A travel demand management strategy: The downtown space reservation system

Y. Zhao, K. Triantis, D. Teodorović, P. Edara

System Performance Laboratory, Grado Department of Industrial and Systems Engineering, Northern Virginia Center, #440, Virginia Tech, 7054 Haycock Road, Falls Church, VA 22043-2311, United States
University of Belgrade, Faculty of Transport and Traffic Engineering, Voyvode Stepe 305, 11000 Belgrade, Serbia
Department of Civil and Environmental Engineering, University of Missouri-Columbia, E3502 Lafferre Hall, Columbia, MO 65211, United States

ABSTRACT

In this paper, a Travel Demand Management strategy known as the Downtown Space Reservation System (DSRS) is introduced. The purpose of this system is to facilitate the mitigation of traffic congestion in a cordon-based downtown area by requiring people who want to drive into this area to make reservations in advance. An integer programming formulation is provided to obtain the optimal mix of vehicles and trips that are characterized by a series of factors such as vehicle occupancy, departure time, and trip length with an objective of maximizing total system throughput and revenue. Based upon the optimal solution, an “intelligent” module is built using artificial neural networks that enables the transportation authority to make decisions in real-time on whether to accept an incoming request. An example is provided that demonstrates that the solution of the “intelligent” module resembles the optimal solution with an acceptable error rate. Finally, implementation issues of the DSRS are addressed.

1. Introduction and context

Various traffic management policies have been introduced to cope with the fast growth of congestion. In recent years, transportation engineers and planners have increasingly embraced strategies that deal with the management of existing facilities, rather than building new infrastructures on account of the investment and maintenance costs associated with these new infrastructures along with environmental considerations. These strategies are generally labeled as Travel Demand Management (TDM) strategies and involve managing travel demand by encouraging travel and land use patterns that lead to less congestion producing ways. In the report TTI prepared for the Federal Highway Administration (FHWA), the following description is included: “... (TDM strategies) include putting more people into fewer vehicles (through ridesharing, increased public transportation ridership, or dedicated highway lanes for high-occupancy vehicles), shifting the time of travel (e.g., through staggered work hours), and eliminating the need for travel altogether (e.g., through telecommuting)” (TTI, 2005, p. 87). As the popular press has indicated, many metropolitan cities have demonstrated a high interest in congestion management after reviewing the success of many congestion pricing implementations in Europe and Asia.

In response to the need of providing innovative and effective solutions for traffic congestion, the design of reservation systems is being considered as an alternative and/or complementary TDM strategy (Edara, 2005). Reservation indicates that a user will follow a booking procedure defined by the reservation system before traveling so as to obtain the right to access a facility or resource (e.g., in this research, travel slots to visit the downtown area). This concept is relatively new to traffic congestion management. However, it is widely used in other domains (e.g., reserving in advance an airline seat, etc.). From the provider’s point of view, a reservation system can function as a control mechanism that oversees the sales of a resource given that traffic congestion, as an emergent phenomenon of individuals’ actions, is hard to understand and predict. Moreover, from the consumer’s perspective the reservation system ensures less uncertainty in obtaining the desired resource.

With the advent and deployment of advanced traveler information systems (ATIS), drivers’ behaviors can be better understood. However, in the absence of appropriate control mechanisms, this information is still of relatively limited use (Levinson, 2003). Dynamic variations in the utilization of transportation capacity can result in serious traffic congestion. Congestion pricing has tried to tackle the problem from an economic disincentive point of view, while a reservation system proposes to primarily address the problem from a rationing policy perspective. “Under rationing, automobile access to portions of the road network is allowed only to users with specific needs assessed by the Public Administration. While, on the contrary, under pricing, passage through some arcs of the network is permitted only to those users willing to pay” (Gentile

* This research has been supported by National Science Foundation (Project #0527252).
* Corresponding author. Tel.: +1 703 538 8446; fax: +1 703 538 8450.
E-mail address: triantis@vt.edu (K. Triantis).

© 2010 Elsevier B.V. All rights reserved.

doi:10.1016/j.ejor.2010.01.026
et al., 2005, p. 3). It is apparent that congestion pricing is more flexible than a reservation strategy, in the sense that people are free to choose whether to travel, as long as they are willing to pay. On the other hand, flexibility generates uncertainty in the demand patterns and thus variation in the road traffic. In this case, a reservation system can be introduced as an additional mechanism to cope with this uncertainty.

In this paper, we explore the application of a reservation concept within a traffic congestion management framework. The purpose of the paper is to provide an analytical model to allow the transportation authority to make decisions as to which requests should be accepted by the system. In particular, the reservation system will be introduced for a cordon-based downtown road network, and hence it is called the downtown space reservation system (DSRS). We use the term “cordon” to define the physical boundary of the system. This means that the vehicles passing through the cordon will be the ones that will be managed by the system. In addition, considering the scope of this research, the current paper will not take into account issues related to dynamic reservation pricing since the price is assumed fixed and known for different types of vehicles and time slots.

This research contributes to the literature in two ways. First, it provides an analytical approach that demonstrates the way that a reservation system can potentially work within a congestion management context for a downtown area. Second, the proposed approach helps decision makers to consider congestion from a resource capacity limitation point of view, where the capacity of a road network is defined. Furthermore, the design and implementation of this system borrows concepts from congestion pricing, especially cordon-based congestion pricing because of some similarities between the two approaches. However, the proposed TDM strategy is not intended to substitute congestion pricing, or any of the other existing TDM strategies. Instead, together with other TDM strategies, it expands the solution domain for congestion management, especially for metropolitan areas.

