A Mixed Integer-Linear Programming Model for Dynamic Traffic Assignment

David E. Kaufman
Jason Nonis
Robert L. Smith

University of Michigan
Ann Arbor
A Mixed Integer-Linear Programming Model for Dynamic Traffic Assignment

David E. Kaufman
Jason Nonis
Robert L. Smith

University of Michigan
Ann Arbor

IVHS Technical Report: #92-13
A Mixed Integer-Linear Programming Model for Dynamic Traffic Assignment

David E. Kaufman
Jason Nonis
Robert L. Smith
Department of Industrial and Operations Engineering
University of Michigan, Ann Arbor, Michigan 48109

July 30, 1992

1 Introduction

We study a problem of dynamic traffic assignment, in which we seek to determine the time-dependent traffic volumes and link travel times that occur in a spatial traffic network with known topology and time-dependent origin-destination travel demands. We discretize the finite interval of time to be studied and model traffic as a continuous-valued multicommodity flow in a time-expanded version of the spatial network. Traffic congestion is modelled by linear capacity constraints, with upper bounds on instantaneous link congestion determined by impedance functions. It is assumed that all vehicles that enter a link during a single time period exit the link together during a single later time period; this is modelled by multiple-choice constraints on 0-1 variables which correspond to the time-dependent travel arcs.

We model the flows that would occur under cooperative behavior, i.e., dynamic system optimality, by attaching total network travel time as an objective function to be minimized. We also present variations on the model that model dynamic user equilibrium, by approximate objective functions and by a more exact criterion which requires additional 0-1 variables.

We present a general linear dynamic impedance model as an extension of static impedance functions. The model can be calibrated off-line from data on time-dependent link entry and exit flows, for example, data collected by loop detectors placed at the beginning and end of spatial links. We impose requirements on the calibration process which ensure that link congestion behaves under stationary dynamic traffic flows as it would under static flows.
2 The system-optimal time-expanded network model

2.1 Network definition

Let $G = (N, A)$ be the traffic network with node set $N$ and link set $A$. We model the problem with a horizon of $h$ periods, each of duration $\Delta t$, as a time-expanded version of $G$, $G(h) = (N, A)$. Corresponding to each node $x$ of $N$, $N$ has $h + 1$ nodes $x(\tau)$, $\tau = 0, 1, \ldots, h$. Corresponding to arc $(x, y)$ of $A$, $A$ has arcs $(x(\tau), y(\tau + s))$, $\tau = 0, \ldots, h - 1$. The time expansion is illustrated in figure 1. Assuming that $N$ contains no self-loops $(x, x)$, we prevent stalling at nodes by not including any arcs $(x(\tau), x(\tau + s))$ in $G(h)$. (Stalling arcs would appear as horizontal arrows in figure 1.)

We represent vehicle flows by nonnegative continuous-valued flow variables $f_d(x(\tau), y(\tau + s))$, the volume of traffic with final destination $d \in N$ which enters link $(x, y)$ at time $\tau$ and exits at time $\tau + s$. We represent the total such flow irrespective of destination by

$$f(x(\tau), y(\tau + s)) = \sum_{d \in N} f_d(x(\tau), y(\tau + s)).$$

We adopt the general notation $g(Z) \equiv \sum_{z \in Z} g(z)$ for vectors $g$ and sets $Z$, and we define $S = \{1, 2, 3, \ldots\}$ and $\bar{S} = S \cup \{0\}$. Sums over time expressed in this way will have upper limits apparent from context, determined by the time horizon. Multiple sets occurring in a single expression denote multiple summation; for example, the system-optimal objective function will include terms

$$f(x(\tau - \bar{S}), y(\tau + S)) \equiv \sum_{u=0}^{\tau} \sum_{s=1}^{\tau - u} f(x(\tau - u), y(\tau + s))$$
which give the number of vehicles on link \((x, y)\) during the \(\tau^{th}\) time period.

### 2.2 Traffic modelling assumptions and implied constraints

We now propose modelling assumptions governing vehicle dynamics and translate them into mathematical programming constraints.

I. **No dispersion of platoons within links.** We assume that all vehicles entering a link in a single period \(\tau\) experience the same link travel time. We enforce the assumption by introducing 0-1 integer variables \(\delta(x(\tau), y(\tau + s))\) for each \((x(\tau), y(\tau + s)) \in \mathcal{A}\). A vehicle platoon entering link \((x, y)\) at time \(\tau\) experiences link travel time \(s\) only if \(\delta(x(\tau), y(\tau + s)) = 1\), by virtue of the constraints

\[
 f(x(\tau), y(\tau + s)) \leq M\delta(x(\tau), y(\tau + s)) \quad (x(\tau), y(\tau + s)) \in \mathcal{A}
\]

where \(M\) is an arbitrarily large constant. We prevent dispersion of the platoon by the **multiple choice constraints**

\[
\delta(x(\tau), y(\tau + S)) \leq 1 \quad (x, y) \in \mathcal{A}, 0 \leq \tau < h
\]

permitting at most one link travel time \(s\) (i.e. one arc \((x(\tau), y(\tau + S))\)) to be chosen from \(S\) for link \((x, y)\) at time \(\tau\). Note that if no vehicles enter link \((x, y)\) at \(\tau\), then it is feasible not to choose any arc, i.e., \(\delta(x(\tau), y(\tau + S)) = 0\). It might be preferable to require a choice so that we would know what link travel time would occur if some increment of flow were rerouted onto \((x, y)\), but the corresponding revision of (2) as strict equality constraints has a drawback, illustrated below, in connection with our next assumption.

