Transportation Research Part B 42 (2008) 771–781

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

An analytical approximation for the macroscopic fundamental diagram of urban traffic

Carlos F. Daganzo [a,*], Nikolas Geroliminis [b]

[a] Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94720, USA
[b] Department of Civil Engineering, University of Minnesota, USA

ARTICLE INFO

Article history:
Received 15 April 2008
Received in revised form 24 June 2008
Accepted 25 June 2008

Keywords:
Macroscopic fundamental diagram
Variational theory
Urban congestion

ABSTRACT

This paper shows that a macroscopic fundamental diagram (MFD) relating average flow and average density must exist on any street with blocks of diverse widths and lengths, but no turns, even if all or some of the intersections are controlled by arbitrarily timed traffic signals. The timing patterns are assumed to be fixed in time. Exact analytical expressions in terms of a shortest path recipe are given, both, for the street's capacity and its MFD. Approximate formulas that require little data are also given.

For networks, the paper derives an upper bound for average flow conditional on average density, and then suggests conditions under which the bound should be tight; i.e., under which the bound is an approximate MFD. The MFD's produced with this method for the central business districts of San Francisco (California) and Yokohama (Japan) are compared with those obtained experimentally in earlier publications.

1. Introduction

It has recently been proposed (Daganzo, 2005a, 2007) that traffic can be modeled dynamically in large urban regions (neighborhoods) at an aggregate level if such regions exhibit two features: a reproducible "macroscopic fundamental diagram" (MFD) relating the number of circulating vehicles (or accumulation) to the neighborhood's average speed (or flow), and a robust relation between the neighborhood's average flow and its total outflow. In this theory the MFD should have a well-defined maximum and remain invariant when the demand changes both with the time-of-day and across days.

The idea of an MFD with an optimum accumulation is quite old (Godfrey, 1969) but the verification of its existence with these dynamic features is recent (Geroliminis and Daganzo, 2007, 2008). Earlier works had looked for MFD patterns in data from lightly congested real-world networks (Godfrey, 1969; Ardekani and Herman, 1987; Olszewski et al., 1995) or in data from simulations with artificial routing rules and static demands (e.g., Williams et al., 1987, Mahmassani et al., 1987, Mahmassani and Peeta, 1993), but the data from all these studies were too sparse and/or contrived to demonstrate that an invariant MFD could dynamically arise in the real world. Although Geroliminis and Daganzo (2007, 2008) have shown that this can indeed happen, more real-world experiments should shed light on the types of networks and demand conditions for which invariant MFD's arise.

Invariance is important because knowledge of a neighborhood's MFD allows decision-makers to use verifiable and robust demand-side policies to improve mobility such as those in Daganzo (2005a, 2007). However, to evaluate changes to the network (e.g., re-timing the traffic signals or changing the percentage of streets devoted to public service vehicles) one needs to know how the MFD is affected by the changes. To begin filling this gap, this paper explores the connection between network structure and a network's MFD for urban neighborhoods controlled at least in part by traffic signals.

* Corresponding author. Tel.: +1 510 642 3853; fax: +1 510 642 1246.
E-mail addresses: daganzo@ce.berkeley.edu (C.F. Daganzo), nikolas@umn.edu (N. Geroliminis).

0191-2615/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.trb.2008.06.008

The task is challenging because the ultimate goal is a universal recipe that can be used for all signal-controlled networks. However, recognizing that networks are complex structures described by many variables, we shall be satisfied with an approximation that uses as few of these variables as possible. The most similar work to what we propose is a simulation study (Gartner and Wagner, 2004) which, in the spirit of earlier works that examined the relationship between flow and density on rings (e.g., Wardrop, 1965; Franklin, 1967), explores how this relationship is affected by placing traffic signals on the ring. Gartner and Wagner (2004) simulated the ring for a limited range of signal timings and unveiled several regularities. These regularities, however, cannot be extrapolated to form a general theory because simulation only speaks to the range of simulated parameters.[1] In view of this, and given the many parameters required to describe a neighborhood, we shall take an analysis approach.

The paper is organized as follows: Section 2 first proves the existence and uniqueness of an exact, concave MFD for any multi-block, signal-controlled street without turning movements using the tenets of variational theory (VT) (Daganzo, 2005b,c). This section also gives exact and approximate recipes for both, the street's capacity and its MFD. Section 3 then introduces a bound for general networks and also explains how the MFD formulae of Section 2 can be scaled up for certain networks. Finally, Section 4 compares the MFDs obtained with this method for both San Francisco (CA, USA) and Yokohama (Japan) with the MFDs observed in Geroliminis and Daganzo (2007, 2008).

