s}^{a,k+1} - \sum_{i,t:(j,i,s,t) \in E} x_{j,i,s,t}^{a,k+1} = \begin{cases} -D_a & \forall a, j = o(a), s = DT^a \\ D_a & \forall a, j = d(a), s = T \\ 0 & otherwise \end{cases} \quad (6\text{-}20)$$

Actually, the reduced cost is:

$\left(c_{i,j,t,s}^{a} \times x_{i,j,t,s}^{a,k+1}\right) - \sum_{(i,j,t,s)}(\pi_{i,j,t,s} \times x_{i,j,t,s}^{a,k+1}) - \pi_a \times \sum_{(i,j,t,s)}\left(c_{i,j,t,s} \times x_{i,j,t,s}^{a,k+1}\right) - \sum_{(i,j,\tau)}(\pi_{i,j,\tau} \times \sum_{t\in\tau} x_{i,j,t,s}^{a,k+1}) - \omega_a$.

If it is negative, add the solution of the sub-problem to the restricted master problem at step 2 and begin next iteration. When the reduced costs of all sub-problems are non-negative, the optimal solution is achieved.

In the initialization step, to find one feasible solution from the shortest path problem as initial extreme points,it is possible to introduce artificial variables for those coupling constraints and solve the problem by the Dantzig-Wolfe decomposition again (Kalvelagen, 2003). For example, corresponding the example in section 3.3, the coupling side constraint is $\sum_j D_{i,j}x_i \leq d_i$, soit is possible to add artificial variable $y_i \geq 0$ to have $\sum_j D_{i,j}x_i - y_i \leq d_i$, and minimize $\sum_i y_i$ as a master problem. Based on the Dantzig-Wolfe decomposition algorithm, when the $\sum_i y_i$ is equal to 0,it is possible to conclude that one feasible solution for the primal problem is obtained and can be used for step 2.

In addition, for the maximum problem, this can be transformed into a minimum problem by changing the positive arc costs to be negative. Since there is no circle in the space-time network, the label correcting algorithm can always be used to find the shortest path.

The analyses on small card data (Trépanier et al., 2007; Barry et al., 2009) show that the data must be thoroughly validated and corrected prior to the practical use. Therefore, it might happen that no feasible solution exists when the observed data are directly used in built models. Even though each observation is tested in the model and can provide feasible solutions, it is still possible to have infeasible solutions when different measurement/observations are considered simultaneously, because the inconsistency among different kinds of sensors may still exist. Hence, the data need to be processed to obtain estimated measurements which are as close as possible to the corresponding measurements under real-world physical constraints. There are different approaches to clean and verify those measurements. The approach adopted here is to minimize the generalized least squares between the observed and corrected measurements, subject to constraints (1), (2) and (5). The model is explained at Appendix A in detail. Therefore, the measurement values μ_a and $\mu_{i,j,\tau}$ in information space generation and uncertainty quantification need to be replaced by μ_a^* and $\mu_{i,j,\tau}^*$ from the proposed prepressing model.

The model is a linearly constrained quadratic programming model. Frank-Wolfe algorithm is usually used to solve the optimization problem where the objective function is convex differentiable real-valued function and the feasible region of side constraints is compact convex (Frank and Wolfe, 1956).

The uncertainty of real-time system state increases the difficulty of real-time state prediction and optimal control. Compared with the offline state estimation in this paper, the challenges in the real-time condition include that (i) the real-time rail transit OD travel information is not available and (ii) the state transition along the time is highly required.

(i) Real-time OD demand estimation: Based on day-to-day historical and accurate dynamic OD demands in urban rail transit systems,it is possible to classify k representatives $OD_{o,d,\tau}^{k}$for each OD pair at different time periods, so the estimated real-time OD demand is $OD_{o,d,\tau} = \sum_k (w_k \times OD_{o,d,\tau}^{k})$ where w_k is a binary variable, which indicates that only one OD candidate k will be chosen. As a result, the dynamic OD travel demand's spatial structure can be well captured, compared with those OD estimation models which mainly optimize one departure time profile for all or one-class total static OD trips. In addition, the real-time trip generation at each station/origin with departure time is available from the smart card data, so $\sum_d OD_{o,d,\tau} = OD_{o,\tau}^{obs}$ provides more information to generate the real-time information space.

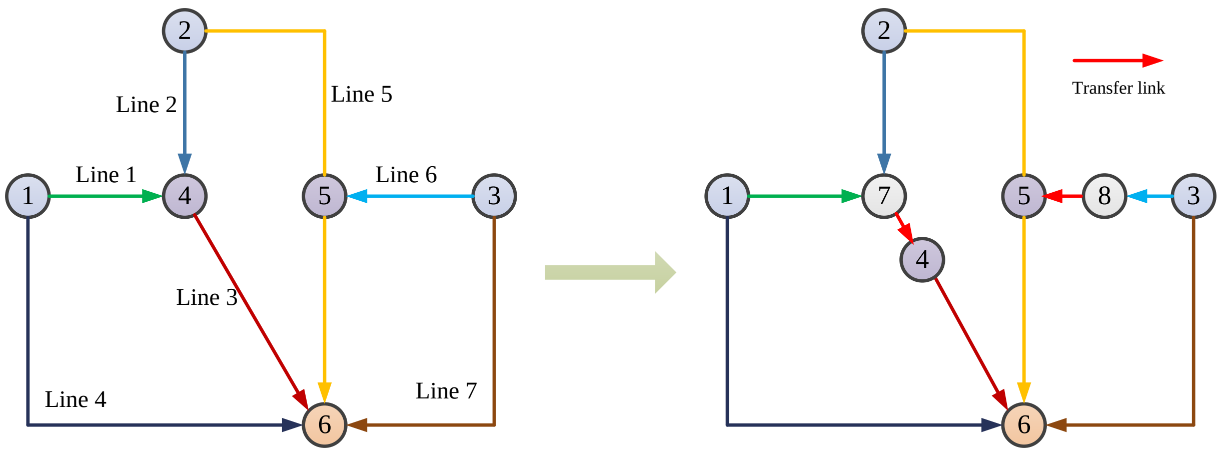
(ii) Real-time state transition: the rolling horizon approach has been widely chosen for real-time transportation operations and control (Peeta and Mahmassani, 1995; Zhou and Mahmassani, 2007; Meng and Zhou, 2011). Under this mechanism, when focusing on one time period, it needs a look-back period and a look-ahead period, because the generated passengers from the look-back period could still in the transit network during the focused time period, and in the look-ahead periodit is possible to assume that all passengers can arrive at their destination for the network modeling. Along the planning time horizon, once some trips are finished at the focused time period, their true OD information can be obtained in real time, so the corresponding estimated OD trips can be replaced by the real ones, which can also reduce the information space for the state uncertainty estimation.

(iii)Like the offline modelling, the states can be flexibly defined based on the managers' analysis goals. The min/max models on one-dimensional state and the MPRE for multi-dimensional states are also available for quantifying the real-time state uncertainty, which provides a fundamental input for the measure of future real-time prediction and optimal control.

6.5 EXPERIMENTS

This section demonstrates the proposed models and algorithms in Sections 4 and implement them in a general purpose optimization package GAMS. All source codes can be downloaded at the website: https://www.researchgate.net/publication/326020738_Observability_Scenarios_1-4.

The experiments are performed in the following transit network shown in Figure 6.9(a), where 7 urban rail lines exist in the transit systems. To model the passenger count observation at transfer corridors, specific transfer links are built as shown in Figure 6.9(b).



(a) Physical transit network

(b) Physical transit network with transfer links

Figure 6.9: Hypothetic urban rail transit network

1) (1) **Error! Reference source not found.** lists the existing transit service arcs based on given the timetable of the seven transit lines, and the corresponding space-time network is constructed in Figure 6.10.
2) (2) The origin, destination, departure time and aggregated trip time of each passenger group are listed in **Error! Reference source not found.**, and each group represents 100 passenger in this test.
3) (3) The vehicle capacity of each line is assumed in

Group No	OD Pair	Departure Time	Average Trip Time	Group No	OD Pair	Departure Time	Average Trip Time
1	1 → 6	0	6	15	1 → 6	3	7.5
2	1 → 6	0	7	16	1 → 6	3	7
3	1 → 6	0	8	17	1 → 6	3	8
4	1 → 6	0	6.5	18	2 → 6	3	6
5	2 → 6	0	7	19	2 → 6	3	7
6	2 → 6	0	7.5	20	2 → 6	3	6.5
7	2 → 6	0	6.5	21	2 → 6	3	7.5
8	2 → 6	0	6	22	2 → 6	3	8
9	3 → 6	0	7	23	2 → 6	3	6.8
10	3 → 6	0	7.5	24	3 → 6	3	7
11	3 → 6	0	8	25	3 → 6	3	7.5
12	1 → 6	3	6	26	3 → 6	3	7.4
13	1 → 6	3	7	27	3 → 6	3	7.8
14	1 → 6	3	6.5	28	3 → 6	3	8

4) , where it can be observed that the capacity of rail transit vehicles could have its adjustment at different time periods by increasing or decreasing the number of train units.
5) (4) The passenger count data from video processed data at transfer corridor $(7,4)$ is available; that is, 450 and 810 passengers are observed at time points 3 and 6.