The paper is organized as follows: The research context is provided in Section 1 whereas Section 2 presents a brief overview of the literature that covers earlier work on reservation systems. Section 3 provides the approach and methodology associated with the DSRS. To illustrate how the system would work in practice, a numerical example is provided in Section 4 along with a brief discussion of travel demand generation. Some implementation issues are provided in Section 5. Finally in Section 6, conclusions and future research recommendations are presented.

2. Background

Although reservation systems have been investigated for a number of industries such as airway transportation (McGill and van Ryzin, 1999; Li, 2001) along with hotels (Badinelli, 2000), for railways (You, 2008), for sea cargo (Ang et al., 2007) and for parking (Teodorovic and Lučić, 2006), it is not until the late 1990s that the concept was introduced into roadway transportation for travel demand management purposes. Nevertheless, very limited research is devoted to this topic. Akahane and Kuwahara (1996) explore the possibility of using a trip reservation system to mitigate traffic congestion at bottlenecks on holidays by adjusting driver departure times. Wong (1997) illustrates the basic concepts of a new way to manage highway traffic by a highway reservation system, where he proposes a booking procedure to govern traffic by controlling the individual driver’s departure time and associated routes by requiring the driver to reserve in advance. He stresses that even with the most advanced driver information system, it is still necessary “to promote sustainable travel patterns and improve road system performance” (Wong, 1997, p. 109). In the end, he suggests that this type of system is promising, even though it will take a long time before it could be successfully implemented. Feijer et al. (2004) perform simulation experiments to justify the advantages of a reservation system by improving travel time reliability and enhancing the effective use of road capacity. This previously described literature primarily focuses on conceptually discussing the merits of a reservation system, until Edara and Teodorovic (2008) discuss a comprehensive model for a highway reservation system, where the system is capable of allocating highway space to a variety of potential users during different time intervals and accepting/rejecting driver travel requests.

The existing literature emphasizes reservation systems for highways. The reason is understandable. As Wong (1997) claims, implementing a reservation system to a freeway should be simple, but it would be much more difficult to implement this type of system for an urban network considering the route choice and administration complexity. Nevertheless, as one explores the tradeoff between the first-best pricing strategies and second-best pricing strategies, it is not always necessary to cover all the streets in the reservation system, and each link within the network does not have to be managed individually. More and more evidence has suggested that cordon-based control can be a viable option for congestion mitigation in a downtown area.

This research differentiates itself from the literature in two aspects. First, the proposed approach is developed for a downtown network characterized by its unique infrastructure and traffic attributes, while other reservation systems only consider a segment of a highway. This requires a unique formulation of the offline optimization module and of the subsequent neural network structure as well as a different method for determining downtown network capacity based on traffic flow theory. The offline optimization module is devised to take into account the objective of maximizing people throughput in the downtown area thus increasing efficiency. Second, instead of emphasizing each origin-destination (O-D) trip and its route in the road network, the actual travel route of each vehicle is ignored, and vehicles are controlled only when passing the cordon boundary. In this way, the formulation of the DSRS is simplified.

3. The downtown space reservation system

3.1. Introduction

Under the DSRS, people who want to drive downtown send their requests to the infrastructure manager before their trips occur. A trip request includes a desired entry time, exit time, reservation price and other necessary trip attributes. Only those who get permission from the transportation authority can drive in the downtown area. From a driver’s perspective, what they need to do is to express their preference via some media (e.g., internet, etc.). The transportation authority evaluates the request against the available resources on the road network and makes the decision on whether to accept an incoming request.

The motivation of tackling the traffic congestion for a downtown area is obvious. The downtown is typically the economic engine of an urban area, and it typically suffers from even worse congestion conditions than the rest of the city. There is very little, if any, open space left between buildings and infrastructure expansion is difficult. Furthermore, the downtown area normally is the oldest part of the city. Transportation infrastructures were typically built long time ago and many of them have fewer and narrower lanes that have severely deteriorated over time. Consequently, these infrastructures do not have the ability to carry any excessive vehicles. Many people work in the area and many of them own and use private cars. Shortage of off-street parking
means that people are parking on the roads or looking for a spot and in this way increase congestion.

Generally, the road network in a downtown area consists of a great number of short links and intersections. For any origin-destination (OD) pair, there are many routes to travel along. Therefore, it is administratively prohibitive to control the traffic flow on each link individually. Furthermore, unlike a highway network that often exhibits bottleneck congestion, congestion in the downtown is typically area-wide. To solve area-wide congestion, it is not necessary to manage the traffic flow on each link individually. Therefore the cordon-based control is a worthwhile approach to investigate. The cordon is a fictional closed loop that defines where the cordon is located. It can be defined by existing roads, bridges or other geographical marks. Questions about the location of the cordon and whether it is a single-loop cordon or multiple-loop cordon have been studied in the existing literature (May et al., 2002a,b). At the initial development stage of the DRS, one could start with cities that have already implemented cordon-based travel demand management strategies such as congestion pricing (e.g., cities such as London, etc.). In terms of city size, the current technology allows one to consider any city size. Nevertheless, as a prototype, one may choose to start with a small area of the city where the level of congestion is high.

3.2. Methodology

If customers are all homogeneous, with identical products, the reservation system can be rather simple. Customer requests would be accepted until resources (products) that are stored are consumed. However in many cases, both customers and products are heterogeneous. This triggers a series of issues. How should the resources (products) be rationed among different customers? Are the products really differentiated and according to which criteria? How can the customer pool be segmented? Answers to these questions are required for the effective design of a reservation system. Each of these issues is worthy of substantial research. Yet, similar questions for airline reservation systems have been widely investigated in the revenue management literature (Pak and Piersma, 2002).