II. **Link consistency.** We assume that vehicles do not pass one another, i.e., that among two platoons traversing a link, the one that enters later does not leave earlier. An inconsistent arc choice is illustrated in figure 2(a).

By observing that a nonzero traffic flow entering \((x, y)\) at \(\tau\) experiences link travel time \(\sum_{s=1}^{h-\tau} s\delta(x(\tau), y(\tau + s))\), we write the consistency constraints

\[
\tau + \sum_{s=1}^{h-\tau} s\delta(x(\tau), y(\tau + s)) \leq \omega + \sum_{s=1}^{h-\omega} s\delta(x(\omega), y(\omega + s))
\]

for all \((x, y) \in \mathcal{A}, 0 \leq \tau < \omega < h\) such that \(\delta(x(\omega), y(\omega + S)) = 1\).

The constraint cannot be applied when \(\delta(x(\omega), y(\omega + S)) = 0\) (whereas, if \(\delta(x(\tau), y(\tau + S)) = 0\), it is vacuous). This restriction is technically nonlinear, but has the simple linear
reformulation
\[
\tau + \sum_{s=1}^{h-\tau} s\delta(x(\tau), y(\tau + s)) \leq \omega + \sum_{s=1}^{h-\omega} s\delta(x(\omega), y(\omega + s)) + M(1 - \delta(x(\omega), y(\omega + S)))
\]
\[(x, y) \in A, 0 \leq \tau < \omega < h\]

for a suitably large value \(M\).

We can now explain why we have chosen the inequality form of the multiple-choice constraint by discussing the example in Figure 2(b). Consider link \((x, y)\) loaded by only those platoons shown descending vertically in the figure. Without jumping ahead and explaining congestion modelling, let us say that the platoons of 15 and 5 vehicles entering empty link \((x, y)\) at times \(\tau - 2\) and \(\tau + 1\), respectively, have link travel times of 3 and 1 periods, respectively. If, as permitted by one form of the congestion model we will present, link travel times are determined simply by the number of vehicles anywhere on the link at the moment of entry, then the zero-flow platoon entering at time \(\tau\) would see a total of 15 vehicles on the link and choose a link travel time of three periods (shown by the dashed arc) if the equality multiple-choice form were in force. To resolve the resulting inconsistency, we would have to delay the link exit of the five-vehicle platoon by one period. We have used the inequality form to prevent this "phantom-vehicle" delay.

III. Finite horizon addressed with fixed trip-completion penalties. We will be permitting vehicles to enter the network throughout the study horizon, so vehicles that enter near time \(h\) may not be able to finish their trips by \(h\). We require that these vehicles occupy some time-expanded node \(y(h)\) at time \(h\), instead of being left out in the middle of a time-expanded arc. This will be accomplished by leaving all arcs of the form \((x(\tau), y(h))\) uncapacitated.
We then penalize each vehicle that failed to complete its trip to node $d$ and instead finished at $y(h)$ by an estimate $\beta(y, d)$ of the travel time required to finish the trip.

Thus the complete system-optimal objective function is

$$\min_{\tau=0}^{h-1} \sum_{(x,y) \in A} f(x(\tau - \delta), y(\tau + \delta)) + \sum_{d \in N} \sum_{(x,y) \in A} \beta(y, d) f_d(x(h - \delta), y(h)) \quad (5)$$

where the first term expresses the actual total travel time in the system (as explained in section 2.1), and the second term adds the trip completion penalties.

IV. Flow conservation except at trip completion. We require that for vehicles with any particular destination $d$, the number departing any time-expanded node $x(\tau)$ ($x \neq d$) is equal to the number entering $x(\tau)$ plus the number that begin their trips at $x(\tau)$. Given that $v_d(x(\tau))$ vehicles enter the network at $x(\tau)$ with destination $d$, we write the corresponding constraints as

$$f_d(x(\tau), N(\tau + S)) - f_d(N(\tau - S), x(\tau)) = v_d(x(\tau)) \quad x, d \in N, x \neq d, 0 \leq \tau < h. \quad (6)$$

We omit conservation constraints for $\tau = h$ since vehicles still on the network at time $h$ are handled via trip-completion penalties. And we omit constraints for $x = d$, allowing vehicles that reach their destinations to drop out of the network and stop increasing the system-optimal objective function.

V. Linear congestion modelling. We now model the delay caused by increasing traffic loads. Our congestion modelling is determined by three assumptions:

V.1 Feasible link travel times for a platoon entering link $(x, y)$ at time $\tau$ depend only on the volume of traffic (and perhaps its spatial distribution) on $(x, y)$ at time $\tau$.

V.2 The dynamic model of link travel time, applied to a traffic network in steady state, agrees with standard static models.

V.3 Congestion can be modelled linearly by constraints

$$f(x(\tau), y(\tau + s)) + \sum_{u=1}^{\tau} \sum_{v=1}^{s} \alpha_{x,y}(u,v) f(x(\tau - u), y(\tau + v)) \leq c(x(\tau), y(\tau + s)) \quad (7)$$

for all $(x(\tau), y(\tau + s)) \in A, \tau + s < h$ such that $\delta(x(\tau), y(\tau + s)) = 1$

where $c$ and $\alpha$ are various externally determined coefficients.
Assumption V.1 is reflected in the structure of constraint (7), which only takes into account flows \( f(x(\tau - u), y(\tau + v)) \), which entered \((x, y)\) before \( \tau \) and will exit after \( \tau \). Assumption V.2 affects the possible choices for \( c \) and \( \alpha \); we defer this topic to section 3.