Table 6.4: Hypothetic transit service arcs lists

Service Arc	Start Time	End Time	Service Arc	Start Time	End Time
(1,7)	0	3	(1,7)	3	6
(7,4)	3	4	(7,4)	6	7
(4,6)	4	6	(4,6)	7	9
(1,6)	0	8	(1,6)	3	11

(2,7)	0	3	(2,7)	3	6
(2,5)	0	4	(2,5)	3	7
(5,6)	4	7	(5,6)	7	10
(3,8)	0	3	(3,8)	3	6
(8,5)	3	4	(8,5)	6	7
(3,6)	0	8	(3,6)	3	11

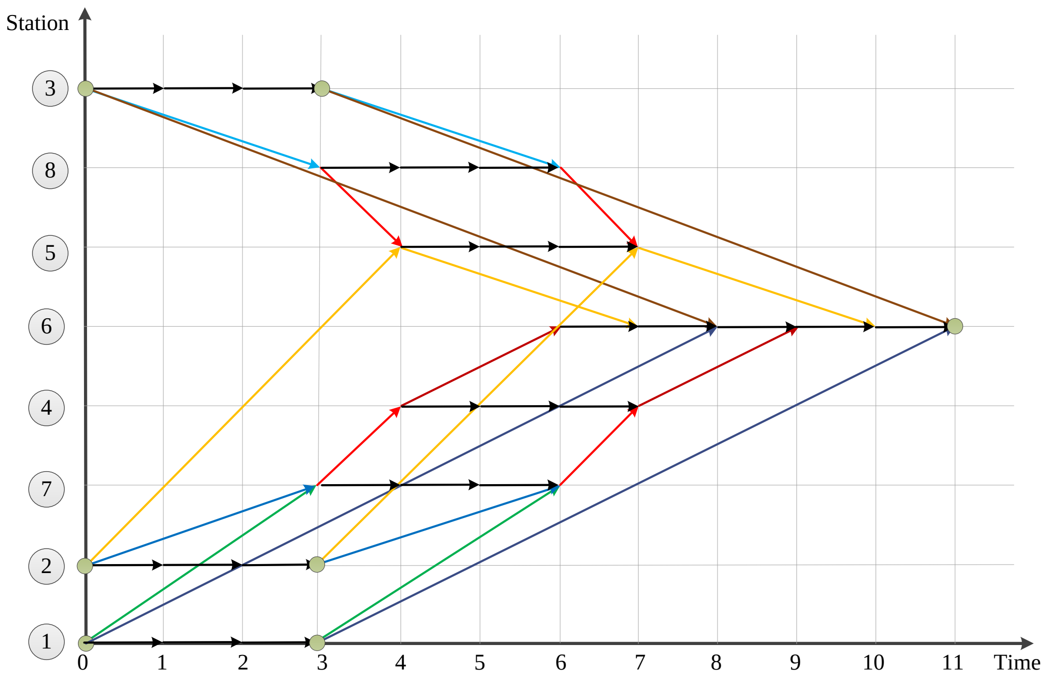


Figure 6.10: The corresponding space-time transit service network

Table 6.5: Trip attributes of each passenger group

Group No	OD Pair	Departure Time	Average Trip Time	Group No	OD Pair	Departure Time	Average Trip Time
1	1 → 6	0	6	15	1 → 6	3	7.5
2	1 → 6	0	7	16	1 → 6	3	7
3	1 → 6	0	8	17	1 → 6	3	8
4	1 → 6	0	6.5	18	2 → 6	3	6
5	2 → 6	0	7	19	2 → 6	3	7
6	2 → 6	0	7.5	20	2 → 6	3	6.5
7	2 → 6	0	6.5	21	2 → 6	3	7.5
8	2 → 6	0	6	22	2 → 6	3	8
9	3 → 6	0	7	23	2 → 6	3	6.8
10	3 → 6	0	7.5	24	3 → 6	3	7
11	3 → 6	0	8	25	3 → 6	3	7.5
12	1 → 6	3	6	26	3 → 6	3	7.4
13	1 → 6	3	7	27	3 → 6	3	7.8
14	1 → 6	3	6.5	28	3 → 6	3	8

Table 6.6: Vehicle capacity of transit lines

Line No	L1	L2	L3	L4	L5	L6	L7
Capacity of vehicles departing at time 0	300	300	600	200	400	300	200
Capacity of vehicles departing at time 3	400	400	800	300	600	400	300

The states for this experiment are as follows.

1) Arc flow state: passenger count (congestion) in transfer corridor $(8, 5)$ at time points 3 and 6, respectively.
2) Path flow state 1: the passenger flow departing at node 2 and time 0 to use line 1.
3) Path flow state 2: the earning collected in the ticket for company line 1 on its first vehicle.
4) Network-level arc flow state: the system-wide passenger count (congestion) on the running vehicles at time point 5.

6.5.1 Scenario Design

As a short summary, based on the available supply and demand data, the tasks are to (i) preprocess the measurements in case there is no feasible solution due to the possible existence of measurement errors in step 1, and (ii) quantify the uncertainty of the focused states in step 2. Five scenarios are designed to demonstrate the value of information based on the proposed models.

- **Scenario 1 (S1: base case)**: it is assumed that the origin, destination, and departure time of each passenger group is given, and no other information is available.
- **Scenario 2 (S2: base case + count)**: based on scenario 1, the passenger count data from video processed data at transfer corridor (7, 4) is available.
- **Scenario 3 (S3: base case + end-to-end travel time)**: based on scenario 1, the averaged group trip time from smart card is available.
- **Scenario 4 (S4: base case + end-to-end travel time + count)**: based on scenario 1, both the passenger count data and average group trip time data are available.
- **Scenario 5 (S5: ground truth)**: since the observed data may have its measurement errors, it is assumed that a ground truth can be obtained and will be compared with other scenarios. The ground truth is assumed as the system conditions based on maximizing the arc flow at time point 3 in scenario 3.

In step 1, the measurement is preprocessed by the proposed model at Appendix A. In step 2, the uncertainty range of states (1)-(3) is computed by maximizing and minimizing the state goals, and state (4) is addressed based on the solutions from the previous three states as a sample-based approximation. Before analyzing different state results in different scenarios, it is important to clearly illustrate the conditions under which those results are obtained from the proposed models.

1) In scenario 1, there is no available sensor data, so the measurement doesn't need to be preprocessed.
2) In scenario 2, the measurement is preprocessed for the passenger count data at transfer corridor (7, 4). The total squared errors in objective function (A.1) in step 1 is not equal to 0, which indicates that there will be no feasible solution if the observed measurement is directly used to step 2. The estimated passenger counts at transfer corridor (7, 4) at time points 3 and

6 is 450 and 800, respectively, compared with the observed values of 450 and 810. The total absolute error for the observed passenger count is 10.

3) In scenario 3, when step 1 is conducted using the observed average trip time, the total error is also not equal to 0 as well. The estimated average group trip time for each group is shown in **Error! Reference source not found.**. The total absolute error for the average group trip time is 2.58.
4) In scenario 4, in step 1, there are two different sensor data, so it will require weights on different measurements. As discussed by Lu et al. (2013), the weights should reflect the degrees of confidence on different observed data and can be represented by the inverses of the variances of the distinct sources of measurements. Therefore, the weights on aggregated average trip time and passenger count are calculated as 2.36 and 0.31, respectively. Finally, the total absolute errors for observed average group trip time and passenger count are 3.83 and 273, which are greater than the absolute errors in scenario 1 and scenario 2, respectively. It shows that the inconsistency among multi-source data makes the model to find a balance among those observation.
5) In scenario 5, the preprocessed group trip time in step 1 is used as the input to maximize the passenger count in transfer corridor $(8, 5)$ at time points 3, and the corresponding system condition is assumed as the ground truth in this dynamic transit system.

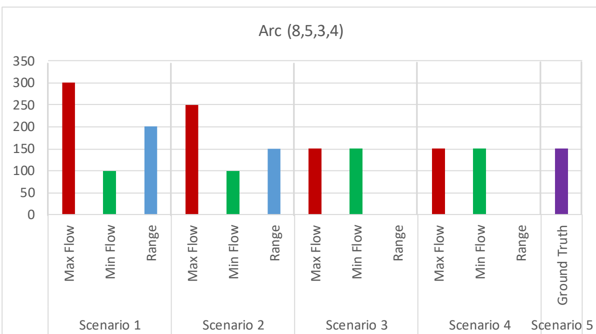
Table 6.7: The observed and preprocessed average group trip time for each passenger group

Passenger group No	Observed values	Preprocessed values in scenario 3	Passenger group No	Observed values	Preprocessed values in scenario 3
1	6	6	15	7.5	7.5
2	7	7	16	7	7
3	8	8	17	8	8
4	6.5	6.5	18	6	6
5	7	7	19	7	6.9
6	7.5	7	20	6.5	6.4
7	6.5	6.5	21	7.5	7
8	6	6	22	8	7
9	7	7	23	6.8	6.7
10	7.5	7.5	24	7	7.08
11	8	8	25	7.5	7.57
12	6	6	26	7.4	7.47
13	7	7	27	7.8	7.87
14	6.5	6.5	28	8	8

Figure 6.11 shows that the estimated maximum and minimal flow rates on each focused arc under different scenarios. As the increase of available information, the uncertainty range of passenger flows on transfer corridor (8, 5) is reduced. Meanwhile, both scenarios 3 and 4 can assert that their estimated state uncertainty is 0 and the state is completely observable. However, the different estimated unique states on arc (8,5,6,7) seem conflicted.

Specifically, in scenario 3, the observed trip time is preprocessed due to its measurement error, and finally the estimated states on transfer corridor (8, 5) is consistent with the states in the ground truth in scenario 5. Note that the estimated states may not be totally consistent with the ground truth, even though the observed data is same as the corresponding data in ground truth, because the observation is only a part reflection of the whole system condition. It is also possible that the corrected measurement is not consistent with that in this ground truth if other measurement correction approaches rather than the least square method are used in step 1.

In addition, in scenario 4, the inconsistency of observed link count data and observed trip time data makes the corrected measurement different with the corresponding data in the ground truth, so the final estimated unique state in step 2 is not the real-world condition anymore. Therefore, in reality, when the transportation system state is estimated by different sensor data, the data quality and assigned weight on each data source in step 1 is important and should be clearly stated. How to obtain a better weight on each observation source is beyond the scope of this paper. For more details on knowledge fusion, readers can refer to the paper (Zheng et al., 2014).



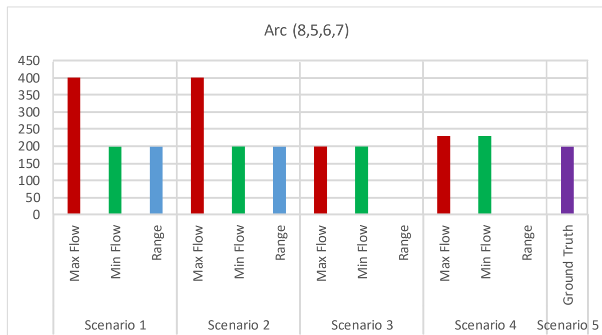


Figure 6.11: The estimated flow uncertainty range on each focused arc

Focusing on the passenger flow departing at node 2 and time 0 to use line 1, it is actually the path flow of path (2,0) →(5,4)→ (6,7). The path flow uncertainty is shown in Figure 6.12. The uncertainty range is similar to the arc flow above. The estimated unique state in scenario 4 is not consistent with the state value in ground truth. In addition, if line 1 is managed by one company and the other lines are managed by other different company, it needs to assign the fare to each company based on their service. However, the number of passengers using one specific line is uncertain in the transit system, so based on the proposed method,it is possible to quantify the uncertainty and estimate the general fare earning for each company rather than just using some simple rules for fare clearing (Gao et al., 2011; Zhou, 2014). For example, one simple rule is to calculate the shortest path and then assume that passengers will choose the shortest path as their selected lines.