The fundamental reservation management decision is whether to accept or reject a request. As airline reservations suggest, requests in higher tariff classes are preferred, while minimizing the chance of flying with vacant seats. This supports the revenue maximization objective. If the demand is deterministic and known at any moment, there would be no problem to achieve maximum revenue. Unfortunately, demand varies with time. For the DRS, we also face the same problem, i.e., when to accept or reject a request facing unknown future demand. To solve this problem, we provide an initial formulation and solution by not considering the dynamic nature of demand. In this module, the problem is formulated as a multiple objective integer programming where different behavioral objectives are given appropriate weights. Then artificial neural network technology is used in an online module that deals with dynamic demand in real time.

3.3. The offline optimization module-model formulation

Some major assumptions are made as follows. Historical demand and reservation information are assumed attainable. This indicates that one knows the exact number of trip requests days before the actual trips take place, along with their associated attributes, such as, vehicle occupancy, scheduled trip departure time, etc. Since a reservation system does not yet exist, it is impossible to obtain the equivalent of “real” information. However, standard traffic counts, parking capacity, and census data, give an initial estimate of reservation demand and capacity. The data simulation approach will be discussed in detail along with a numerical example in Section 4.

Travel demand market can be divided into various segments. Each trip has its own unique attributes. When decision makers allocate a common resource to them, their uniqueness and commonalities have to be acknowledged. The goal that a system intends to achieve plays an important part in the demand market segmentation. As in revenue management for the airline industry, profit is the primary goal. Then customers are classified into various classes, and customers in different classes have to pay different fares. For the DRS, travel demand segmentation has to closely respond to its goal of alleviating traffic congestion and improving people's mobility, which could be achieved by increasing average vehicle occupancy (including encouraging public transit rides) while decreasing single occupancy vehicles. Hence, vehicle occupancy could be chosen as one of the criteria to segment travel demand, which in this research, includes four categories, single occupancy vehicles, high-occupancy vehicles, buses and trucks. A certain portion of the downtown network capacity needs to be allocated to each of them. The reservation price is assumed fixed and known for each type of vehicle. The price also depends on what time of the day a motorist enters and exits the congestion charging area. Travelling during peak hour intervals is assumed to be more expensive.

The notation used in this paper is as follows: Decision variables:

\[ x_j^i \]

\[ \begin{array}{ll}
  1, & \text{if jth request of demand category i is accepted} \\
  0, & \text{otherwise}
\end{array} \]

Parameters:

\[ psg_j^i \]

\[ \text{total number of passengers in the vehicle of jth request of demand category i} \]

\[ pcu_j^i \]

\[ \text{the passenger car units of jth vehicle of demand category i} \]

\[ p_m \]

\[ \text{the reservation price of time slot m for demand category i} \]

\[ s_m^i \]

\[ \begin{array}{ll}
  1, & \text{if jth request of demand category i is included in the mth time interval} \\
  0, & \text{otherwise}
\end{array} \]

\[ e_j^i \]

\[ \text{the entry time of jth request of demand category i} \]

\[ f_j^i \]

\[ \text{the departure time of jth request of demand category i} \]

\[ t_0 \]

\[ \text{the control starting time} \]

\[ \Delta t \]

\[ \text{the length of each time interval} \]

\[ C_m \]

\[ \text{capacity during time slot m} \]

\[ U_m \]

\[ \text{upper bound of the demand category i in time slot m} \]

\[ L_m \]

\[ \text{lower bound of demand category i in time slot m} \]

\[ R_j^i \]

\[ \text{jth request in demand category i} \]

Under the DRS, traffic authority personnel make decisions on who can finally drive into the downtown at their desired time under the condition that, at any moment, the total number of vehicles within the cordon area does not exceed the capacity limit (C). The total number of vehicles is the accumulation of the vehicles entering minus those departing (flow over time). We denote the entering traffic flow at time t as inflow \( f_{in}(t) \), and departing flow as outflow \( f_{out}(t) \). The total number of vehicles within the cordon area at time t, \( TV(t) \) equals:

\[ TV(t) = \int_{t_0}^{t} f_{in}(t)dt - \int_{t_0}^{t} f_{out}(t)dt \leq C. \]  

Eq. (1) ensures the capacity is not violated. This means that the inflow and/or the outflow have to be managed continuously over time. The continuous assessment of the capacity violation Eq. (1) is administratively prohibitive. Nevertheless, to make the problem manageable, let us consider \( \Delta t \) as \( \Delta t \), where \( \Delta t > d_t \). The time period of \( t_0 \) to \( t \) is divided into k time intervals \( \Delta t \). Eq. (1) can be rewritten as:
TV(t) = \sum_{i=1}^{k} f_{in}(\Delta t_i)\Delta t_i - \sum_{i=1}^{k} f_{out}(\Delta t_i)\Delta t_i \leq C. \quad (2)

Therefore, instead of dealing with a continuous formulation, the system can be monitored in a discrete manner. The magnitude of \( \Delta t \) is determined by two factors, i.e., the variation of the traffic flow with time and the desired precision of the final results. Hereafter, \( \Delta t \) is called a time slot. For illustrative purposes, let us assume the control period starts at 7:00am in the morning and ends at 7:00pm in the evening, and the entire period is divided into 1-hour intervals, which means \( \Delta t = 1 \) (Fig. 1). To facilitate the expression of Eq. (2) and make it more straightforward to understand, let us introduce the parameter \( d_{jm} \), indicating the status of a time slot with respect to a request.

\[
d_{jm}^i = \begin{cases} 
1 & \text{if} \ (e_j^i, f_j^i) \cap (t_0^i + (m - 1)\Delta t, t_0^i + m\Delta t) \neq \emptyset, \\
0 & \text{otherwise}
\end{cases} 
\quad (3)
\]

Each arrow in Fig. 1 represents a request submitted by users. For example, \( R_1 \) is the second request in class 1. For this specific trip, the driver desires to drive into the downtown area around 8:30 in the morning and leave the area around 1:30 during the same day. The time interval (8:30am to 1:30pm) when the vehicle is residing within the cordon intersects with time slots defined by \( m = 2, 3, 4, 5, 6, 7 \). Hence, there are: \( d_{2m}^1 = 1 \) for \( m = 2, 3, \ldots, 7 \), \( d_{1m}^1 = 0 \) for \( m = 1, 8, 9, \ldots, 12 \). And for each request there is a corresponding time slot vector expressed in a similar fashion as: \( d_j^i = (011111110000) \).