Note that by summation over \( s \), (7) is equivalent to (but has \( O(h) \) times as many constraints as)

\[
\begin{align*}
  f(x(\tau), y(\tau + S)) & + \sum_{u=1}^{h-\tau} \sum_{v=1}^{r-1} \alpha_{xy}(u, v) f(x(\tau - u), y(\tau + v)) \\
  & \leq \sum_{s=1}^{h-\tau} \delta(x(\tau), y(\tau + s)) c(x(\tau), y(\tau + s)) 
\end{align*}
\]

(8)

for all \((x, y) \in A, 0 \leq \tau < h \) such that \( \delta(x(\tau), y(\tau + S)) = 1 \)

which itself must be implemented using the big-\( M \) value as were constraints (3). Also, the compressed version requires \( c(x(\tau), y(h)) = M \) for all \((x, y) \in A, 0 \leq \tau < h \) so that arcs ending at time \( h \) are uncapacitated, as specified under Assumption III.

The case \( \alpha \equiv 1 \) is of special interest. We then have the distribution-independent congestion constraints

\[
\begin{align*}
  f(x(\tau), y(\tau + S)) + f(x(\tau - S), y(\tau + S)) & \leq \sum_{s=1}^{h-\tau} \delta(x(\tau), y(\tau + s)) c(x(\tau), y(\tau + s)) 
\end{align*}
\]

(9)

under which a platoon entering \((x, y)\) at time \( \tau \) is permitted link travel time \( s \) only if the total number of vehicles already on the link \([f(x(\tau - S), y(\tau + S))]\) (regardless of their location on the link) plus the number entering \([f(x(\tau), y(\tau + S))]\) is no larger than \( c(x(\tau), y(\tau + s)) \), which we call the capacity for the entering platoon and which is presumed increasing in \( s \). Strictly speaking, \( c(x(\tau), y(\tau + s)) \) is not a capacity for arc \((x(\tau), y(\tau + s))\) in the usual network programming sense, but is, instead, the capacity for the spatial link \((x, y)\) during period \( \tau \), given that flows entering at time \( \tau \) will have link travel time \( s \).

2.3 The mathematical program

The mixed integer-linear mathematical program for system-optimal dynamic traffic assignment consists of objective function (5) subject to non-dispersion ((1) and (2)), consistency ((3) or (4)), flow conservation (6), and linear congestion modelling ((8) or (9), or the corresponding big-\( M \) true linear form), given the decision variables as specified in section 2.1.

3 Coefficients for the linear congestion model

In the literature of dynamic traffic modelling, no consensus has been reached on how vehicle speeds should be determined. McGurrin and Wang [7] have constructed a microscopic
traffic simulation built on car-following models which would not produce link travel times in agreement with our non-dispersion assumption. Furthermore, car-following models do not have equivalent or closely approximating macroscopic forms. Chen and Mahmassani [4] and Van Aerde [10] have constructed more macroscopic simulations in which vehicle speeds are determined by static speed functions applied to some average congestion level, but these approaches may produce link inconsistency and they ignore much of the dynamic character of the problem. Janson [6] takes a similar approach in a nonlinear integer programming model with a mixed discrete/continuous time character; the model may suffer from inconsistency and the discretization is based on relatively long intervals, losing the dynamic character of congestion within links. Merchant and Nemhauser [8], Carey [2,3], Friesz et al [5] and Wie et al [11], and Boyce et al [1] have proposed models where nonunique (i.e. dispersionary) link travel times are determined implicitly by link outflow functions, which give the number of vehicles to depart a link as a function of instantaneous link volume. The link outflow function approach ignores nonuniform distribution of traffic on links and may manifest increased flow rates when congestion increases. Further, the construction and calibration of link outflow functions have not been addressed.

Our linear congestion model constitutes a new proposal for dynamic congestion modelling. In this section, we demonstrate that the external coefficients of the linear form can be shown to satisfy a simple steady-state relationship, which completely determines the capacity coefficients in the distribution-independent case and restricts the admissible coefficients in the general (distribution-dependent) case. We propose a calibration method which determines the congestion coefficients in the general case, requiring only time-series data on the entries to and exits from a link, by solving a single convex mathematical program.

### 3.1 Distribution-independent model of a link in steady state

Under assumption V.2, we require that for a link in steady state with constant inflow rate and total volume over time, the dynamic model predicts the same link travel time that would arise in static modelling. Static traffic modelling provides speed functions \( \sigma_{xy}(\cdot) \), which give the vehicle speed on link \((x, y)\) as a function of the time rate of flow across the link, assumed constant over time. We assume that \( \sigma \) is a positive, continuous, decreasing function.

To make the dynamic model act as a true generalization of the static model, we construct a steady-state loading in our time-expanded network \( G(h) \) where \( \lambda \) vehicles enter link \((x, y)\) in each period. The speed on \((x, y)\) is therefore \( \sigma_{xy}(\lambda) \), constant over time. We denote the physical length of \((x, y)\) by \( d(x, y) \), and thus

\[
\sigma_{xy}(\lambda) = \frac{d(x, y)}{s}
\]

where \( s \) is the steady-state link travel time. The speed function \( \sigma \) has a well-defined inverse,
and thus by a trivial application of Little’s law for conservative systems (the link always contains s platoons of size \(\lambda\)), the steady-state number of vehicles on the spatial link \((x, y)\) is \(s\sigma^{-1}_{xy}(d(x, y)/s)\). This fact determines the values of the capacity coefficients \(c\) in the distribution-independent case, except for time-expanded arcs ending at time \(h\), which we still leave uncapacitated.