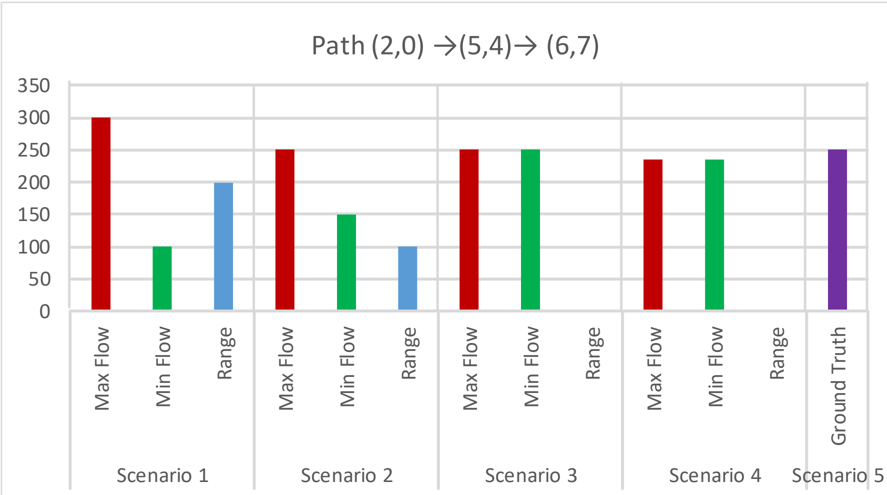


Figure 6.12: Estimated flow uncertainty range on the focused path

In each scenario, the passenger flow on arc (8,5,3,4), arc (8,5,6,7), and path of line 1 are maximized and minimized as 6 cases, respectively, so six feasible solutions of $x_{i,j,t,s}^{a}$ can be obtained as sample points to estimate the defined system-level state uncertainty. For state 4, the system-wide passenger count (congestion) on the running vehicles at time point 5 could be represented by the passenger flow on arcs (1,6,3,11), (1,6,0,8), (1,7,3,6), (2,5,3,7), (2,7,3,6), (3,6,3,11), (3,6,0,8), (3,8,3,6), (4,6,4,6), and (5,6,4,7). The results under six objectives in scenario 1 are listed in Table 6.8.

Table 6.8: Estimated passenger flows on arcs under six objectives in scenario 1

Arc(i,j,t,s)	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
(1,6,3,11)	300	300	300	300	200	200
(1,6,0,8)	200	200	200	200	100	100
(1,7,3,6)	300	300	300	300	400	400
(2,5,3,7)	400	400	200	400	200	200
(2,7,3,6)	200	200	400	200	400	400
(3,6,3,11)	300	300	100	300	100	100
(3,6,0,8)	0	200	200	200	0	200
(3,8,3,6)	200	200	400	200	400	400
(4,6,4,6)	500	300	300	300	600	400
(5,6,4,7)	400	400	400	400	400	400

Based on the definitions of Maximal Possible Relative Error (MPRE) and Estimation Reliability (Re), the values of MPRE and Re are 16.38 and 5.753%, respectively. As shown in **Error! Reference source not found.**, the possible flow on arc (3,6,0,8) is from 0 to 200, which creates a huge uncertainty and makes the estimation reliability extremely low. One possible reason is that the proposed models don't assume any travel behavior, so all solutions are based on the physical constraints and available sensor observations.

In scenario 2, with passenger count information, the estimated results of 6 cases for the system-level state are listed in **Error! Reference source not found.**. The corresponding values of MPRE and Re are 0.267 and 78.93%, respectively. It shows that the estimation reliability gets significantly improved when passenger counts from one key location (transfer corridor) are available, which could avoid a large uncertainty range occurred in scenario 1. Scenarios 3 and 4 are performed and their values of MPRE and Re are 0 and 100%, respectively, but it is still emphasized that the MRPR and Re should be clearly explained with its correspondingly different measurement preprocessing errors (assigned weights) and adopted approach.

Table 6.9: Estimated passenger flows on arcs under six objectives in scenario 2

Arc(i,j,t,s)	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
(1,6,3,11)	200	200	300	300	200	200
(1,6,0,8)	200	100	200	200	200	100
(1,7,3,6)	400	400	300	300	400	400
(2,5,3,7)	200	200	200	400	200	200
(2,7,3,6)	400	400	400	200	400	400
(3,6,3,11)	300	300	100	300	100	100
(3,6,0,8)	50	200	200	200	50	150
(3,8,3,6)	200	200	400	200	400	400
(4,6,4,6)	450	450	300	300	450	450
(5,6,4,7)	400	350	400	400	400	400

6.5.2 Results from Frank-Wolfe algorithm and Dantzig-Wolfe decomposition

In this section, the Frank-Wolfe algorithm is implemented in step 1 and the Dantzig-Wolfe decomposition algorithm in step 2 in GAMS. The case of minimizing the passenger flow on arc (8,5,3,4) in scenario 4 is treated as an example to analyze the performance of those algorithms. The source code can be downloaded at this link (https://www.researchgate.net/publication/324809217_F-W_and_D-W_Observability_Quantification).

Previously, the case was solved by the solver MINOS in GAMS directly. In step 1, the solved model is a non-linear programming model, and the minimal total generalized least square error in the objective function is 5.968. When the model is solved by Frank-Wolfe algorithm as a linear programming model, the result shown in Figure 6.13 finally converge to 7.069 after 20 iterations. The gap is probably caused by the optimal step size, which is found as a constant value at each iteration rather than a constant value vector for each variable. Hence, it could make the final solution converge to a local optimal solution.

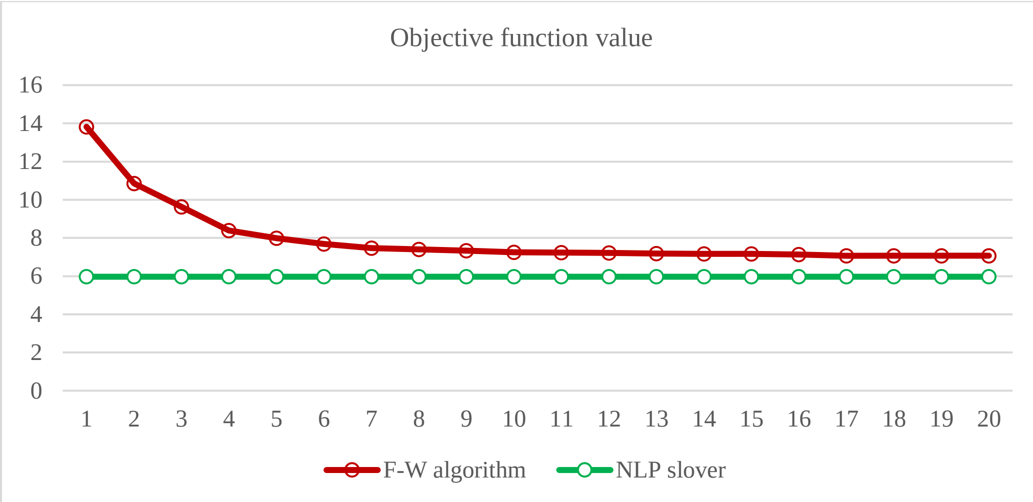


Figure 6.13: Objective function values under different solving approaches

In step 2, when the linear programming is directly solved by CPLEX, the objective function (minimal passenger flow on arc (8,5,3,4)) is 150. On the other hand, when Dantzig-Wolfe decomposition is applied to generate extreme points for time-dependent *OD* pairs, the minimal passenger flow is 142.5 based on the preprocessed measurements by Frank-Wolfe algorithm rather than by the NLP solver. The generated extreme points (feasible paths) and the correspondingly optimal weights are listed in Table 6.10. However, if the preprocessed measurements in step 1 are directly obtained from the NLP solver, the final minimal passenger count on arc (8,5,3,4) from Dantzig-Wolfe algorithm is 150 as well.

Table 6.10: Generated extreme points and optimal weights in Dantzig-Wolfe decomposition

Passenger Group No	Extreme points (path node sequence (ii, tt))	Optimal weights on extreme points	Passenger Group No	Extreme points (path node sequence (ii, tt))	Optimal weights on extreme points
1	(1,0)→(6,8);	0.06	15	(1,3)→(7,6) →(4,7) → (6,9);	0.30
	(1,0) → (7,3) → (4,4) → (6,6);	0.94		(1,3)→(6,11);	0.70
2	(1,0)→(6,8);	0.54	16	(1,3)→(7,6) →(4,7) → (6,9);	0.51
	(1,0) → (7,3) → (4,4) → (6,6);	0.46		(1,3)→(6,11);	0.49
3	(1,0)→(6,8);	0.93	17	(1,3)→(7,6) →(4,7) → (6,9);	0.07
	(1,0) → (7,3) → (4,4) → (6,6);	0.07		(1,3)→(6,11);	0.93
4	(1,0)→(6,8);	0.29	18	(2,3)→(7,6) →(4,7) → (6,9);	1
	(1,0) → (7,3) → (4,4) → (6,6);	0.71		(2,3)→(5,7) →(6,10);	0
5	(2,0)→(5,4) →(6,7);	0.93	19	(2,3)→(7,6) →(4,7) → (6,9);	0.15
	(2,0) → (7,3) → (4,4) → (6,6);	0.07		(2,3)→(5,7) →(6,10);	0.85
6	(2,0)→(5,4) →(6,7);	0.93	20	(2,3)→(7,6) →(4,7) → (6,9);	0.62
	(2,0) → (7,3) → (4,4) → (6,6);	0.07		(2,3)→(5,7) →(6,10);	0.38
7	(2,0)→(5,4) →(6,7);	0.58	21	(2,3)→(7,6) →(4,7) → (6,9);	0
	(2,0) → (7,3) → (4,4) → (6,6);	0.42		(2,3)→(5,7) →(6,10);	1
8	(2,0)→(5,4) →(6,7);	0.13	22	(2,3)→(7,6) →(4,7) → (6,9);	0