2. A single street with no turns

Considered here is a street of length L with a fixed number of lanes but any number of intersections. The intersections can be controlled by stop lines, roundabouts, traffic signals or any type of control that is time-independent on a coarse scale of observation; i.e., large compared with the signal cycles. We are interested in solutions where the flow at the downstream end of the street matches the flow at the upstream end; e.g., as if the street formed a ring, because then the average density does not change. To this end we will consider an initial value problem (IVP) on an extended version of our street, obtained as in Fig. 1 by placing end-to-end an infinite number of copies of the original street. The problem will be treated with VT. We first show how to evaluate the capacity of this system; then turn our attention to the MFD.

2.1. Street capacity

The centerpiece of variational theory is a relative capacity ("cost") function (CF), $r(u)$, that describes each homogeneous portion of the street. This function is related to the FD of kinematic wave theory Q by the complex conjugation operation: $r(u) = \sup\{Q(k) - ku | \forall k \in [0, \kappa]\}$, where k is the density and κ the jam density. Physically, the CF gives the maximum rate at which vehicles can pass an observer moving with speed u; i.e., the street's capacity from the observer's frame of reference. We assume in this paper that the CF is linear, as shown in Fig. 2, and characterized by the following parameters: k_0 (optimal density), u_f (free flow speed), κ (jam density), w (backward wave speed), q_m (capacity), and r (maximum passing rate). Linear CFs correspond to triangular FDs.

In VT, the street can also have any number of time-invariant and/or time-dependent point bottlenecks with known capacities; e.g., at intersections controlled by traffic signals. The bottlenecks are modeled as lines in the t, x plane on which the "cost" per unit time equals the bottleneck capacity, $q_B(t)$. As an illustration, hypothetical red periods of signal-controlled intersections are indicated by short lines in the "original" swath of Fig. 1. These lines, replicated in all the copies, would have zero cost. During the green periods (G) the bottleneck capacity is the saturation flow of the intersection s, i.e., $q_B = s$, which could be equal or less than q_m.

A second element of VT is the set of "valid" observer paths on the (t, x) plane starting from arbitrary points on the boundary at $t = 0$ and ending at a later time, $t_0 > 0$. A path is "valid" if the observer's average speed in any time interval is in the range $[-w, u_f]$. Let $\mathscr{P}$ be one such path, $u_{\mathscr{P}}$ be the average speed for the complete path, and $\Delta(\mathscr{P})$ the path's cost. This cost is evaluated with $r(u)$, treating any overlapping portions of the path with the intersection lines as shortcuts with cost $q_B(t)$. (Of course, $q_B = 0$ during the red periods.) By definition, $\Delta(\mathscr{P})$ bounds from above the change in vehicle number that could possibly be seen by observer $\mathscr{P}$. Thus, the quantity

$$R(u) = \lim_{t_0 \to \infty} \inf_{\mathscr{P}} \{\Delta(\mathscr{P}) : u_{\mathscr{P}} = u\}/t_0 \tag{1}$$

is an upper bound to the average rate at which traffic can overtake any observer that travels with average speed u for a long time. Note that (1) is a shortest path problem, and that $R(0)$ is the system capacity. Thus, the problem of evaluating the capacity of long heterogeneous streets with short blocks and arbitrary signal timings turns out to be conceptually quite simple.

It is also practically simple. It has been shown (Daganzo, 2005c) that for linear CF's an optimal path always exists that is piece-wise linear: either following an intersection line or else slanting up or down with slope u_f or $-w$. This is illustrated by Fig. 3a, which depicts block i of our street (with length, l_i). In this figure arrows denote the possible directions of an optimal path, with associated costs shown in parentheses. Consideration shows that if all the blocks of our street are sufficiently long

[1] An example of a regularity in Gartner and Wagner (2004) that cannot be extrapolated is the reported independence between system capacity and signal offsets because, as is well known, offsets do affect capacity when intersections are closely spaced.

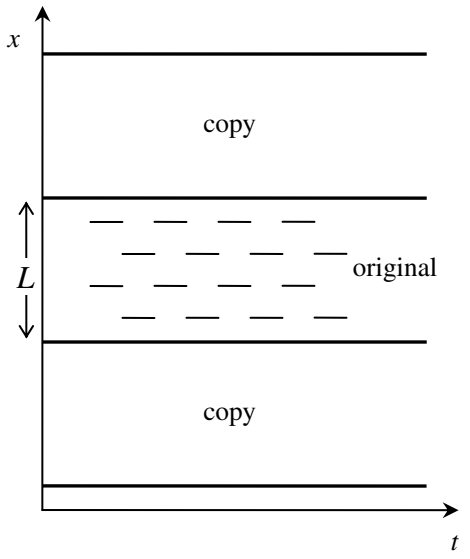

Fig. 1. The periodic IVP for a single street of length L: short segments are red phases at intersections.