**Theorem 1** In the distribution-independent linear congestion model (constraint (9)), Assumption V.2 implies

\[
c(x(\tau), y(\tau + s)) = s\sigma^{-1}_{xy} \left( \frac{d(x, y)}{s} \right) \quad (x(\tau), y(\tau + s)) \in \mathcal{A}, \tau + s < h.
\]

**Proof:** Within the proof, we suppress notation for link \((x, y)\) and time \(\tau\). For any value of \(c(s)\), it is feasible under constraint (9) for \(\lambda = c(s)/s\) vehicles to enter (and depart) the link in each period, with constant link travel time \(s\). The argument preceding the theorem shows that \(\lambda = \sigma^{-1}(d/s)\) must be feasible, hence \(c(s) \geq s\sigma^{-1}(d/s)\). If \(c(s) > s\sigma^{-1}(d/s)\), we find that \(\sigma(\lambda) < d/s\). Then the link travel time is strictly greater than \(s\), a contradiction proving the desired result. \(\blacksquare\)

### 3.2 Distribution dependence

It is easy to see why we may wish to apply congestion coefficients \(\alpha \neq 1\). In the distribution-independent case, link travel times depend on the number of vehicles already on the link without reference to when they entered or how far they have travelled along the link. It seems reasonable to expect that platoons that have entered the link more recently and are nearer the beginning of the link tend to impede an entering platoon more than traffic far downstream, which is about to exit the link. In this event, measuring the congestion on a link strictly by the number of vehicles present is clearly inappropriate.

In the distribution-dependent case, theorem 1 generalizes to a relationship between \(\alpha\) and \(c\), which leaves all of the coefficients free within a restricted set of admissible values.

**Theorem 2** In the general linear congestion model (constraint (7)), assumption V.2 implies

\[
c(x(\tau), y(\tau + s)) = \left[ 1 + \sum_{u=1}^{s-1} \alpha_{xy}(u, s - u) \right] \sigma^{-1}_{xy}(d(x, y)/s) \quad (x(\tau), y(\tau + s)) \in \mathcal{A}, \tau + s < h.
\]

**Proof:** Substituting into (7) the steady state case of \(\lambda\) vehicles entering per period with link travel time \(s\) periods yields

\[
\lambda + \sum_{u=1}^{s-1} \alpha_{xy}(u, s - u) \lambda \leq c(x(\tau), y(\tau + s))
\]
constraining the flow rate $\lambda$. As in theorem 1, we find $c$ by substituting in the largest flow rate which has a link travel time of $s$ periods according to the static speed function $\sigma_{xy}$, namely $\lambda = \sigma_{xy}^{-1}(d(x, y)/s)$, proving the theorem.

Theorem 1 now appears as a special case by substituting $\alpha \equiv 1$. In the general case, we are free (subject to theorem 2) to fit the congestion coefficients to empirical data. For the calibration procedure, suppose our data consists of a time series of platoon sizes $\hat{f}_t$ entering link $(x, y)$ at time $\tau$ and their link travel times $\hat{s}_\tau$, over some very long study horizon $\tau = 0, \ldots, T$. (We defer discussion of how to collect this data.) We assume that the physical condition of $(x, y)$ does not change over the study horizon, i.e., no incidents blocking flow, so that we may consider $c(x(\tau), y(\tau + s)) \equiv c_{xy}(s)$ independent of $\tau$. We expect actual vehicles to proceed at the fastest speed permitted by the congestion model, so we expect to find for each time $\tau$ that

\begin{align}
  c_{xy}(\hat{s}_\tau - 1) &< \hat{f}_\tau + \sum_{u \geq \tau - \hat{s}_\tau} \alpha_{xy}(u, \hat{s}_\tau - u) \hat{f}_{\tau-u} \\
  &\leq c_{xy}(\hat{s}_\tau) 
\end{align}

(10) (11)

since the second inequality expresses constraint (8) for the observed data, and if the first inequality were violated, we would expect the flow $\hat{f}_\tau$ to have chosen the feasible link travel time $\hat{s}_\tau - 1$. The extent to which (10) and (11) are violated measures the inaccuracy of the linear congestion model with coefficients $\alpha$ and $c$. Let $\hat{\eta}_\tau$ and $\hat{\eta}_\tau^{+}$ be the values of the violations of (10) and (11) respectively (note that for each $\tau$ at most one of these is positive). Then we propose calibration program $(P)$ to find congestion coefficient values minimizing the total error.