	(2,0) → (7,3) → (4,4) → (6,6);	0.87		(2,3)→(5,7) →(6,10);	1
9	(3,0)→(6,8);	0.08	23	(2,3)→(7,6) →(4,7) → (6,9);	0.36
	(3,0) → (8,3) → (5,4) → (6,7);	0.92		(2,3)→(5,7) →(6,10);	0.64
10	(3,0)→(6,8);	0.56	24	(3,3)→(6,11);	0.09
	(3,0) → (8,3) → (5,4) → (6,7);	0.44		(3,3)→(8,6) →(5,7) →(6,10);	0.91
11	(3,0)→(6,8);	0.93	25	(3,3)→(6,11);	0.55
	(3,0) → (8,3) → (5,4) → (6,7);	0.07		(3,3)→(8,6) →(5,7) →(6,10);	0.45
12	(1,3) → (7,6) → (4,7) → (6,9);	0.94	26	(3,3)→(6,11);	0.45
	(1,3)→(6,11);	0.06		(3,3)→(8,6) →(5,7) →(6,10);	0.55
13	(1,3) → (7,6) → (4,7) → (6,9);	0.50	27	(3,3)→(6,11);	0.85
	(1,3)→(6,11);	0.50		(3,3)→(8,6) →(5,7) →(6,10);	0.15
14	(1,3) → (7,6) → (4,7) → (6,9);	0.76	28	(3,3)→(6,11);	0.93
	(1,3)→(6,11);	0.24		(3,3)→(8,6) →(5,7) →(6,10);	0.07

6.5.3 Tests in a large-scale network

To address the computational challenges in large-scale networks, an approximation-based approach is employed, which provides a *k*-shortest path set as extreme points for each passenger group (in each *OD* pair with time-dependent departure time) in advance rather than using Dantzig-Wolfe decomposition to generate extreme point iteration by iteration. In this section, the public Google Transit Feed Specification (GTFS) data from Alexandria Transit Company in 2015 is used as the tested large-scale transit network (https://transitfeeds.com/p/alexandria-transit-company). As shown in Figure 6.14, it has 12 routes, 1638 trips (866 trips on weekdays, 423 trips on Saturdays, 261 trips on Sundays, and 88 trips on Christmas day), and 629 stops.

Figure 6.14: Alexandria transit network read from GTFS, in Virginia, USA

In this experiment, the trips on weekdays are considered the provided schedule. Then, 32,029 vertexes and 713,650 arcs are generated in the corresponding space-time network for one whole weekday. The arcs include vehicle running arcs, passengers' walking arcs from origin to transit stops and from transit stops to destination, transfer arcs, and waiting arcs. The space-time arc generation rules contain that (i) the trip (path) travel time is less than 120min; (ii) the maximum number of transfer times is 3; (iii) the maximum transfer/walking time is 30min; (iv) the maximum transfer/walking distance is 0.5mile.

To obtain the time-dependent transit demand, the traffic analysis zones in the city of Alexandria are mapped to the transit network as the activity locations. As a result, 42 OD pairs are matched. Plus, the time period of 7:00am to 9:00am is divided by 36 time intervals, so the time-dependent OD demand is defined by each 5 mins. Finally, 1484 time-dependent OD pairs are obtained based on the arc generation rules above.

In addition, as an approximation for those extreme points in Dantzig-Wolfe decomposition for each time-dependent *OD* pair, 3-shortest paths are generated using the developed *k*-shortest path algorithm. Finally, 4452 paths were generated with 7,868 arcs in the space-time network. The *k*-shortest path algorithm for each time-dependent *OD* pair was shown as follows.

(i) Based on the origin vertex (origin node and departure time) in the space-time network, the label correcting algorithm is used to generate a shortest path tree from origin vertex to all possible vertexes selected based on the space-time arc generation rules.
(ii) According to the destination physical location,it is possible to find many candidate vertexes (stop id and stop time in schedule) connecting the destination node by walking arcs. Thenit is possible to add the label costs of those candidate vertexes and its corresponding walking arc costs to the destination, so the destination will have many vertexes (destination node and arrival time) with different label cost.
(iii)Sort those label costs of the destination node and select *k* least-cost destination vertexes and back trace to the origin vertex. As a remark, at each vertex, the transfer state is recorded (the number of transfer times), so when back tracing the path to origin vertex,it is possible to obtain different paths from one same vertex with same label cost but with different transfer states. Finally, the k-shortest path set can be generated for each time-dependent *OD* pair.

For simplicity, it is assumed that all transit vehicle capacity is 35 and the walking, waiting and transfer arc capacity is 9999. Also, the time-dependent demand of each *OD* pair is assumed to be 1, which means that one passenger will arrive every 5 mins for each *OD* pair. The observed passenger trip time is assumed and generated as a random value between the minimal and the maximal path costs of 3-shorest paths. the focused state is the uncertainties of passenger flow count on transfer links from stop 4000644 to stop 4000863 and from stop 4000745 to stop 4000509 based on the 3-hour transit demand. Then, the four models are solved by CPLEX in GAMS as a linear programming problem by a workstation with Intel(R) Xeon(R) CPU E2680 v2 @ 2.8GHz processors. For each model, there are 10,837 equations and 4,452 variables, and the computation time is around 19 seconds. The results are shown in Figure 6.15.

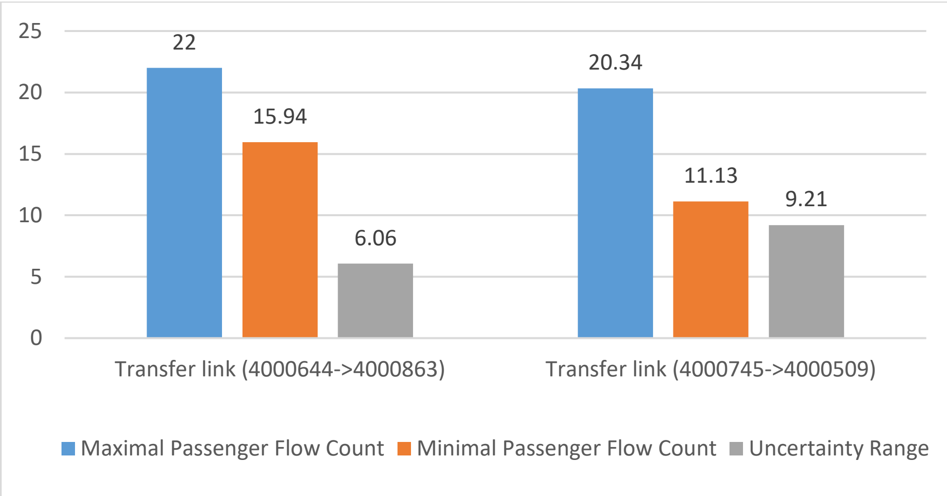


Figure 6.15: Uncertainties of passenger flow count at transfer links

6.6 SUMMARY

USDOT (2015) listed optimizing traffic flow on congested freeways and arterial streets as one of fundamental urban mobility challenges for Smart City and pointed out that outdated traffic signal timing causes more than 10 percent of all traffic delay on major routes in urban areas. In the subway system of Beijing, 96 stations implement the passenger flow control policy to relieve the system congestion during peak hours in 2018. As explained before, the basic question is how well the system can be observed, and thenit is possible to provide the best control to reach the goals, whatever it is in supply side to optimally control signal timing, lane use, speed limit, vehicle rescheduling or in the demand side to influence travelers' departure time choice, route choice and trip generation. Note that most previous studies mainly focused on most likely system state estimation rather than system observability quantification. Hence, this study has aimed to develop a modeling framework capable of incorporating multi-source sensor data to address various system state uncertainty quantification in the whole transportation system. The contributions of the effort are:

(i) The information space is generated by a system of linear equations and inequality constraints based on the multi-source sensor data and physical transportation system representation in a unified framework.

(ii) Different projection functions are proposed to map the unique information space with the focused different system states (e.g., passenger density on the station platform, in the vehicle, or in the transfer corridor in urban rail transit system) for further system observability quantifications;

(iii) The proposed space-time network flow models are finally solved as a simplified linear programming model by using Dantzig-Wolfe decomposition and Frank-Wolfe algorithm, which improves the computational efficiency in theory.

(iv) The observation errors are also considered by a least square model to correct the directly observed measurements, such as, trip time of grouped passengers from transit smart card, aggregated passenger count from video systems, path choice information from cell phone trajectory data.

The developments presented in this section provide insights on the relationship among multi-source information, information space, state estimation, and system observability quantification by taking the urban transit systems as the analysis object. The information space and information errors are highly respected for state estimation, and projection- functions-based approaches are presented to quantify the uncertainty of different states under same information space. The proposed models can explain that the value of information highly relies on its aimed specific estimated states and sensor location rather than its volume. It provides the analysis base for how to better use available information for different state estimates and how to design the sensor network for future estimate improvement.

It should be remarked that, the observability quantification based on different states is just the first step for better observing and controlling the system. The following questions are currently under the considerations for future research: (1) what is the balance among the system observability, the minimally needed information, and the required accuracy of future controls? (2) What is the balance of the sensor data cost, value of information and its computational efficiency in proposed models and algorithms? (3) How to integrate the heterogeneous sensor network design with the real-time system control? (4) How to visualize the real-time uncertainty of different system states in a straightforward way for the public.

7.0 SUMMARY AND FUTURE WORK

7.1 SUMMARY

This study has explored new and creative ways to use sensor data to 1) enhance freight-related path choice, both pre-trip and en-route, and 2) improve the performance of urban networks more generally from a freight perspective. While a significant body of literature exists on both path choice and traffic assignment, this study presents new and creative ways to address these topics predicated on real-time data and a freight-first mentality. These new methods can lead to better freight-focused routing decisions and network operating conditions whose performance for freight is improved and can be assessed statistically.

Sensor data, from vehicles and facilities, is revolutionizing how urban transportation systems operate. Pre-trip route choices can be informed by network status, en-route path choices can be predicated on evolving conditions, prices can influence path choice decisions, and more robust network operating conditions can be obtained. Careful placement of sensing equipment can enhance system observability in and controllability.

Two specific research objectives were targeted. The first was creation of new data-driven, truck-oriented path choice algorithms. The second was a data-informed, freight-focused traffic assignment model. Both these efforts have produced results that can enhance freight flows and at the same time mitigate congestion. The efforts built on previous research efforts in which the authors were involved plus findings from projects in which they have collaborated.