Then relation (2) can be reformulated as:

\[
\sum_{i=1}^{I} \sum_{j=1}^{J} d_{jm}^i \leq C_m \quad \text{for} \ m = 1, 2, \ldots, 11, 12. \quad (4)
\]

As it has been stated that \( R_2 \) is only a request, in other words, only if this request is accepted by the system, the associated requested trip will turn into an actual travel activity in the near future. Therefore, more precisely, what the decision maker is interested in:

\[
\sum_{i=1}^{I} \sum_{j=1}^{J} d_{jm}^i x_j^i \leq C_m \quad \text{for} \ m = 1, 2, \ldots, 11, 12. \quad (5)
\]

Relation (5) is the primary constraint the system is subject to. In this paper, two objectives are chosen to test the reservation procedure. On the one hand, the DRSR as a travel demand management strategy, shares a common goal with other TDM strategies, i.e., placing more people into fewer vehicles. On the other hand, revenue maximization is also included. The revenue generated potentially provides a funding source that can be used to improve roadway conditions and the performance of public transportation. However, with respect to the choice of the objective function, more work is needed to study the impacts of other objectives.

The objective function for the reservation system (multi-objective programming) is denoted as:

\[
\text{Maximize} \quad Z = \left[ Z_1(X), Z_2(X) \right], \quad (6)
\]

where

\[
Z_1(X) = \sum_{i=1}^{I} \sum_{j=1}^{J} x_j^i \quad (7)
\]

\[
Z_2(X) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{m=1}^{M} p_m^i d_{jm}^i x_j^i \quad (8)
\]

\( Z_1(X) \) stands for the objective of increasing system throughput, measured in persons, and \( Z_2(X) \) is the revenue objective, measured in dollars. There are various solution techniques for the multi-objective programming problem, such as multi-attribute utility functions, minimum distance from the ideal solution, goal programming, and prior assessments of weights (Cohon, 1978). Prior assessment of weights is adopted in this research for the sake of the ease of the interpretation of the relationship between the two objective functions. The multi-objective is actually reduced to a single-objective problem.

\[
\text{Max} \quad \sum_{i=1}^{I} \sum_{j=1}^{J} \left( \frac{psg_j^i}{pcu_j^i} \right) x_j^i + w \sum_{m=1}^{M} (p_m^i d_{jm}^i) x_j^i \quad \text{s.t.} \quad \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{psg_j^i}{pcu_j^i} x_j^i \leq C_m \quad m = 1, \ldots, T, \quad (9)
\]

\[
x_j^i = \{0, 1\}, \quad i = 1, \ldots, I; \quad j = 1, \ldots, J. \quad (10)
\]

The weight \( w \) is equivalent to the identification of a desirable tradeoff between \( Z_1 \) and \( Z_2 \), where \( w \) is measured as people/dollar and indicates the tradeoff of serving additional people in relation to one additional dollar in revenue.

The objective function (9) represents the total number of persons passing through the cordon area over the defined control
period along with the computed revenue. Each request is allowed to make only one reservation for one vehicle trip. Eq. (10) is essentially the same as the capacity constraint Eq. (5), except for the inclusion of the number of PCUs. This assumes that capacity is measured in PCUs

\[ \sum_{j=1}^{J} \frac{x_{ij} \epsilon_{m}}{d_{m}} \leq U_{m}, \quad i = 1, \ldots, I; \quad m = 1, \ldots, T, \]  

(12)

\[ \sum_{j=1}^{J} \frac{x_{ij} \epsilon_{m}}{d_{m}} \geq L_{m}, \quad i = 1, \ldots, I; \quad m = 1, \ldots, T. \]  

(13)

In addition, two auxiliary constraints (12) and (13) can be introduced to take into account some practical issues. For instance, according to the solution from the above integer programming formulation, certain types of vehicles, such as trucks may be completely eliminated from the area. Although the solution maximizes the intended objective function, in order to ensure the necessary goods are supplied, a minimum portion of downtown capacity needs to be protected for freight vehicles. The definition of the limits associated with these constraints is not determined by the reservation system but are pre-determined by the decision makers. Therefore, the two constraints are not mandatory, instead, they are proposed as auxiliary.

3.3.1. Network capacity determination

In highway management, capacity has been measured in passenger car units (pcus)/h for a highway segment. The actual capacity can be measured at a certain point of the highway by counting the number of vehicles passing through that location during the hour. For a network with one origin, one destination, and one user class, the capacity is determined by the minimum of cut-set capacities\(^1\) (Bell and Iida, 1997). For more general networks with multiple origins and destinations, determining an area-wide road network capacity is still an open question in literature. Wong and Yang (1997) measured a road network capacity using the greatest common multiplier for existing flows that could be accommodated subject to capacity constraints, cycle time and other restrictions by solving a network equilibrium problem. Yang et al. (2000) determined the capacity of urban transportation networks through equilibrium traffic assignment. Chen et al. (2002) expanded capacity analysis on the basis of reliability, and introduced the concept of capacity reliability as a network performance index in conjunction with connectivity and the reliability of travel time. In this research, we propose an alternative method of dealing with capacity determination based on the fundamental traffic flow theory based on volume, density, and speed.