\begin{align}
  \text{min} & \quad \sum_{\tau=0}^{T} (\hat{\eta}_\tau + \hat{\eta}_\tau^{+})^a \\
  \text{s.t.} & \quad \hat{\eta}_\tau, \hat{\eta}_\tau^{+} \geq 0 \quad 0 \leq \tau \leq T \quad (P.1) \\
  & \quad \hat{\eta}_\tau \geq c_{xy}(\hat{s}_\tau - 1) - \left( \hat{f}_\tau + \sum_{u=1}^{\tau} \alpha_{xy}(u, \hat{s}_\tau - u) \hat{f}_{\tau-u} \right) \quad 0 \leq \tau \leq T \quad (P.2) \\
  & \quad \hat{\eta}_\tau^{+} \geq \left( \hat{f}_\tau + \sum_{u=1}^{\hat{s}_\tau - 1} \alpha_{xy}(u, \hat{s}_\tau - u) \hat{f}_{\tau-u} \right) - c_{xy}(\hat{s}_\tau) \quad 0 \leq \tau \leq T \quad (P.3) \\
  & \quad c_{xy}(s) = \left[ 1 + \sum_{u=1}^{s-1} \alpha_{xy}(u, s-u) \right] \sigma_{xy}^{-1} \left( \frac{d(x, y)}{s} \right) \quad 1 \leq s \leq T \quad (P.4) \\
  & \quad c_{xy}(s_1) \leq c_{xy}(s_2) \quad 1 \leq s_1 < s_2 \leq T \quad (P.5) \\
  & \quad \alpha_{xy}(u_1, v) \geq \alpha_{xy}(u_2, v) \quad 1 \leq u_1 < u_2 \leq T \quad (P.6)
\end{align}

9
\[
\begin{align*}
\alpha_{xy}(u, v_1) & \leq \alpha_{xy}(u, v_2) & 1 \leq v_1 < v_2 \leq T \quad (P.7) \\
\alpha_{xy}(u, v) & \geq 0 & 1 \leq u, v \leq T \quad (P.8) \\
c_{xy}(s) & \geq 0 & 1 \leq s \leq T \quad (P.9)
\end{align*}
\]

for some choice of the error exponent \( a \). If \( a \geq 1 \), (P) is convex. In particular, (P) is solvable as a single linear program with \( a = 1 \), and \( a = 2 \) makes the objective function differentiable. Within the calibration problem, constraints 1-3 ensure that the \( \eta \) variables take on the correct error values, constraint 4 enforces theorem 2, constraints 5, 8, and 9 are needed for obvious reasons, and constraints 6 and 7 express the intuitive notion that if a platoon downstream on the link entered more recently \( (u_1 < u_2) \) or will take longer to exit the link \( (v_2 > v_1) \), it is presumably nearer the beginning of the link and will more strongly impede an entering flow, thus deserving a higher \( \alpha \) value.

Note that the size of this problem is of relatively little importance. We would like to be able to solve the assignment MILP rapidly to use in real-time traffic control, but the congestion coefficients can be seen as independent of time and traffic demands and dependent only on the physical condition of the link itself. Thus we only need solve (P) for a few scenarios (all lanes open, one lane blocked, shoulder blocked, ...) for each link off-line and keep the corresponding coefficient values handy to plug into the assignment MILP.

### 3.3 Dynamic speed functions

We have based our calibration technique for dynamic congestion modelling on agreement with static speed functions. We can generalize this approach to produce dynamic speed functions, which can then be used to improve the accuracy of macroscopic traffic simulation.

Define the state of link \((x, y)\) at time \( \tau \) as \( g_{xy}(\tau) = ((f_0, s_0), (f_1, s_1), \ldots, (f_{\tau-1}, s_{\tau-1}) \), where for each time \( \omega = 0, \ldots, \tau - 1 \) \( f_\omega \) gives the size of the entering platoon during period \( \omega \) and \( s_\omega \) is its link travel time. Let the dynamic speed function \( \psi_{xy}(f_\tau, g_{xy}(\tau)) \) be the speed of a platoon of size \( f_\tau \) entering \((x, y)\) which is in state \( g_{xy}(\tau) \) at \( \tau \). As an example, consider

\[
\psi_{xy}(f_\tau, g_{xy}(\tau)) = \left\{ \frac{d(x, y)}{s} : \frac{d(x, y)}{s} = \sigma_{xy} \left( \frac{f_\tau + \sum_{\omega=0}^{\tau} f_\omega}{s} \right) \right\}
\]

which is well defined because the set has exactly one element for \( \sigma_{xy} \) continuous and decreasing. The numerator of the argument of \( \sigma \) is simply the number of vehicles \( F \) that will be on the link once \( f_\tau \) enters. This speed function implicitly assumes that the platoons are uniformly distributed on the link, and finds the link travel time \( s \) (perhaps not integer) corresponding to a steady-state flow rate of \( F/s \). In the MILP this choice of \( \psi_{xy} \) leads directly
to the distribution-independent model (9). The dynamic speed function corresponding to the distribution-dependent model is

$$\psi_{xy}(f, g_{xy}(\tau)) = \left\{ \frac{d(x, y)}{s} : \frac{d(x, y)}{s} = \sigma_{xy} \left( \frac{\mu + \sum \alpha_{xy}(\tau - \omega, \omega + s_{\omega} - \tau) f_{\omega}}{c_{xy}(s)} \right) \right\}.$$  

Furthermore, there is no need to preserve the linearity of the congestion form for use in simulation if nonlinear forms are found to agree better with empirical data. The denominator of the argument of \( \sigma \) would then be determined by a nonlinear generalization of theorem 2.

### 3.4 Data collection

We now suggest a simple way to gather empirical data consisting of platoon sizes \( \hat{f} \) and link travel times \( \hat{s} \) for the calibration procedure. The data is collected by detectors that record the times at which any vehicle passes the detector. One such detector is placed at the point of entry to the link and one at the exit. The raw data can then be examined as a graph of the cumulative number of link entries and exits over time, as in figure 3. We can read off the link travel time of each vehicle to traverse the link from the graph, as pictured for vehicle \( k \), if we make the simplifying assumption that vehicles enter and exit the link.
in first-in first-out (FIFO) order, i.e., link consistency is satisfied. When the data does not support the non-dispersion assumption, i.e., when vehicles which enter the link in the same period depart in different periods, we combine the various possible link travel times into a single value for \( s \) by averaging or some other rule.