The path choice problem has been addressed using algorithms that deal with multiple objectives. This is important because trucks are rarely just concerned with one aspect of the trip-making task. One algorithm finds the k-shortest paths based on cost and risk, as illustrations of two objectives that often surface as being important. Another identifies routes on a probabilistic basis, with each route having a likelihood of being selected for use. The third explicitly finds the non-dominated multi-objective set of paths for any origin-destination (*OD*) pair. This third algorithm makes it possible for decision makers to look explicitly at the options available and select the path that seems to achieve the best compromise among the objectives.

The traffic assignment model uses pricing strategies to encourage the choice of "desired" paths. The network prices specifically facilitate truck flows. The prices focus on improving the quality of the truck trips; encouraging trucks to use facilities (freeways, arterials, etc.) that have high-quality performance. The pricing strategies are thus, multi-class vectors, with different prices by vehicle class. Moreover, with an interest in improving network performance robustness, the pricing strategies endeavor to reduce volume-to-capacity *v/c* ratios on the arcs. This improves network resilience; that is, the network's ability to deal with unforeseen (and unpreventable) conditions caused by incidents, bad weather, unanticipated maintenance work, special events, etc. This resilience is critical to freight. It reduces the likelihood that door-to-door travel times will change dramatically if the network conditions that unfold are not exactly consistent with those used in making path choices.

The literature review shows that while much has been done on both routing and traffic assignment, the approaches suggested here are new and unique. But, much of that work focuses on path choice in the abstract, simply assuming there is an origin, a destination, a network, and paths to be identified.

The notion that the path choice problem should emphasize trucks, especially, or path choice factors that are of special concern to trucks is uncommon. One exception is the HazMat literature, that explicitly focuses on path choice for trucks. Admittedly, the commodity is special, and it has unique concerns, but the findings are generalizable to trucks more generally. This is especially true in two senses. First, there are only some links (arcs) in the network that are available for use. Many urban areas designate a truck network that can (must) be used except for local pick-ups and deliveries. Second, multiple objectives are considered. In the case of HazMat, the post common ones are cost and risk. But, there can be others, such as exposure to accidents or challenging geometric conditions.

The section presents three ways to develop truck-focused paths. The first follows the paradigm created by Dial *et al.* (xxx) that emphasizes the distribution of flows among paths based on their relative "impedances" or costs. That is, paths with better (lower) costs would see a higher percentage of the flows and those with worse (higher) costs would see less. A drawback is that the procedure does not examine the objectives individually, it combines them through a generalized cost function and uses that function to compute the "cost" of each route. Also, the method is "weak" in that it does not identify paths explicitly. Rather, it ascertains the percentage of each OD flow that will use specific arcs. These values are commonly called arc utilizations and are employed heavily in traffic assignment procedures. In a truck context, the main nuance is that the generalized cost functions are different for the flow classes. Autos are focused on cost (based on travel time and distance) and to some degree tolls. But, the trucks are far more focused on tolls; and, to some degree risks. So, in identifying paths by class, it is important that these different generalized cost functions be employed.

The second methodology identifies the K-shortest paths for each OD pair. As might be obvious, these paths range from the "shortest" (lowest cost) to the "longest" (highest cost). In the sense that a (linear) generalized cost function is employed to identify the paths, all K of the paths reflect the weights employed by that function. The non-dominated paths that reflect tradeoffs among those objectives are not identified. In fact, as Figure 3.2 shows, the relative combination of the objectives can vary widely from one of the K shortest paths to another. This means the decision maker needs to be comfortable with seeing the ratios between the objectives vary considerably as the paths progress from $k = 1$ to $k = K$.

The second methodology explicitly identifies the non-dominated set of paths for each OD pair. If there are two objectives, then the result is akin to Figure 1.3, as is shown by Figure 3.3, where a set of four non-dominated paths were found for a specific OD pair. An advantage to this methodology is that it identifies the set of paths that has the "optimal" tradeoffs among the objectives. That is, if switching from path A to path B allows one objective's value to be made better at the expense of making another objective's value "worse", the least damage to the second objective is done by selecting path B. Also, unlike the first path choice algorithm, the paths are identified explicitly. The two significant drawbacks are 1) the algorithm, unlike Dial's (1971), provides no guidance about the probabilities that the paths should be chosen. Or, alternately put, the percentage of flow that should use any path. Also, if flow is moved from one path to another, then a shift must be made to a path that proportionally involves a different weighted combination of the objectives, because it lies on the non-dominated surface, and those paths, inherently, span the space between and among the solutions that optimize the objectives one at a time.

It is not that one of these path choice options is "best" or "correct". They have strengths and weaknesses. Any one of them provides path choices for the traffic assignment problem that are useful and valuable. Also, they can all be sensitive to multiple objectives, either through a generalized cost function, or through an explicit identification of the non-dominated multi-objective path options. Which one is best to use depends on how the traffic assignment problem is approached. In this study, the third one has been carried forward because it provides a set of paths that lie on the non-dominated surface.

In traffic assignment, much work has also been done. It may be the most-explored topic in all transportation research literature. But, the number of papers that explicitly focus on putting freight (trucks) first is very limited. And, the number that focus on multi-commodity assignment is also limited. Moreover, the number that incorporate capacity constraints into the scope of the problem is also limited. Those are among the important aspects of the problem considered here. So, this work is a contribution to the state-of-the-art in that regard.

Also limited at the number of traffic assignment studies that have focused on network resilience; specifically, managing the *v/c* ratios among the arcs in the network. This idea is very uncommon. But, it is critical, the study team suggests, in the context of freight(truck) routing. It is not so much that the v/c ratios need to be kept low for the truck assignments to be either feasible or optimal. But, rather, that keeping the *v/c* ratios low helps ensure that the network can deal with unforeseen situations that arise, from accidents, incidents, weather, or other situations. Hence, a min max objective is incorporated that endeavors to keep the *v/c* ratios for the arcs below target values.

Also unusual is the use of a gap function to measure the achievement of the user optimal solution. Beckmann's formulation is used far more often, with an emphasis on satisfying the KTT conditions that their objective function creates. Instead, in this study a gap concept is used where the travel times on the path are compared with a target path time that is desirable. The target might be the travel time (or cost) associated with the user equilibrium solution. Or, it could be some other travel time that is a policy objective from the perspective of the network operator. An advantage to adopting this perspective is that the user optimal objective function value obtained in any solution can be compared against the value of 0 that would be obtained were the user optimal solution achieved. The Beckmann objective function cannot be used for this purpose.

This study effort simply suggests that multiple objectives are typically important. And, it uses cost, risk, and tolls to motivate that thought. In fact, the tolls are a control mechanism that network managers can use to encourage specific tendencies in the selection of truck routes. The idea is somewhat counter-intuitive since the "objective" is to keep the trucks off local streets and encourage them to use high0type facilities. This is important, because the more common purpose in introducing tolls is to rationalize the use of capacity-challenged facilities, like bridges and tunnels, that have limited capacity but high demand. Those tolls tend to discourage traffic from using facilities that are "good" from the perspective of truck paths. Hence, this motivates the idea that two toll structures might be useful. One would focus on auto trips, and discourage highly peaked demands, and highly-peaked flows, and the use of capacity-challenged facilities where other (lower class facility) options are available. In contrast, the pricing structure for the trucks should encourage the use of high-type facilities (e.g., freeways) and discourage the use of low-type facilities (like local streets). It is important to see that this difference in perspective is important and to utilize this insight to design the network pricing structure.

All the path choice methods can be made sensitive to real-time information about the evolving network conditions. They can all make use of emerging trends in the network travel times (rates). And, they can be sensitive to the inclusion / exclusion of specific arcs because of use restrictions (no trucks), either permanent, temporary, or condition (time-of-day) dependent; and they can be sensitive to changes in the "impedances" for the arcs in the sense of varying tolls. The one unexpected insight is that the tolls pertaining to trucks might be different from and motivated by a different objective that those that pertain to autos. In fact, the two may have opposite trends. While the auto-focused tolls may be intended to discourage peaking and the use of capacity-challenged arcs, the truck tolls may want to discourage use of local streets and low-type facilities; and encourage trucks to use the high-quality facilities, which may, in fact, be the ones that are capacity-challenged. This means the trucks are (or should be) encouraged to use the "best" facilities so that they do not "rat run" through the network to avoid tolls on the best facilities available. Rather, the tolls should encourage the trucks to stay on the best arcs in the network so that they do not seek paths that use local streets. This is a pricing strategy that has significant, "freight-first" implications.

This section has described a realization of the traffic assignment problem in which trucks are represented separate from autos. The problem is non-linear in that the travel times on the arcs are sensitive to the arc flows. And that, in turn, leads to one of the objectives being quadratic. The terms involve the multiplication of one choice variable times another. Fortunately, for small problems, there are generalized solvers that can deal with such situations. For large-scale problems, like the ones that are faced for most metropolitan areas, such a non-linear formulation is not practical to solve explicitly. But, for illustration purposes, a simplified network for the Albany, NY metropolitan area is used in the case study to show the type of results obtained.

The most important nuance in the model is the fact that the path choices for the trucks are different from those for the non-trucks (autos). The generalized cost function is different and the network over which the trucks can travel is more restrictive. The trucks take into consideration the risk associated with traversing the arcs (as in exposure to accidents) and the tolls are given considerable weight. Also, the toll structure is different for trucks than it is for autos. The tolls for autos discourage use of capacity-challenged facilities like bridges, tolls, and heavily used freeway links, encouraging the vehicles to use other, comparable, but "lower quality" facilities, like the arterials, in their paths. For the trucks, however, the toll structure discourages the use of local streets and other "lower class" facilities so that neighborhoods are not exposed to unnecessary truck traffic. And, they are encouraged, through low tolls, to use the "high type" facilities, such as freeways, except for local pickups and deliveries. This bi-pronged pricing strategy helps to put "freight-first" in the context of the traffic assignment. And, it leads to solutions that the public is "more likely" to accept because it discourages trucks from using local streets and highways.

The mathematical equations that are involved in the traffic assignment model are presented and described as well as the LINGO problem statement which implements them. A case study example is presented, with excerpts of the results, so that the reader can gain a sense of the results obtained. (The complete, machine readable workspaces exist and are available for anyone wishing to use them.)