Road traffic in a specific state is always characterized by flow (q), density (k) and speed (u). Flow rate represents the number of vehicles that passes a certain cross section per unit of time. Density reflects the number of vehicles per kilometer of road. Speed, flow, and density are related to each other. When the total number of vehicles operating on the link reaches a certain upper limit, the flow and speed fall to zero, and traffic stops. This point is known as jam density (k_j). On the other hand, when density is extremely low, a free flow speed (u_f) can be obtained. Several models that relate the variables q, u and k with each other have been developed based on various assumptions regarding the manner in which traffic streams flow. The most widely used of these is Greenshields equation developed in 1935 (Daganzo, 1997). Greenshields assumed that there is a linear relationship between speed and density. The fundamental relationships are used here to derive a maximum number of vehicles on a link, so that the traffic flow reaches its maximum at this point. The combination of this number for a set of links (A) is called as the network capacity hereafter.

As shown in the Appendix A, n_max which is the number of vehicles that achieves maximum throughput is given as:

\[ n_{\text{max}} = \frac{L_{\text{max}}}{2}. \]  

(14)

If the total number of vehicles moving along the link l exceeds n_max, the traffic condition deteriorates and dampsens the throughput of the system. Assuming the interactions among various links are negligible, for a network consisting of a set of links (A), the maximum flow, thus the capacity of the network is calculated as:

\[ n_{\text{max}} = \sum_{i \in A} n_i. \]  

(15)

For the downtown area, this only represents the capacity of the road network. Additionally, the parking space has to be incorporated, since it can accommodate a portion of the vehicles. Parking data can be easily obtained through a survey or field data collection over all the parking garages and street parking spots within the cordon area, with questions such as what is the maximum number of cars that can be parked in the garage, and what is the occupancy percentage during each time slot (e.g., hourly period). Therefore, at any time, one could compute the total number of vehicles in the downtown area that are parked somewhere, rather than moving on the streets. If we take this into account, and denote the number of available parking spaces during each time slot m in downtown as P_m, the maximal number of vehicles in downtown equals:

\[ C_m = n_{\text{max}} + P_m. \]  

(16)

In the above calculation, certain assumptions are made. One is that density is homogenous over the downtown network. This is a simplified assumption; in a network which is at capacity, certain links can be significantly below capacity. However, considering that we have a relative small network and in a relative crowded downtown with area-wide congestion, we obtain an upper bound network capacity based on the simplified assumption in this research. In the micro-traffic simulation that is developed in the follow-on research, we do not impose the assumption of homogenous density. Second, the interactions between links are not considered along with the influence of intersection signals. When considering all these factors, the total maximum number of vehicles ought to be less than the value obtained using the above approach. However, this approach still provides a basic reference point of downtown network capacity, and could serve as a basis for a further study of downtown capacity.

Several other directions can provide alternative methods regarding the determination of capacity. One approach is to derive capacity from the intersection capacity as transportation professionals argue that traffic speed and flow on urban streets are determined primarily by intersection capacity. Another way of dealing with the determination of capacity is to take into account the routing problem, which is essentially the traffic assignment model. The potential approaches are not confined to these. To evaluate which method would provide a more effective capacity analysis, further comparison research is needed.

3.4. Online decision making module-neural network model

If we can precisely predict the cumulative demand of all the trips over the next day, making real time decisions would not be difficult and would be achieved by solving the integer programming formulation (Section 3.3). However, demand is dynamic. In this research, in order to take into account the dynamic nature of

---

\(^1\) Cut-sets is a concept in network flow analysis. The definition of cut-set is: Let X be any set of nodes in the network such that X contains node 1 but not node m. Let \( \bar{X} = N - X \) where N is the total number of nodes in the network. Then \( (X, \bar{X}) \) = \( \{(i, j) : i \in X, j \in \bar{X}\} \) is called a cut-set separating node m from node 1 (Bazaraa et al., 1990).
demand, an artificial neural network will be used. A neural network is an adaptable system that can learn relationships through repeated presentation of empirical data, and is capable of generalizing to new data. The neural network model adopted in this paper is the multilayer perceptron (MLP), which consists of three layers with a feed forward and back propagation learning algorithm. Research (Trappenberg, 2002) shows that generally a three-layer structure can approximate any function arbitrarily well. For the DSRS, the MLP neural network is constructed as follows (Fig. 2).

The architecture of a neural network is determined by the number of hidden layers, the number of neurons in each layer, and the transfer function used. The number of neurons in the input and output layer is determined by the number of input and output variables. Two category inputs are included, specific trip attributes associated with the request (i.e., entry/exit time, vehicle class, occupancy, reservation price), and the capacity availability status from the system. One output variable is needed, i.e., the acceptance/rejection indicator, representing whether to accept or reject the request. There is no straightforward method to find the best number of neurons for a hidden layer. It is generally determined by trial and error. A transfer function is the mathematical formula used to produce the output of a neuron. One very popular nonlinear transfer function is the sigmoid function. Many neural network applications have shown that the sigmoid function should at least be a baseline model to measure results. A general rule of thumb is that the sigmoid will produce the most accurate model (Duch and Jankowski, 1999). In the current paper, we start with a sigmoid transfer function, and compare the learning curve with a hyperbolic transfer function learning curve. The numerical example in Section 4 shows that the hyperbolic transfer function outperforms the sigmoid function. Therefore, the hyperbolic transfer function is chosen in this study.