The FIFO assumption will be violated somewhat due to vehicles passing one another, but much more strongly due to vehicles making stops mid-link. Unless we place detectors extremely close to each other in areas with many entry/exit points (shopping malls, for example) then stopped time in parking lots may wrongly be counted as vehicle delay. Special techniques may be required for these areas.

4 User-optimality modifications

We now discuss modifications to the assignment MILP which model noncooperative driver behavior, under which each driver seeks to minimize his personal trip duration regardless of the effect on total system time.

Definition 1 An assignment is said to be user-optimal if for any time-expanded node \( x(\tau) \) \( (\tau < h) \) and destination \( d \), all vehicles entering at \( x(\tau) \) with destination \( d \) experience the same trip duration, and that duration is no greater than the duration of any path not chosen by any vehicles from \( x(\tau) \) to \( d \).

In a more complete treatment, we might describe this rather restrictive definition as strong user-optimality. A definition of weak user-optimality would allow different durations to be experienced for the same trip from \( x(\tau) \) to \( d \), so long as among all paths that are used by some flow from \( x(\tau) \) to \( d \), all except the longest path in duration are full, i.e., prevented from accepting additional flow with the same trip duration by the congestion constraints. Due to the capacitated nature of the problem, strong user-optimal solutions may not exist. However, we will be introducing user-optimality considerations into the objective function and not the constraints of the MILP, so we should find near-optimal solutions when no user-optimal solution exists.

We begin by adding constraints to the MILP which effectively calculate the optimal paths and trip times in the network, and then we write a new objective function to be minimized, consisting of penalties on suboptimal individual route choice.
4.1 Endogenous travel time determination

We denote by $a_{xy}(\tau)$ the link travel time experienced by a platoon entering $(x, y)$ at time $\tau$, and note that this can be calculated by

$$a_{xy}(\tau) = \sum_{s=1}^{\infty} s \delta_{xy}(\tau, \tau + s)$$

as in the consistency constraints of the MILP. Henceforth we assume that $\delta(x(\tau), y(\tau + S)) = 1$ for all $(x, y) \in A$, $0 \leq \tau < h$ so that this link travel time is never zero.

We can now construct optimal-path costs through dynamic programming (DP) constraints. We introduce $\pi_x^d(\tau)$ to represent the optimal trip duration from $x$ to $d$ beginning at time $\tau$. We would like enforce the interpretation by writing anticipatory DP constraints

$$\pi_x^d(\tau) \leq a_{xy}(\tau) + \pi_y^d(\tau + a_{xy}(\tau)) \quad (x, y) \in A, d \neq x, 0 \leq \tau < h$$

with the boundary conditions $\pi_x^d(\tau) = 0$ for $d \in N, 0 \leq \tau \leq h$. However, this form is nonlinear due to the variable argument of $\pi_y^d$.

Instead, we proceed by static DP in the time-expanded network. We associate with each time-expanded arc $(x(\tau), y(\tau + s))$ an arc cost $s + M(1 - \delta_{xy}(\tau, \tau + s))$. We then constrain

$$\pi_x^d(\tau) \leq s + M(1 - \delta_{xy}(\tau, \tau + s)) + \pi_y^d(\tau + s) \quad (x(\tau), y(\tau + s)) \in A, d \neq x$$

with the same boundary conditions. Since $(1 - \delta_{xy}(\tau, \tau + s)) = 0$ for exactly one $s$, only one of these constraints is in force for any combination of $(x, y, \tau)$. For that $s$, $a_{xy}(\tau) = s$ and thus constraints (13) are equivalent to constraint (12). The drawback is that (13) adds $O(T)$ times as many constraints as (12).

4.2 User equilibrium formulation

We introduce $\eta_{xy}^d(\tau)$ to represent the slacks in constraint (12):

$$\eta_{xy}^d(\tau) = [a_{xy}(\tau) + \pi_y^d(\tau + a_{xy}(\tau))] - \pi_x^d(\tau) \quad (x, y) \in A, d \neq x, 0 \leq \tau < h.$$  

$\eta_{xy}^d(\tau)$ is the delay incurred by beginning a trip from $x(\tau)$ to $d$ on link $(x, y)$ instead of choosing the fastest path. (Although $\eta$ are incorporated as continuous variables, in fact they will always take integer values.) Then $(x, y)$ is an optimal next link from $(x, \tau)$ to $d$ if and only if $\eta_{xy}^d(\tau) = 0$. Therefore, user equilibrium is characterized by the complementarity expressions

$$f_{xy}^d(\tau, \tau + S) \cdot \eta_{xy}^d(\tau) = 0 \quad (x, y) \in A, d \neq x, 0 \leq \tau < h.$$  

As we noted above, we have no assurance that a solution satisfying (15) exists. Therefore we move the complementarity terms into the objective function. We define the new problem $(UE)$ differing only from the system-optimal MILP in that it

13
• includes constraints (13) and the following linear version of (14):

\[
\begin{align*}
\eta^d_{xy}(\tau) & \leq s + \pi^d_y(\tau + s) - \pi^d_x(\tau) + M(1 - \delta_{xy}(\tau, \tau + s)) \quad (x(\tau), y(\tau + s)) \in A, \, d \neq x \\
\eta^d_{xy}(\tau) & \geq s + \pi^d_y(\tau + s) - \pi^d_x(\tau) - M(1 - \delta_{xy}(\tau, \tau + s)) \quad (x(\tau), y(\tau + s)) \in A, \, d \neq x
\end{align*}
\]

• has an entirely new nonlinear objective function

\[
\min \sum_{(x,y) \in A} \sum_{d \neq x} \sum_{\tau=0}^{h-1} f^d_{xy}(\tau, \tau + S) \cdot \eta^d_{xy}(\tau). \tag{16}
\]

It is evident that a feasible solution to \((UE)\) is user-optimal if and only if it has value zero in (16).