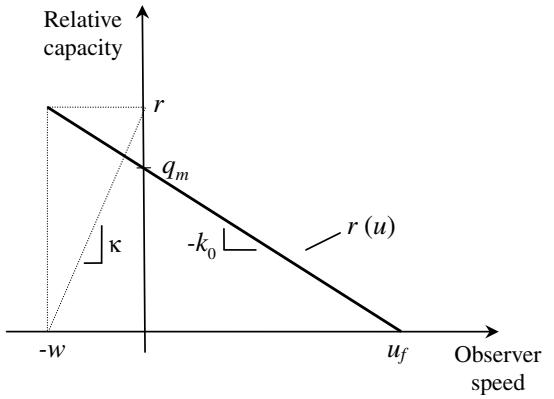

Fig. 2. The linear cost function.

(such as the one in Fig. 3c) then the shortest path (SP) is a horizontal line along the trajectory of one of the intersections; and the capacity is simply: $R(0) = \min_i\{s_i G_i/C_i\}$, where G_i is the effective green time and C_i the cycle time. However, if some of the blocks are short then there could be shortcuts that use red periods at more than one intersection, as shown in Fig. 3b. In this case the capacity is smaller.

Example. As an illustration, we evaluate the capacity, c, of a homogeneous ring road with two diametrically opposed and identically timed signals. Let $2l$ be the length of the road, and assume $s = q_m$. In order to obtain a complete solution with as few degrees of freedom as possible without losing generality we measure time at every location starting with the passage of a free flowing vehicle (so we can assume $u_f = \infty$) and then choose the units of time, distance and vehicular quantity so that $C = 1$, $s = 1$ and $w = \kappa = 1$. Now, the solution only depends on three parameters (G, l, δ); but we shall only consider the two symmetric cases where the offsets are the same for both signals: $\delta = 0$ and $\delta = C/2$.

The reader can verify using the shortest path method described above that the complete solution to this problem is as displayed in Fig. 4. This solution matches the known capacity formulae for pairs of intersections. Note that offsets affect capacity considerably when blocks are short: $l < G$. Appendix A gives capacity formulae for a few additional cases.

2.2. The street's MFD

Consider now an IVP with a periodic initial density, with average k. This problem is known to have a unique solution with meaningful densities everywhere (Daganzo, 2006) and, since all its input data are periodic in space, this solution must be periodic – with period L. Thus, our original street has the same inflows and outflows: it behaves as a ring, as desired.

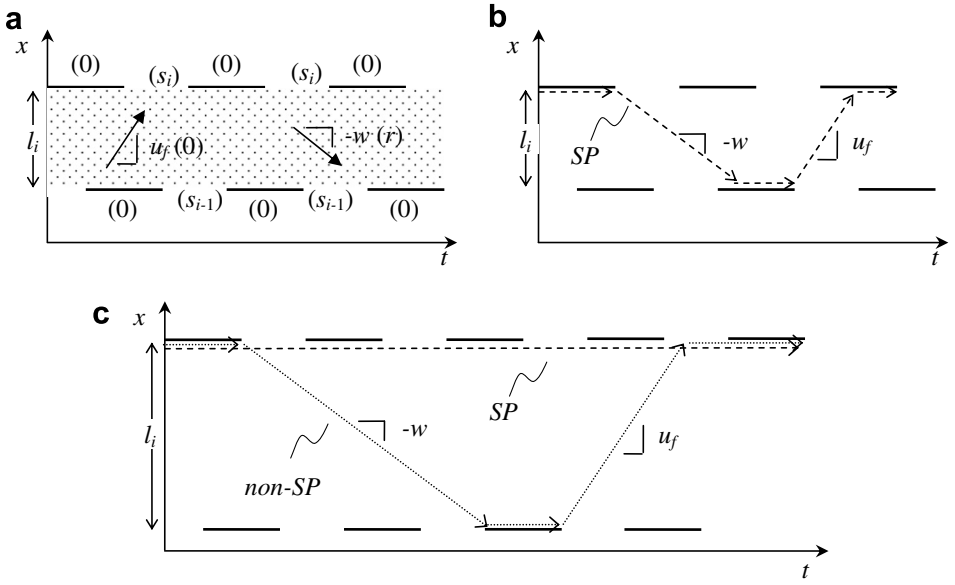

Fig. 3. Estimation of capacity according to VT: (a) costs; (b) short block; and (c) long block.