The procedure used by the neural network to calculate output from the given inputs is illustrated with a simple network (Fig. 3). As the inputs \( x_1 \) and \( x_2 \) pass through the input layer of neurons, the weights \( W_1 \) and \( W_2 \) are assigned arbitrarily by the network. The linear combiner calculates the weighted sum of these inputs, \( \text{NET} = W_1 x_1 + W_2 x_2 \). The range of \( \text{NET} \) is compressed by a transfer function such that the output signal \( Y \) never exceeds a relatively low level regardless of the value of \( \text{NET} \). It enables the receiving and processing of very weak and strong signals. Let the threshold be defined as \( \theta \). Now considering the following numerical example: \( x_1 = 0, x_2 = 1, W_1 = 0.2, W_2 = -0.1, \theta = 0.2 \), and the step function as the transfer function \( (Y = 0, \text{if } \text{NET} < \theta, \text{else } Y = 1) \). The output is calculated as follows: \( W_1 x_1 + W_2 x_2 = 0.2 \cdot 0 + (-0.1) \cdot 1 = -0.1 < 0.2 \), then \( Y = 0 \). In this way, the output is calculated by the neural network, the weights are adjusted by learning and the same procedure is followed to calculate the output using new weights.

Generally, there are two basic phases in neural network operation. In the training phase, a set of examples are shown to the network, and the learning process is realized by changing the weights to adapt to the desired outputs. After training, knowledge embedded in the examples is stored in the neural network. In the testing phase, the neural network is frozen with the fixed weights obtained from the training phase, and a new set of data is fed into the trained network to produce network output. If the testing indicates that the neural network has generalized what it has learned from the examples, the network is ready for use. If not, one may need to go back to the previous steps to revise the network parameters and repeat the process (Fig. 4).

4. A numerical example

A hypothetical example is provided. For illustration purposes, we choose downtown DC as an example. Traffic data from the Advanced Interactive Traffic Visualization System (AITVS) (Spatial Data Management Laboratory, 2007) provides a platform of real time and historical traffic pattern analysis for I-66 and I-95.
However, this example does not intend to replicate downtown DC to any extent considering the scope of the example (only 5000 trips were generated in this example).

The demand data are generated according to the available aggregate level transportation data. The major task of the generation is to simulate vehicles entering and exiting a downtown area at any moment. Thereafter, the cumulative number of vehicles residing in the area can be derived. This is an important parameter that the DSRs is aimed to control. Expanding on this notion, for a cordon-based TDM strategy, the traffic entering and exiting the area is the summation of the traffic over all the roadways that lead to the downtown (Fig. 5), as Eqs. (17) and (18) below indicate (where \( f_i \) indicates inflow along path \( i \) and \( f_o \) represents the outflow along path \( o \)).

\[
\begin{align*}
    f_{in}(t)\Delta t &= f_1(t)\Delta t + f_2(t)\Delta t + f_3(t)\Delta t + f_4(t)\Delta t, \\
    f_{out}(t)\Delta t &= f_1(t)\Delta t + f_2(t)\Delta t + f_3(t)\Delta t + f_4(t)\Delta t.
\end{align*}
\]

Taking downtown DC as an example, and assuming that traffic flow on I-66 at a location close enough to downtown DC, adequately represents the traffic inflow and outflow patterns for downtown DC, two traffic observation stations at milepost 51.2 are selected (Fig. 6). Each station is for one direction. The data acquired at the traffic observation stations are traffic volume, aggregated on average every 5 minutes (for more details on the data collection, please see http://spatial.nvc.cs.vt.edu/traffic/). The source data of ATIS are provided by Virginia Department of Transportation. As the data from ATIS shows, generally, a relatively stable pattern exists for week days and weekends. The control begins at 7:00am in the morning until 7:00pm in the evening on weekdays, and the period is divided into twelve hourly intervals. Cumulative traffic volume (in percentage) for each time interval is derived from the volume data at the inbound direction (entry). The generated demand has the same percentage distribution. The departure time for each trip depends on its entry time and staying length (departure time = entry time + staying length). Therefore, instead of generating departure time directly, staying length is generated assuming that vehicles entering during different time intervals follow different distributions. For example, for vehicles that enter the downtown area during time intervals in the morning peak hours, 40% of them will stay a length of time that is normally distributed with \( \mu = 1, \sigma = 0.1 \), and 60% of them will stay a length of time that is normally distributed with \( \mu = 8, \sigma = 0.8 \). The parameters of the normal distribution are iteratively determined by comparing the generated outbound traffic volume (departure) with the real traffic flow data (outbound) from I-66. The set of parameters are chosen to approximate the demand distribution of real traffic.

Data regarding vehicle occupancy, vehicle classes (Table 1), are approximated from two sources, the Census Transportation Planning Package (US Census Bureau, 2005) and the Virginia Department of Transportation (VDOT) Daily Traffic Volume Estimates including Vehicle Classification Estimates (VDOT, 2005). Price varies for different time intervals and vehicle types, but remains fixed within the interval. These prices are only used for illustration purposes in this example (Table 2). Since revenue maximization is one of the two objectives, the price discrimination for the peak period penalizes the peak period users by having them pay more than the non-peak period users. If the system throughput is the only objective, one could conceivably achieve the same system throughput without adjusting the reservation prices for peak periods by setting a cap for each vehicle class during each time period. The generated demand data are shown in Table 3. The multi-objective optimization programming (Eqs. (9)-(11)) is solved using AMPL/Cplex. Weight \( w \) is assumed 1/3 in this example. The optimal results are shown in last column, where 0 indicates the corresponding request is rejected, and 1 indicates it is accepted.

In the next step, solutions from the optimization problem are used to train the neural network. As stated in the Section 3.4, the neural network is trained with the sigmoid transfer function first, and then with hyperbolic tangent function. As the learning curve shows in Fig. 7, although the sigmoid function outperforms the hyperbolic tangent function at the beginning of the training, after about 150 epochs the average cost (computed as mean squared error) of the hyperbolic tangent function is lower than the sigmoid function. A lower average cost indicates a better performance for a neural network. Therefore, the hyperbolic tangent transfer function is adopted in this research.