4.3 Solution methods

4.3.1 Search in discrete space

One general approach to solving \((UE)\) is to treat the problem as having a purely integer domain of arc choices, i.e., complete choices for the 0-1 \(\delta\) variables, and to search in this domain by methods like simulated annealing or tabu search. The unique advantage these methods have in solving \((UE)\) is that having chosen a \(\delta\) solution, we can dispense with constraints (12), (13), and (14) and calculate the values of \(\eta\) directly, which makes it possible to evaluate the current \(\delta\) solution by substituting \(\delta\) into the constraints and \(\eta\) into the objective function and solving a single linear program. The disadvantage of search methods applied to \((UE)\) is the exceptional size of the search space as a function of the network size and the length of the time intervals.

4.3.2 MILP reformulation with surrogate criterion

We cannot transform \((UE)\) into a MILP, but we can so transform a related problem. We add 0-1 variables \(\phi\) to \((UE)\) to act as the transformation of \(\eta\) to 0-1:

\[
\phi^d_{xy}(\tau) = \begin{cases} 
0 & \text{if } \eta^d_{xy}(\tau) = 0 \\
1 & \text{if } \eta^d_{xy}(\tau) \geq 1
\end{cases}
\]

We enforce this relationship by adding constraints

\[
\phi^d_{xy}(\tau) \geq \frac{1}{M} \eta^d_{xy}(\tau) \quad (x, y) \in A, \, d \neq x, \, 0 \leq \tau < h.
\]

14
We will replace objective function (16) by

$$\min \sum_{(x,y) \in A} \sum_{d \neq z} \sum_{\tau=0}^{h-1} f_{xy}^d(\tau, \tau + S) \cdot \phi_{xy}^d(\tau).$$  \hfill (17)

As in (UE), a feasible solution satisfies user equilibrium if and only if it has value zero in the new objective function. We make this linear by introducing continuous variables $\gamma_{xy}^d(\tau)$ for all $(x, y) \in A$, $d \neq x$ and $0 \leq \tau < h$, each constrained by

$$\begin{align*}
\gamma_{xy}^d(\tau) &\geq f - M(1 - \phi_{xy}^d(\tau)) \\
\gamma_{xy}^d(\tau) &\geq 0.
\end{align*}$$

We complete the user-optimal MILP $(UE_1)$ with the new objective function

$$\min \sum_{(x,y) \in A} \sum_{d \neq x} \sum_{\tau=0}^{h-1} \gamma_{xy}^d(\tau) \hfill (18)$$

which at optimality is clearly equivalent to (17).

The great disadvantage of this formulation is that we have added 0-1 variables $\phi$ superscripted by $d$, so that for every 0-1 $\delta$ there are $D$ 0-1 variables $\phi$, where $D$ is the number of destinations.

### 4.3.3 Approximating MILP formulations

We close by presenting two MILP formulations that approximate $(UE_1)$ without explicitly including the extra 0-1 variables $\phi$. Both reformulations can be shown to be exact under separate special conditions. The technique is a variation on that of Oral and Kettani [9], and the proofs of theorems 3 and 4 are based on the proofs in that paper.

We require an a priori upper bound $f_{xy}^+$ on the flows $f_{xy}^d(\cdot, \cdot)$. We do not use our usual big-$M$ notation here because the error of our approximations will be shown to increase with $f_{xy}^+$. We form the new program $(UE_2)$ from $(UE_1)$ by adding continuous variables $\zeta_{xy}^d(\tau)$ for all $(x, y) \in A$, $d \neq x$ and $0 \leq \tau < h$, each constrained by

$$\begin{align*}
\zeta_{xy}^d(\tau) &\geq -f_{xy}^d(\tau, \tau + S) + f_{xy}^+(1 - \eta_{xy}^d(\tau)) \\
\zeta_{xy}^d(\tau) &\geq 0
\end{align*}$$

and writing the new objective function

$$\min \sum_{(x,y) \in A} \sum_{d \neq x} \sum_{\tau=0}^{h-1} f_{xy}^d(\tau, \tau + S) - f_{xy}^+(1 - \eta_{xy}^d(\tau)) + \zeta_{xy}^d(\tau). \hfill (19)$$
Theorem 3 The optimal solution to \((UE_2)\) also minimizes the objective function

\[
\sum_{(x,y)} \sum_d \sum_{\tau} \left[ f^d_{xy}(\tau, \tau + S) \phi^d_{xy}(\tau) + f^+_x(1 - \eta^d_{xy}(\tau)) \right]
\]

subject to the same constraints.