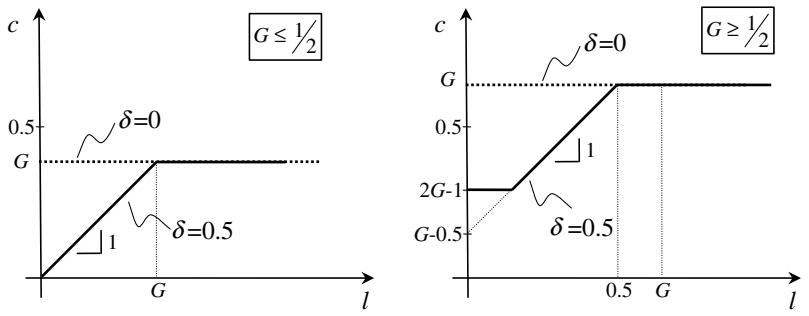

Fig. 4. Capacity of a symmetric ring with two signals.

Consider next the average flow from $t = 0$ to $t = t_0$ at some location (say $x = 0$), and denote it by $q(t_0)$. Because our IVP is periodic so that vehicles are conserved, $q(t_0)$ approaches a location-independent limit, q, as $t_0 \to \infty$. This limit will, of course, depend on the initial density distribution. We now show that q is connected with the initial density distribution only through its average; i.e., that an MFD function Q, $q = Q(k)$, exists. We also show that Q is concave.

Proposition. *A ring's MFD, $q = Q(k)$, is concave and given by:*

$$q = \inf_u \{ku + R(u)\}. \tag{2}$$

Proof. Recall from VT that the vehicle number at a point is the greatest lower bound of the numbers that could have been computed by all valid observers, $\mathscr{P}$, by adding each observer's $\Delta(\mathscr{P})$ to its given initial number (at the boundary). We now evaluate with this recipe the vehicle number, n_0, observed when $t = t_0 \to \infty$ at the location where the initial vehicle number is 0. We do this by considering observers ending their trips at the location in question but traveling with different long term average speeds u (and of course emanating from different points on the boundary). By using (1) and noting that the initial vehicle number for an observer with average speed u is in the range $kut_0 \pm kL$ we find that $n_0 = \inf_u\{kut_0 \pm kL + R(u)t_0\}$, where $t_0 \to \infty$. Thus, on dividing both sides by $t_0 \to \infty$ we obtain (2). To conclude the proof we need to show that (2) is concave. But this is clear because (2) is the lower envelope of a set of straight lines, which is always a concave curve. □

The term $R(u)$ can be obtained with the SP recipe of Section 2.1. Fig. 5 illustrates that (2) is the lower envelope of the one-parameter family of lines on the (k, q) plane defined by $q = ku + R(u)$ with u as the parameter. We call these lines "cuts" because they individually impose constraints of the form: $q \leqslant ku + R(u)$ on the macroscopic flow-density pairs that are feasible on our street. This inequality should be intuitive because it applies to a scale of observation so large that our street appears to

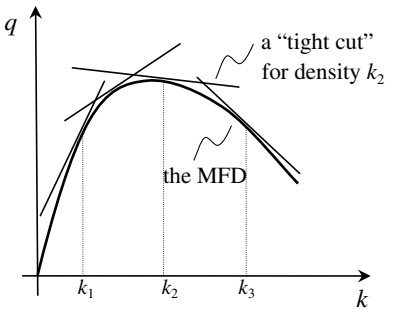

Fig. 5. The MFD defined by a 1-parameter family of "cuts".

be on a steady state (k, q) – as a result, an observer traveling at speed u must be passed at a sustained rate q_r such that $q = ku + q_r$, where $q_r \leqslant R(u)$ (as we showed in Section 2.1). Less obvious is that according to our proposition there always is a "tight" cut that yields the average flow for any given density, such as those shown for k_1, k_2 and k_3 in the figure.

2.3. Practical approximations

Because evaluating $R(u)$ in (2) for all u can be tedious, we propose instead using three families of "practical cuts" that jointly bound the MFD from above, albeit not tightly. The approximate MFD is denoted by T instead of Q. Note T is concave, and $T \geqslant Q$. Our practical cuts are based on observers that can move with only three speeds: $u = u_f$, 0, or $-w$. Recall that an observer's cost rate is $q_B(t)$ if the observer is standing at intersection with capacity $q_B(t) \leqslant q_m$ and otherwise it is as given by Fig. 2; i.e., it is either 0, s or r.