7.2 FUTURE WORK

Much future work can be carried out based on the analyses conducted so far. Some important examples of these efforts are as follows:

Real-World Tests. As is often the case, the methodological advances presented here have been tested using a blend of empirical data and hypothetical situations. One natural extension for future work is to test these methods based on datasets that are more representative and reflective of real-world conditions. This pertains to all the methods presented, from the assessment of reliability for segments and routes to the selection of plans for truck routing and locations for distribution centers.

Additional Sensitivity Analyses. As is often the situation, the case study analyses focus on some aspects of the problem and not others. In this instance, there are several aspects of the problem that were not explored. In the instance of Section 3, the impact of variations in the objective weights would be interesting to explore. Only a few combinations were tested. Also, the weight for the multinomial logit model in Dial's (1971) algorithm could be adjusted parametrically to see what affect it has. In the case of the *k*-shortest path algorithm, the value of *K* could be varied, and the weight values changed, individually and in combination. In Section 4, the weights among the objectives could be altered, the arcs for which trucks are prohibited could be changed, and the impacts of tolling strategies could be explored more thoroughly. The latter is particularly interesting because of the observation that the tolling strategy for the trucks should, probably, be different than it is for autos. In Section 5, it would be very interesting to see if general trends can be discerned for the combinations of network parameter values that produce the paradox; and, if there are related values for the tolls that alleviate its occurrence.

REFERENCES

Abkowitz M., and P. Cheng, 1988. "Developing a Risk-Cost Framework for Routing Truck Movements of Hazardous Materials," Accident Analysis and Prevention, vol. 20, pp. 39-51.

Alsger, A., Assemi, B., Mesbah, M., Ferreira, L., 2016. Validating and improving public transport origin–destination estimation algorithm using smart card fare data. Transportation Research Part C, 68, 490-506.

Ban, X.J., Hao, P. and Sun, Z., 2011. Real time queue length estimation for signalized intersections using travel times from mobile sensors. Transportation Research Part C: Emerging Technologies, 19(6), 1133-1156.

Ban, X., Herring, R., Margulici, J.D., Bayen, A., 2009. Optimal sensor placement for freeway travel time estimation. Transportation and Traffic Theory, Chapter 34 (W.H.K. Lam, S.C. Wong, H.K. Lo eds.), Springer, 697-721.

Barber, E., and L. Hildebrand, 1980. Guidelines for Applying Criteria to Designate Routes for Transporting Hazardous Materials, Report No. FHWA-IP-90-15.

Barry, J., Freimer, R. and Slavin, H., 2009. Use of entry-only automatic fare collection data to estimate linked transit trips in New York City. Transportation Research Record: Journal of the Transportation Research Board, (2112), 53-61.

Batta, R. and S. Chiu, 1985. Optimal Obnoxious Paths on a Network: Transportation of Hazardous Materials, Working Paper No. 85-106, Department of Industrial Engineering, SUNY at Buffalo, Buffalo, NY.

Beckmann, M., C.B. McGuire, and C.B. Winsten, 1956. Studies in the Economics of Transportation, New Haven: Yale University Press.

Bennett, L., 1993. The existence of equivalent mathematical programs for certain mixed equilibrium traffic assignment problems, European Journal of Operational Research, Vol 71, pp. 177-187.

Bianco, L., Confessore, G., Reverberi, P., 2001. A Network Based Model for Traffic Sensor Location with Implications on O/D Matrix Estimates. Transportation Science 35(2), 50-60

Bierlaire, M., 2002. The total demand scale: a new measure of quality for static and dynamic origin–destination trip tables. Transportation Research Part B: Methodological, 36(9), 837-850.

Blumsack, S. and M. Ilić, 2006. The Braess Paradox in Electric Power Systems, Carnegie Mellon Electricity Industry Center Working Paper.

Boyles, S., Waller S. T., 2011. Optimal information location for adaptive routing. Networks and Spatial Economics 11, 233-254.

Braess, D., 1969. Über ein Paradoxon aus der Verkehrsplanung. Unternehmensforschung 12, 258–268.

Braess, D., A. Nagurney, and T. Wakolbinger, 2005. On a Paradox of Traffic Planning, Transportation Science, Vol. 39, pp. 446–450.

Brogan, J., and J. Cashwell, 1985. "Routing Models for the Transportation of Hazardous Materials - State Level Enhancements and Modifications," Transportation Research Record 1020, pp. 19-22.

California, State of; Department of Motor Vehicles, referenced on July 31, 2018, California Driver Handbook – Laws and Rules of the Road, https://www.dmv.ca.gov/portal/dmv/detail/pubs/hdbk/speed_limits++.

Capital Area Metropolitan Planning Organization, 2018. 2045 Metropolitan Transportation Plan, http://www.campo-nc.us/transportation-plan/2045-metropolitan-transportation-plan.

Castillo, E., Conejo, A.J., Pruneda R., Solares C., 2007. Observability in linear systems of equations and inequalities: Applications. Computer & Operations Research, 34, 1708-1720.
Castillo, E., Conejo, A.J., Menéndez, J.M. and Jiménez, P., 2008. The observability problem in traffic network models. Computer ‐ Aided Civil and Infrastructure Engineering, 23(3), 208-222.
Castillo, E., Rivas, A., Jimenez, P., Menendez, J. M., 2012. Observability in Traffic Networks. Plate Scanning Added by Counting Information. Transportation 39 (6), 1301-1333.
Castillo, E., Grande, Z., Calviño, A., Szeto, W.Y. and Lo, H.K., 2015. A state-of-the-art review of the sensor location, flow observability, estimation, and prediction problems in traffic networks. Journal of Sensors, 2015.
Cats, O., West, J. and Eliasson, J., 2016. A dynamic stochastic model for evaluating congestion and crowding effects in transit systems. Transportation Research Part B: Methodological, 89, 43-57.
Ceapa, I., Smith, C. and Capra, L., 2012. Avoiding the crowds: understanding tube station congestion patterns from trip data. In Proceedings of the ACM SIGKDD international workshop on urban computing (134-141). ACM.
Cetin, M., List, G., and Zhou, Y. (2005) "Factors affecting the minimum number of probes required for reliable travel time estimation", Journal of the Transportation Research Board, Vol. 1917, pp 37-44.
Chase, R.T., B.M. Williams, N.M. Rouphail, and S.Kim, 2012. Comparing Reported Travel Times between Temporally-Stitched and Simultaneous Methods at Various Aggregation Intervals. 91st Annual Meeting of the Transportation Research Board, Washington, D.C.
Chen, H.K. and Hsueh, C.F., 1998. A model and an algorithm for the dynamic user-optimal route choice problem. Transportation Research Part B: Methodological, 32(3), 219-234.
Chen, X., Zhou, X., and List, G. (2011), "Using Time-Varying Tolls to Optimize Truck Arrivals at Ports", Transportation Research, Part E, 47:6 pp. 965-982.
Codina, E. and Rosell, F., 2017. A heuristic method for a congested capacitated transit assignment model with strategies. Transportation Research Part B: Methodological, 106, 293-320.
Cox, R., 1984. Routing and Scheduling of Hazardous Materials Shipments: Algorithmic Approaches to Managing Spent Nuclear Fuel Transport, Ph.D. Dissertation, Cornell University, Ithaca, NY.
Cox, R. and M. Turnquist, 1986. Scheduling Truck Shipments of Hazardous Materials in the Presence of Curfews, Transportation Research Record 1063, pp. 21-26.
Current, R., C.S. ReVelle and J.L. Cohon, 1990. "An Interactive Approach to Identify the Best Compromise Solution for Two Objective Shortest Path Problems," Computers and Operations Research, vol. 17, pp. 187-198.
Dafermos, S. and F. Sparrow, 1969. The traffic assignment problem for a general network, J. Res. Natl. Bur. Stand., Sect, Vol 73, pp. 91–118.
Daganzo, C.F., 2007. Urban gridlock: Macroscopic modeling and mitigation approaches. Transportation Research Part B: Methodological, 41(1), 49-62.
Danczyk, A., Liu, H., 2011. A mixed-integer linear program for optimizing sensor locations along freeway corridors. Transportation Research Part B, 45 (1) 208–217.
Dantzig, G.B. and Wolfe, P., 1960. Decomposition principle for linear programs. Operations research, 8(1), 101-111.
Desrosiers, J. and Lübbecke, M.E., 2005. A primer in column generation. In Column generation (1-32). Springer US.
Dial, R., 1971. A Probabilistic Multipath Traffic Assignment Model which Obviates Path Enumeration, Transportation Research, 5:2, pp. 83-111.

Dial, R., 1996. Bicriterion traffic assignment: basic theory and elementary algorithms, Transportation Science, 30:2, pp. 93-111.

Dijkstra, E. W. (1959). "A note on two problems in connexion with graphs" (PDF). Numerische Mathematik. 1: 269–271.

Drissi-Kaïtouni, O., Hameda-Benchekroun, A., 1992. A dynamic traffic assignment model and a solution algorithm. Transportation Science 26, 119–128.

Eisenman, S., M., Fei, X., Zhou, X., Mahmassani, H. S., 2006. Number and Location of Sensors for Real-time Network Estimation and Prediction: A Sensitivity Analysis. Transportation Research Record 1964, 253-259.

Eisenman, S. and List, G. (2004) "Using Probe Data to Estimate OD Matrices", Proceedings, 7th International IEEE Conference on Intelligent Transportation Systems, Washington, DC.

Federal Highway Administration, 2013. National Performance Management Research Data Set (NPMRDS)Technical Frequently Asked Questions, https://ops.fhwa.dot.gov/freight/freight_analysis/perform_meas/vpds/npmrdsfaqs.htm.

Frank, M. and Wolfe, P., 1956. An algorithm for quadratic programming. Naval Research Logistics (NRL), 3(1‐2), 95-110.

Ford Jr, L.R. and Fulkerson, D.R., 1958. A suggested computation for maximal multi-commodity network flows. Management Science, 5(1), 97-101.

Gao, L., Wang, B., and Zhang, C., 2011. Comparison of urban rail transit fare clearing model based on travel survey. Modern urban transit, 11, 97-99.

Gentili, M., Mirchandani, P. B., 2005. Locating Active Sensors on Traffic Networks. Annals of Operations Research 136 (1): 229-257.

Gentili, M. and Mirchandani, P.B., 2012. Locating sensors on traffic networks: Models, challenges and research opportunities. Transportation research part C: emerging technologies, 24, 227-255.

Glickman, T., and D. Rosenfield, 1984. "Risks of Catastrophic Derailments involving the Release of Hazardous Materials," Management Science, Vol. 30, No. 4, No. 503-511.