Proof: Since \(\zeta\) appears in the minimization objective, at optimality \((UE_2)\) satisfies

\[
\zeta^d_{xy}(\tau) = \max \{0, -f^d_{xy}(\tau, \tau + S) + f^+_x(1 - \eta^d_{xy}(\tau))\} \quad \forall (x,y), d, \tau.
\]

If \(\eta^d_{xy}(\tau) = 0\), then \(\zeta^d_{xy}(\tau) = -f^d_{xy}(\tau, \tau + S) + f^+_x(1 - \eta^d_{xy}(\tau))\). Thus, the objective function contribution for \(((x,y), d, \tau)\) is

\[
0 = f^d_{xy}(\tau, \tau + S) \phi^d_{xy}(\tau) + f^+_x(1 - \eta^d_{xy}(\tau)) - \phi^d_{xy}(\tau).
\]

If \(\eta^d_{xy}(\tau) \geq 1\), then \(\zeta^d_{xy}(\tau) = 0\) and \(\phi^d_{xy}(\tau) = 1\). The objective function contribution for

\[
((x,y), d, \tau)\]

is

\[
f^d_{xy}(\tau, \tau + S) - f^+_x(1 - \eta^d_{xy}(\tau)) = f^d_{xy}(\tau, \tau + S) \phi^d_{xy}(\tau) + f^+_x(1 - \eta^d_{xy}(\tau)) - \phi^d_{xy}(\tau).
\]

The objective function term for \(((x,y), d, \tau)\) in \((UE_2)\) is, thus, the same as that for \((UE_1)\) if \(\eta^d_{xy}(\tau) = 0\) or 1. But if \(\eta^d_{xy}(\tau) \geq 2\) the \((UE_2)\) term is \(f^+_x(\eta^d_{xy}(\tau) - 1)\) too large. A linear approximate correction to \((UE_2)\) is then to subtract \(f^+_x(\eta^d_{xy}(\tau) - 1)\) for all values of \(\eta^d_{xy}(\tau)\).

We call the resulting program \((UE_3)\); it is identical to \((UE_2)\) except in its objective function

\[
\sum_{(x,y) \in A} \sum_{d \neq x} \sum_{\tau = 0}^{h-1} f^d_{xy}(\tau, \tau + S) + \zeta^d_{xy}(\tau).
\]

(20)

Theorem 4 The optimal solution to \((UE_3)\) also minimizes the objective function

\[
\sum_{(x,y)} \sum_d \sum_{\tau} \left[ f^d_{xy}(\tau, \tau + S) \phi^d_{xy}(\tau) + f^+_x(1 - \phi^d_{xy}(\tau)) \right]
\]

subject to the same constraints.

Proof: Again, at optimality in \((UE_3)\),

\[
\zeta^d_{xy}(\tau) = \max \{0, -f^d_{xy}(\tau, \tau + S) + f^+_x(1 - \eta^d_{xy}(\tau))\} \quad \forall (x,y), d, \tau.
\]

If \(\eta^d_{xy}(\tau) \geq 1\), then \(\zeta^d_{xy}(\tau) = 0\) and \(\phi^d_{xy}(\tau) = 1\), hence \(((x,y), d, \tau)\) contributes

\[
f^d_{xy}(\tau, \tau + S) \phi^d_{xy}(\tau) + f^+_x(1 - \phi^d_{xy}(\tau))
\]

to the objective function.

If \(\eta^d_{xy}(\tau) = 0\), then \(\zeta^d_{xy}(\tau) = -f^d_{xy}(\tau, \tau + S) + f^+_x\) and \(\phi^d_{xy}(\tau) = 0\), hence \(((x,y), d, \tau)\) contributes

\[
f^+_x = f^d_{xy}(\tau, \tau + S) \phi^d_{xy}(\tau) + f^+_x(1 - \phi^d_{xy}(\tau))
\]

to the objective function. ■

16
The objective function of \((UE_3)\) exceeds that of \((UE_1)\) by

\[
\sum_{(x,y)} \sum_d \sum_{\tau} f^+_{xy} \cdot \text{[# of optimal next links from \((x,\tau)\) to \(d\)].}
\]

Although, in general, both reformulations have objective functions that only approximate \((UE_1)\), we can provide conditions for each to be exact. As we observed above, the error that separates \((UE_1)\) from \((UE_2)\) has the effect of adding a penalty, for each destination \(d\) at each time \(\tau\), for each link \((x,y)\) whose optimal trip to \(d\) is at least 2 periods longer than the fastest trip. The penalty, of magnitude \(f^+_{xy}(\eta^d_{xy}(\tau) - 1)\), has the effect of favoring solutions that sacrifice equilibrium flow in order to reduce the difference among the optimal trip times in the forward star from each space-time node \((x,\tau)\) to destination \(d\). In networks that strongly favor multipath routing, however, this effect is less severe.

**Theorem 5** The optimal solution to \((UE_2)\) is an equilibrium flow pattern such that from any \((x,\tau)\), the optimal trip times to any particular \(d\) associated with each possible first decision \((x,y)\) differ from each other by at most one, if any such flow pattern exists.

**Proof:** All flow patterns have nonnegative value in \((UE_2)\), and a flow pattern of the kind described in the statement of the theorem has value zero. ■

On the other hand, \((UE_3)\) would appear to be preferable in networks that strongly favor single-path routing.

**Theorem 6** The optimal solution to \((UE_3)\) is an equilibrium flow pattern such that from any \((x,\tau)\), there is only one optimal next link to any particular destination \(d\), if any such flow pattern exists.

**Proof:** From any \((x,\tau)\) to any \(d\) there is at least one optimal next link \((x,y)\) such that \(\phi^d_{xy}(\tau) = 0\). Therefore the value of any flow in \((UE_3)\) is at least \(\sum_x \sum_d \sum_{\tau} f^+_{xy}\). This is exactly the value of a flow pattern as described in the statement of the theorem. ■
References