To test the network, let us generate another set of demand data. First, in a similar fashion we solve the optimization problem using AMPL/Cplex, and save the results for future use. Then, the data are fed into the trained neural network. The outputs produced are continuous numbers. However, when making a decision, a request is accepted or rejected. The results are binary. There is no status between accept and reject. In this case, the continuous outputs from neural network have to be converted to a binary representation. For analysis purposes, we separate the data into two sets according to the optimal solution, one for accepted requests (1s), and the other for rejected requests (0s). In terms of the neural network output, Table 4 indicates that about 92.7% output from the neural network fall in the interval of [0.8, 1], about 5% in between [0.6, 0.8].

---

**Table 1**

Traffic composition by travel mode.

<table>
<thead>
<tr>
<th>Traffic composition</th>
<th>69%</th>
<th>23%</th>
<th>3%</th>
<th>5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Truck</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Similarly, for the rejected requests, 71% are in \([-1,0]\), 15% in \([0,0.2]\), and about 9% in \([0.2,0.6]\). We set the neural network output to one if the value of the output is greater than 0.5 then the associated request is accepted, otherwise rejected. Then the error rate, which represents the probability of a request being incorrectly accepted/rejected, is about 4.5%.

According to our example, the % accepted according to the offline optimal results for SOVs is about 80%, and for HOVs and buses, the % accepted is greater than 90%, and only 30% of trucks are accepted. As the results show, the high-occupancy vehicles including buses have higher chance being accepted by the DSRS, and larger vehicles with low occupancy are rejected more frequently. This echoes the major system objective that maximizes the people throughput in the system. In order to save more spaces for HOVs and buses, the system chooses to reject more trucks and SOVs. The neural network provides similar results. However, this example is intended to demonstrate how the downtown reservation space system can be potentially configured. It is not the intent of this example to provide insights of how a downtown reservation system could impact traffic congestion. This is the focus of subsequent research work.

5. Implementation issue

Although we are yet to reach the implementation phase of this system, it is still possible to identify some implementation issues that need to be considered. First of all, the implementation of any significant TDM policy requires a fundamental life style adjustment. Therefore, an appropriate channel to convey the benefits of such a system would be helpful for public acceptance. Moreover, an effective TDM, like the DSRS, is not only about designing the program itself. It involves a variety of participants within the community. Pedestrian improvements, transit improvements, alternative work schedules often enhance the synergistic result of a TDM program. One critical point is that motorists have to be provided enough opportunity to take alternative transportation modes. In London, the congestion pricing policy along with improved transit services has played an indispensable part in assuring its success. Similarly for the DSRS, it is essential to make alternative transportation modes, especially public transit, attractive. In this case, those trips without reservation (rejected trips) have three choices, e.g., using public transportation, shifting the time of travel, and eliminating the travel need (for example, instead of going to a local store to shop, utilize online shopping). The DSRS gives greater priority to high-occupancy vehicles such as carpools and transit vehicles. Special needs like emergency trips, disabled people, can be taken into account in the system design by setting aside slots for the emergency vehicles and disabled people. Some of the relevant issues are discussed below.
Reserving spaces: The communication channel(s) via which travellers can make reservation requests could include, telephone, internet, text messaging, among others. The time window available to make a request also is a key variable. For example, the reservation system can be opened 24 hours in advance and closed 1 hour prior to the actual departure times. The size of this window depends on the request processing capabilities of the DSRS server.

Early/late arrivals: In the event that a vehicle arrives at its entrance (say freeway ramp) before or after its scheduled arrival time the DSRS needs to have slack in the system to accommodate such vehicles. This can be accomplished by setting aside a certain portion of the entrance capacity to such arrivals. The early arriving vehicle can be issued a warning in the first violation with repeat offenders being ticketed and/or preventing them from making reservations for a certain number of days.

Late departures: It may happen that the departure plans of travellers change after entering the downtown region. This could happen for a variety of reasons, e.g., a meeting running late. Again, in such situations where drivers violate their original departure times may be treated the same way early/late arrivals are treated.

Enforcement: License plate numbers provided by motorists while reserving spaces can be used for enforcement purposes. These numbers can be read at highway speeds with the current technology and processed in milliseconds. If the number does not match the system accepted number for that time interval then a violation ticket is issued. To detect the number of occupants inside the vehicle, technology such as infrared and laser can be used (e.g., Forti’s road bridge in Scotland has tested such equipment for use in a variable pricing project in 2007. The system gives discounts to car pools based on the number of occupants.)

Daily commuters: People who work in the downtown area and commute on every week day can be treated in a different way. A portion of the total capacity can be devoted to drivers that are willing to make reservations for a longer period of time (e.g., few months to up to a year). Since they help reduce the uncertainty in demand, such requests can be given preferential treatment, e.g., reducing their reservation price.

Trip cancellations/no-shows: The DSRS should also establish a policy on treating trip cancellations and drivers that do not show up for their trips. Assigning penalties on future reservations is one way of dealing with it. Drivers missing their reservations for valid reasons such as a medical emergency may be allowed to show the necessary proof to get the penalty waived. It is to be noted that these are just potential ideas of how to deal with such violations.

Request rejections and re-requests: Travellers whose travel requests are turned down for a certain time interval may keep checking (re-requests) in hope of someone cancelling their trips or may choose to reserve for the next available departure time. The number of re-requests is an important factor for the design of the reservation server and should be taken into account to avoid service outages due to the overloading of the server.