Gorard, S., 2005. Revisiting a 90-year-old debate: the advantages of the mean deviation. British Journal of Educational Studies, 53(4), 417-430.

Handler, G.Y., and I. Zang, 1980. "A Dual Algorithm for the Constrained Shortest Path Problem," Networks, Vol. 10, pp. 293-310.

Harker, P., 1988. Multiple Equilibrium Behaviors on Networks, Transportation Science, 22:1, pp. 39-46.

Haurie, A. and P. Marcotte, 1985. On the relationship between Nash—Cournot and Wardrop equilibria, Networks, Volume 15, pp 295–308.

Hearn, D.W., 1980. Bounding flows in traffic assignment models, Research Report 80-4, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL.

Henig, M.I. (1985), "The Shortest Path Problem with Two Objective Functions," European Journal of Operations Research, Vol. 25, pp. 281-291.

Herrera, J.C., Work, D.B., Herring, R., Ban, X.J., Jacobson, Q. and Bayen, A.M., 2010. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. Transportation Research Part C: Emerging Technologies, 18(4), 568-583.

Hu, S. R., Peeta, S., Chu, C., 2009. Identification of Vehicle Sensor Locations for Link-Based Network Traffic Applications. Transportation Research Part B 43 (8), 873-894.

H. Huang, and S. Gao, 2012. "Optimal paths in dynamic networks with dependent random link travel times", Transp. Res. Part B: Method., vol. 46, pp. 579-598.

Inouye, H., 1987. Traffic equilibria and its solution in congested road networks, R. Genser (Ed.), Proc. IFAC Conf. on Control in Transportation Systems, pp. 267–272 Vienna.
Ivancie, F., 1984. Hazardous Materials Routing Study: Final Report, prepared by the Portland Office of Emergency Management, Portland, OR.
Ji, Z., Kim, Y., and A. Chen, 2011. Multi-objective α-reliable path finding in stochastic networks with correlated link costs: A simulation-based multi-objective genetic algorithm approach (SMOGA). Expert Systems with Applications, 38(3), pp. 1515–1528.
Jiang, Y. and Szeto, W.Y., 2016. Reliability-based stochastic transit assignment: Formulations and capacity paradox. Transportation Research Part B: Methodological, 93, 181-206.

Jorgensen, 1963. Some aspects of the urban traffic assignment problem, Graduate Report, Institute of Transportation and Traffic Engineering, University of California, Berkeley, CA.
Kalman, R., 1959. On the general theory of control systems. *IRE Transactions on Automatic Control 4*(3), 110-110
Kalvelagen, E., 2003. Dantzig-Wolfe Decomposition with GAMS.
Kessler, D., 1986. "Establishing Hazardous Materials Truck Routes for Shipments Through the Dallas-Fort Worth Area," Recent Advances in Hazardous Materials Transportation Research: An International Exchange, TRB State of the Art Report 3.
Koenker, R. and Hallock, K.F., 2001. Quantile regression. Journal of economic perspectives, 15(4), 143-156.
Korilis, Y., A. Lazar, and A. Orda, 1999. Avoiding the Braess paradox in non-cooperative networks, Journal of Applied Probability,Volume 36, pp. 211-222.
Korilis, Y., A. Lazar, and A. Orda, 1997. Capacity Allocation Under Noncooperative Routing, IEEE Transactions on Automatic Control, VOL. 42, pp. 309-325.
Korilis, Y., A. Lazar, and A. Orda, 1995. Architecting non-cooperative network, IEEE journal on selected areas on communication, Vol 13, pp. 1241-1251.
Kusakabe, T. and Asakura, Y., 2014. Behavioural data mining of transit smart card data: A data fusion approach. Transportation Research Part C: Emerging Technologies, 46, 179-191.
Kusakabe, T., Iryo, T. and Asakura, Y., 2010. Estimation method for railway passengers' train choice behavior with smart card transaction data. Transportation, 37(5), 731-749.
Lam, W.H. and Yin, Y., 2001. An activity-based time-dependent traffic assignment model. Transportation Research Part B: Methodological, 35(6), 549-574.
Larsson, T. and Patriksson, M., 1992. Simplicial decomposition with disaggregated representation for the traffic assignment problem. Transportation Science, 26(1), 4-17.
Larsson, T., Patriksson, M. and Rydergren, C., 2004. A column generation procedure for the side constrained traffic equilibrium problem. Transportation Research Part B: Methodological, 38(1), 17-38.
LaValle, S.M., 2012. Sensing and filtering: A fresh perspective based on preimages and information spaces. Foundations and Trends® in Robotics, 1(4), 253-372.
Lei, H., Zhou, X., List, G., and Taylor, J., 2015. "Characterizing Corridor-level Travel Time Distributions Based on Stochastic Flows and Segment Capacities", Cogent Engineering, 2: 990672.
Leurent, F., 1994. "Cost Versus Time Equilibrium over a Network.", 1994 Annual Conference of the Transportation Research Board. Washington D.C.
Li, P., Mirchandani, P., Zhou, X., 2015. Solving simultaneous route guidance and traffic signal optimization problem using space-phase-time hypernetwork. Transportation Research Part B 81(1), 103–130.

Li, X., Ouyang, Y. 2011. Reliable sensor deployment for network traffic surveillance. Transportation Research Part B 45 (1), 218–231.
List, G., P.B. Mirchandani, M.A. Turnquist and K.G. Zografos, 1991. "Modeling and Analysis for Hazardous Materials Transportation: Risk Analysis, Routing/Scheduling and Facility Location," Transportation Science, vol. 25, no. 2, pp. 100-114.
List, G., 1993. "Siting Emergency Response Teams: Tradeoffs Among Response Time, Risk, Risk Equity and Cost," Transportation of Hazardous Materials: Issues in Law, Social Science and Engineering, L.N. Moses and D. Lindstrom, Eds., Kluwer Academic Publishers, pp. 117-134.
List, G., E. Williams, B. Yang, and L., Yu, 2015. State of the Art in Freight Reliability Analysis and Modeling, prepared for the National Transportation Center at Maryland.
List, G., 2016. Task 6.3 Truck Flow Forecasting Model: Description and Application, prepared for the Capital Area Metropolitan Planning Organization.
List, G., E. Williams, J. Addison, and A. Morsali, 2017. Improving the Efficiency and Reliability of Freight Transportation, prepared for the National Transportation Center at Maryland, Project INTC2014-SU-R-06.
List, G., and M. Turnquist, M., 1991. Logistics for Hazardous Materials Transportation: Scheduling, Routing, and Siting. Prepared for the Revion II University Transportation Research Center. ,
Liu, J., Kang, J.E., Zhou, X. and Pendyala, R., 2017. Network-oriented household activity pattern problem for system optimization. Transportation Research Part C: Emerging Technologies.
Liu, J., Zhou, X., 2016. Capacitated transit service network design with boundedly rational agents. Transportation Research Part B: Methodological, 93, 225-250.
Lu, C.C., 2016. Robust multi-period fleet allocation models for bike-sharing systems. Networks and Spatial Economics, 16(1), 61-82.
Lu, C.C., Liu, J., Qu, Y., Peeta, S., Rouphail, N.M., Zhou, X., 2016. Eco-system optimal time-dependent flow assignment in a congested network. Transportation Research Part B 94, 217–239.
Lu, C.C., Zhou, X., Zhang, K., 2013. Dynamic origin–destination demand flow estimation under congested traffic conditions. Transportation Research Part C, 34, 16-37.
Ma, X.L., Liu, C., Liu, J.F., Chen, F. Yu, H., 2015. Boarding stop inference based on transit IC card data. Journal of Transportation Systems Engineering and Information Technology, 15(4), 78-84.
Ma, J., Smith, B.L., and Zhou, X., 2015. “Personalized Real-time Traffic Information Provision: Agent-based Optimization Model and Solution Framework”, Transportation Research, Part C.
Ma, X.L., Wang, Y.H., Chen, F., Liu, J.F., 2012. Transit smart card data mining for passenger origin information extraction. Journal of Zhejiang University Science C, 13(10), 750-760.
Ma, X.L., Wu, Y.J., Wang, Y., Chen, F. and Liu, J., 2013. Mining smart card data for transit riders’ travel patterns. Transportation Research Part C: Emerging Technologies, 36, 1-12.
Mahmoudi, M. and Zhou, X., 2015. Finding Optimal Solutions for Vehicle Routing Problem with Pickup and Delivery Services with Time Windows: A Dynamic Programming Approach Based on State-space-time Network Representations, Draft Manuscript, Arizona State University.
Masuda, Y. and S. Whang, 1999. Dynamic Pricing for Network Service: Equilibrium and Stability, Management Science, Vol. 45, pp. 857-869.
McCord M., and A.Y-C. Leu, 1995. “Sensitivity of Optimal Hazmat Routes to Limited Preference Specifications,” INFOR, 33:2, pp. 68-83.

Meng, L. and Zhou, X., 2011. Robust single-track train dispatching model under a dynamic and stochastic environment: a scenario-based rolling horizon solution approach. Transportation Research Part B: Methodological, 45(7), 1080-1102.

Milchtaich. I., 2006. Network topology and the efficiency of equilibrium, Games and Economic Behavior,Vol 57, pp. 321–346.

Mirchandani, P. and H. Soroush, 1985. "Optimal Paths in Probabilistic Networks: A Case with Temporary Preferences," Computers and Operations Research, Vol. 12, No. 4, pp. 365-381.

Mirchandani, P., and M. Wiecek, 1993. Routing with nonlinear multi-attribute cost functions, Applied Mathematics and Computation, 54:2-3, 215-239.

Munizaga, M.A. and Palma, C., 2012. Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. Transportation Research Part C: Emerging Technologies, 24, 9-18.

Nair, R., Miller-Hooks, E., Hampshire, R.C. and Bušić, A., 2013. Large-scale vehicle sharing systems: analysis of Vélib'. International Journal of Sustainable Transportation, 7(1), 85-106.

Nassir, N., Hickman, M. and Ma, Z.L., 2015. Activity detection and transfer identification for public transit fare card data. Transportation, 42(4), 683-705.

New York City, referenced on July 31, 2018, Trucks and Commercial Vehicles, http://www.nyc.gov/html/dot/html/motorist/trucks.shtml.

Ng, M., 2012. Synergistic sensor location for link flow inference without path enumeration: A node-based approach. Transportation Research Part B: Methodological, 46(6), 781-788.

Nozick, L. G. List, and M. Turnquist, 1997. Integrated Routing and Scheduling in Hazardous Materials Transportation, Transportation Science, 31:3, pp. 200-215.

Nunes, A.A., Dias, T.G. and e Cunha, J.F., 2016. Passenger journey destination estimation from automated fare collection system data using spatial validation. IEEE transactions on intelligent transportation systems, 17(1), 133-142.

Pan, B., Zheng, Y., Wilkie, D. and Shahabi, C., 2013. Crowd sensing of traffic anomalies based on human mobility and social media. In Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (344-353). ACM.

PC-Miler, 2018. Introducing PC-Miler 32, https://www.pcmiler.com/.

Pelletier, M.P., Trépanier, M. and Morency, C., 2011. Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies, 19(4), 557-568.

Peeta, S. and Mahmassani, H.S., 1995. Multiple user classes real-time traffic assignment for online operations: a rolling horizon solution framework. Transportation Research Part C: Emerging Technologies, 3(2), 83-98.

Port Authority of New York and New Jersey, July 31, 2018, Traffic Restrictions, Bridges and Tunnels, https://www.panynj.gov/bridges-tunnels/lincoln-tunnel-traffic-restrictions .html.

Poorzahedy H. and M. Turnquist, 1982. Approximate algorithms for the discrete network design problem, Transportation. Research, Vol 16, pp. 45-55.

Roughgarden, T., 2006. On the severity of Braess's Paradox: Designing networks for selfish users is hard, Journal of Computer and, System Sciences, Volume 72, pp. 922–953.

Seaborn, Catherine, John Attanucci, and Nigel H. M. Wilson.2009. Analyzing Multimodal Public Transport Journeys in London with Smart Card Fare Payment Data. Transportation Research Record: Journal of the Transportation Research Board 2121, 55–62.

Shang, P., Li, R., Liu, Z., Yang, L. and Wang, Y., 2018. Equity-oriented skip-stopping schedule optimization in an oversaturated urban rail transit network. Transportation Research Part C: Emerging Technologies, 89, 321-343.

Sheffi, Y., 1985. Urban Transportation Networks, Prentice-Hall.

Shier, D.R., 1976. "Iterative Methods for Determining the k Shortest Paths in a Network," Networks, Vol. 6, pp. 205-229.
Shobrys, D.E., 1981. A Model for the Selection of Shipping Routes and Storage Locations for a Hazardous Substance, Ph.D. dissertation, John Hopkins University.
Srinivasan, K. K., Prakash, A. A., & Seshadri, R. (2014). Finding most reliable paths on networks with correlated and shifted log–normal travel times. Transportation Research Part B: Methodological, 66, pp. 110–128.
Sun, Y., Xu, R., 2012. Rail transit travel time reliability and estimation of passenger route choice behavior: Analysis using automatic fare collection data. Transportation Research Record (2275), 58-67.
Szeto, W.Y. and Jiang, Y., 2014. Transit assignment: Approach-based formulation, extragradient method, and paradox. Transportation Research Part B: Methodological, 62, 51-76.
Tang, J., Song, Y., Miller, H.J. and Zhou, X., 2016. Estimating the most likely space–time paths, dwell times and path uncertainties from vehicle trajectory data: A time geographic method. Transportation Research Part C: Emerging Technologies, 66, 176-194.
Thiagarajan, A., Ravindranath, L., LaCurts, K., Madden, S., Balakrishnan, H., Toledo, S. and Eriksson, J., 2009. VTrack: accurate, energy-aware road traffic delay estimation using mobile phones. In Proceedings of the 7th ACM conference on embedded networked sensor systems(85-98). ACM.
Tong, L., Zhou, X. and Miller, H.J., 2015. Transportation network design for maximizing space–time accessibility. Transportation Research Part B: Methodological, 81, 555-576.
Transportation Research Board, 2016. Highway Capacity Manual, 6th Edition.
Trépanier, M., Tranchant, N. and Chapleau, R., 2007. Individual trip destination estimation in a transit smart card automated fare collection system. Journal of Intelligent Transportation Systems, 11(1), 1-14.
Turnquist, M., 1987. "Routes, Schedules and Risks in Transporting Hazardous Materials," Strategic Planning in Energy and Natural Resources, B. Lev, et al., Eds., North-Holland, Amsterdam, pp. 289-302.
USDOT, 2015. Smart City Challenge: Lessons for Building Cities of the Future. Report.
Van Vuren, T. and D. Watling, 1991. `A multiple user class assignment model for route guidance, Transportation research, record 1306, pp. 22-31, 1991. ,
Vythoulkas, P.C., 1990. A dynamic stochastic assignment model for the analysis of general networks. Transportation Research Vol. 24B, No 6, pp. 453-470.
Wardrop, J., 1952. Some theoretical aspects of road traffic research. In Proceedings of the Institution of Civil Engineers, Volume 1, pp. 325-378.
Wei, Y., Avcı, C., Liu, J., Belezamo, B., Aydın, N., Li, P.T. and Zhou, X., 2017. Dynamic programming-based multi-vehicle longitudinal trajectory optimization with simplified car following models. Transportation Research Part B: Methodological, 106, 102-129.
Wu, X., Guo, J., Xian, K., Zhou, X., 2018. Hierarchical travel demand estimation using multiple data sources: A forward and backward propagation algorithmic framework on a layered computational graph. Submitted
Wu, Tzu-li, G.F. List, and J.L. Adler, 1999. "Multiple-Objective Perspectives of Traffic Assignment Models for the Traffic Network Management Problem," Proceedings, Fourth Conference of the Hong Kong Society for Transportation Studies, Harbour Plaza Hotel, Hong Kong.
Xing, T., and X. Zhou, 2011. Finding the most reliable path with and without link travel time correlation: A Lagrangian substitution based approach. Transportation Research Part B: Methodological, 45(10), pp. 1660–1679.

Xing, T., Zhou, X., Taylor, J., 2013. Designing Heterogeneous Sensor Networks for Estimating and Predicting Path Travel Time Dynamics: An Information-Theoretic Modeling Approach. Transportation Research Part B, 57, 66-90.
Xu, X., Lo, H.K., Chen, A. and Castillo, E., 2016. Robust network sensor location for complete link flow observability under uncertainty. Transportation Research Part B: Methodological, 88, 1-20.
Yang, H., Iida, Y. and Sasaki, T., 1991. An analysis of the reliability of an origin-destination trip matrix estimated from traffic counts. Transportation Research Part B: Methodological, 25(5), 351-363.
Yang, H. and Meng, Q., 1998. Departure time, route choice and congestion toll in a queuing network with elastic demand. Transportation Research Part B: Methodological, 32(4), 247-260.
Yang, H., Zhou, J., 1998. Optimal Traffic Counting Locations for Origin-Destination Matrix Estimation. Transportation Research Part B 32, 109-126.
Yen, J. Y., 1971. "Finding the k-Shortest Loopless Paths in a Network", Management Science, 17: 712–716.
Yuan, N.J., Wang, Y., Zhang, F., Xie, X. and Sun, G., 2013. Reconstructing individual mobility from smart card transactions: A space alignment approach. In Data Mining (ICDM), 2013 IEEE 13th International Conference, 877-886.
Yuan, J., Zheng, Y., Xie, X. and Sun, G., 2011. Driving with knowledge from the physical world. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (316-324). ACM.
Zawack, D.J. and Thompson, G.L., 1987. A dynamic space-time network flow model for city traffic congestion. Transportation Science, 21(3),153-162.
Zhang, X., H. Yang, and H.-J. Huang, 2008. Multiclass multicriteria mixed equilibrium on networks anduniform link tolls for system optimum, European Journal of Operational Research, Volume 189, Pages 146–158.
Zeleny, M., 1982. Multiple Criteria Decision Making, McGraw Hill.
Zhang, F., Yuan, N.J., Wilkie, D., Zheng, Y. and Xie, X., 2015. Sensing the pulse of urban refueling behavior: A perspective from taxi mobility. ACM Transactions on Intelligent Systems and Technology (TIST), 6(3), 37.
Zhao, J., Rahbee, A., Wilson, N.H., 2007. Estimating a Rail Passenger Trip Origin‐Destination Matrix Using Automatic Data Collection Systems. Computer‐Aided Civil and Infrastructure Engineering, 22(5), 376-387.
Zhao, S. and Zhang, K., 2017. Observing individual dynamic choices of activity chains from location-based crowdsourced data. Transportation Research Part C: Emerging Technologies, 85, 1-22.
Zheng, Y., Capra, L., Wolfson, O. and Yang, H., 2014. Urban computing: concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 5(3), 38
Zhou, F., Xu, R.H., 2012. Model of passenger flow assignment for urban rail transit based on entry and exit time constraints. Transportation Research Record (2284), 57-61.
Zhou, L., 2014. Verification and optimization of network clearing model. Urban mass transit, 11, 59-66.
Zhou, X., and List, G.F., 2010. An information-theoretic sensor location model for traffic origin-destination demand estimation applications. Transportation Science, 44(2), 254-273.
Zhou, X. and Mahmassani, H.S., 2007. A structural state space model for real-time traffic origin–destination demand estimation and prediction in a day-to-day learning framework. Transportation Research Part B: Methodological, 41(8), 823-840.

Zhu, N., Fu, C., Ma, S., 2018. Data-driven distributionally robust optimization approach for reliable travel-time-information-gain-oriented traffic sensor location model. Transportation Research Part B: Methodological, 113, 91-120.

Zhu, Y., Koutsopoulos, H.N. and Wilson, N.H., 2017a. A probabilistic Passenger-to-Train Assignment Model based on automated data. Transportation Research Part B: Methodological, 104, 522-542.

Zhu, Y., Koutsopoulos, H.N. and Wilson, N.H., 2017b. Inferring left behind passengers in congested metro systems from automated data. Transportation Research Part C: Emerging Technologies.

Zimmerman, J., Tomasic, A., Garrod, C., Yoo, D., Hiruncharoenvate, C., Aziz, R., Thiruvengadam, N.R., Huang, Y. and Steinfeld, A., 2011. Field trial of tiramisu: crowd-sourcing bus arrival times to spur co-design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (1677-1686). ACM.

Zografos, K. and C. Davis, 1989. "Multiobjective Programming Approach for Routing Hazardous Materials," Journal of Transportation Engineering, vol. 115, pp. 661-673.