Privacy issues: Travellers may be asked to provide their license plate numbers during reservations. The agency should take sufficient care not to disclose the identity of drivers. This situation is very similar to the automated tolling applications around the world and may not be a major implementation issue.

Fairness issues: For a broader sense of fairness, the DSRS is fair in the sense that drivers who reserve well in advance (and help reduce demand uncertainty) do not have to pay the same amount as the drivers reserving at a later point in time. In congestion pricing, road users belonging to the same vehicle class (e.g., single occupant vehicles) pay the same toll amount regardless of their income levels. However, the DSRS to some extent considers the income concerns of low income drivers by providing an option to reserve early. The DSRS also gives greater priority to high-occupancy vehicles such as carpools and transit vehicles (similar to congestion pricing). In addition, the DSRS could also have different policies for the needs of citizens with special needs such as the disabled or senior citizens. Furthermore, once the DSRS is implemented, one could conduct a meta analysis to study the ramifications of rejecting specific requests at the margin, i.e., requests that were equivalent in terms of their characteristics to other accepted requests but were rejected. This could provide the basis for a policy that provides some mechanism that ensures that the DSRS is fair most of the time.

6. Conclusions and future research

In this paper, we explore the applicability of reservation concept for congestion mitigation, particularly focused on the downtown of a city. Although it is called downtown space reservation system, its application is not rigorously restricted to a downtown area. Streets of a central business district resembling an area could use a similar approach. The proposed DSRS consists of two modules, an offline optimization module and an online decision making module. In the offline module, an optimization problem is solved based on historical travel information. Multiple objectives are incorporated into the optimization problem. The optimization solutions show that under known demand patterns, the system is able to pick the “best” trips and achieve the “best” system performance in terms of people throughput and revenue generation. Based on the optimal results from the offline module, an “intelligent” system is developed using neural network technology. Not only is the system able to learn and generalize from the optimal solutions, but the system is also able to update itself, as the new information arrives. With an extensive training, the “intelligent” system can facilitate decision makers to make real time decisions on whether to accept/reject a reservation request. The intelligent system may not guarantee the optimal results compared with the results from the offline module. However, for a known traffic pattern, it does show consistency. In the numerical example, one can catch a glimpse of the way that the system in principle works and to what extent the neural network resembles optimality in the offline module. Leaving aside the implementation aspects, the system can be promising.

Nevertheless, the proposed approach is based on fundamental assumptions concerning the relative importance of revenue generation versus people throughput, the calculation of downtown capacity, the configuration of the neural network, and the consideration of static versus dynamic pricing. Furthermore, the alternatives available for the requests that are rejected are also a legitimate issue that needs to be addressed in later research that focuses on the evaluation of the system. All of these issues need to be explored in more depth especially given the implementation challenges discussed in the previous section. Finally, a more comprehensive performance evaluation framework needs to be defined to assess the relative merits of the DSRS.

At last, the authors want to emphasize that the current work is an early exploration of the concept of using a downtown space reservation system for congestion mitigation. Its development and implementation have a long way to go and require that many details need to be carefully investigated. The performance of the DSRS will depend on issues such as, whether it is used alone or with other policies (e.g., congestion pricing, intelligent parking systems, etc.) and by the choice of the objective function (e.g., the degree to which one maximizes social welfare and/or the efficiency of the transportation system, etc.). All of these issues are subject to further research.
Appendix A

Greenshields assumed a number of fundamental relationships among speed (u), density (k) and flow (q) (Fig. 8). The fundamental relationships are used here to derive a maximum number of vehicles on a link. The flow (q) by definition is number of automobiles (n) passing a point over time (t). We denote speed and density respectively by u and k.

\[ q = \frac{n}{t} \]  
(19)

Travel time along observed link l equals:

\[ t = \frac{L}{u} \]  
(20)

After substituting (20) into (19) we get:

\[ q = \frac{n}{u} = n \cdot \frac{1}{u} = ku \]  
(21)

Speed-density relationship in Greenshields model is denoted as:

\[ u = u_f \left( 1 - \frac{k}{k_f} \right) \]  
(22)

Combining relations (21) and (22), we get:

\[ q = ku_f \left( 1 - \frac{k}{k_f} \right) \]  
(23)

According to the definition of density,

\[ k = \frac{n}{T} \]  
(24)

Relation (23) can be written as:

\[ q = q(n) = ku_f \left( 1 - \frac{k}{k_f} \right) = n \cdot \frac{v_f}{T} \left( 1 - \frac{k}{k_f} \right) = \frac{v_f}{T} n \left( 1 - \frac{q}{k_f} \right) \]  
(25)

In relation (25), traffic flow rate is a function of total number of vehicles in the link at that time. For any link, a maximum throughput is desired from a system view of point. To achieve this objective, maximum flow rate needs to be reached. The first order derivative of q(n) is:

\[ \frac{dq}{dn} = \frac{v_f}{T} - 2n \frac{v_f}{T} \frac{v_f}{k_f} \]  
(26)

The second order derivative:

\[ \frac{d^2q}{dn^2} = - \frac{2v_f}{T} \frac{v_f}{k_f} \leq 0. \]  
(27)

Relation (27) indicates the concavity of function q(n). Therefore we can obtain optimal solution at:

\[ \frac{dq}{dn} = \frac{v_f}{T} - 2n \frac{v_f}{T} \frac{v_f}{k_f} = 0. \]  
(28)

Solving Eq. (28) for n, we can get:

\[ n^* = \frac{k_f}{2} \]  
(29)

This is the maximum flow rate the link can reach. Under the maximum flow rate, the link achieves its maximum throughput. The combination of this number for a set of links (A) is called as a network capacity.

References


Virginia Department of Transportation, 2005. Daily traffic volume estimates including vehicle classification estimates, Special Locality Report 100, City of Alexandria.