Family 1: The first family uses stationary observers at different locations, and out of these, we choose the one standing at the most constraining intersection. This leads to the first cut

$$q \leqslant q_B = \min_i \{s_i G_i / C_i\}, \tag{3}$$

where q_B is the average capacity of the most constraining intersection.

Family 2: Now consider observers that move forward at speed u_f, except where delayed by a red phase at an intersection. In order to control the average speed of these observers in a way that allows us to evaluate the maximum passing rate, we assume that all the red phases R_i have been artificially extended at the front end by an amount εG_i, where $\varepsilon \in [0,1]$ is a parameter. (We also assume that the delayed observer always departs the intersection at the end of the red, even when $\varepsilon = 1$.) Let $u(\varepsilon)$ be the average speed of this observer and $f_i(\varepsilon)$ the fraction of time that it spends artificially stopped in green phases of intersection i (and its copies) because of extended reds. This observer can be passed at most at rate s_i during $f_i(\varepsilon)$, and not at all at other times. Thus, traffic can pass it on average at a rate $q_r \leqslant \sum_i s_i f_i(\varepsilon)$ on average, and the moving observer formula yields our second family of cuts:

$$q \leqslant ku(\varepsilon) + \sum_i s_i f_i(\varepsilon) \quad \text{for } 0 \leqslant \varepsilon \leqslant 1. \tag{4a}$$

If the street has uniform width, with the same q_m on all its blocks ($q_m \geqslant s_i$), one may use the rougher cut

$$q \leqslant ku(\varepsilon) + q_m f(\varepsilon) \quad \text{for } 0 \leqslant \varepsilon \leqslant 1, \tag{4b}$$

where $f(\varepsilon) = \sum_i f_i(\varepsilon)$ is the fraction of time that the vehicle is stopped on extended red phases.

Family 3: The third and last family is the mirror image of the second, with the observer traveling in the opposite direction, at speed w instead of u_f, and also stopping for the red phases. Now we use $w(\varepsilon) > 0$ for the average speed of the observer, $b_i(\varepsilon)$ for the fraction of time it spends in extended red phases of intersection i and $h_i(\varepsilon)$ for the fraction of time it spends moving toward i. This observer can be passed at most at rate r_i when moving. Therefore, it can be passed in total at most at an average rate $\sum_i [s_i b_i(\varepsilon) + r_i h_i(\varepsilon)]$, so that the resulting set of cuts arising from the moving observer formula is

$$q \leqslant -kw(\varepsilon) + \sum_i [s_i b_i(\varepsilon) + r_i h_i(\varepsilon)] \quad \text{for } 0 \leqslant \varepsilon \leqslant 1. \tag{5a}$$

Again, if the street is uniform in width, with the same r and the same $q_m (q_m \geqslant s_i)$ on all its blocks, one may prefer to use the rougher cut

$$q \leqslant -kw(\varepsilon) + q_m b(\varepsilon) + r\frac{w(\varepsilon)}{w} \quad \text{for } 0 \leqslant \varepsilon \leqslant 1, \tag{5b}$$

where $b(\varepsilon) = \sum_i b_i(\varepsilon)$ is the fraction of time that the observer is stopped on extended red phases, and $w(\varepsilon)/w = \sum_i h_i(\varepsilon)$ is the fraction of time that the observer is moving.

Eqs. (4b) and (5b) can be further simplified for any homogeneous street (i.e., with uniform width, block lengths and signal settings) because in this case the observers follow simple periodic paths with one stop per period. These paths only differ in

the number of blocks, $\gamma = 1,2,\ldots,\gamma_{max}$, traversed per stop, where γ_{max} may be infinite. Thus, for homogeneous streets γ can be used as a discrete parameter instead of ε to express all the cuts in (4b) and (5b) in terms of constants: $u_\gamma, f_\gamma, b_\gamma$ and w_γ. Appendix B expresses these constants in terms of l, G, C, and the offset δ.

How good are these simplifications? The reader can verify without too much effort that for the symmetric ring of Section 2.1 the five simple cuts given by (3) and the two extreme cases of (4b) and (5b) (with $\gamma = 1$ and $\gamma = \gamma_{max}$) define an approximate MFD, T, with a capacity that matches exactly the one predicted in Section 2.1. Furthermore, it is possible to show that these five simple cuts always predict exactly the capacity of a homogeneous street with two signals.[2] Therefore, we conjecture that (3–5) should be good approximations in general. These cuts will also be used with networks.

3. Application to urban areas: networks

Three new issues arise with networks. First, unlike our ring, real streets never contain a perfectly invariant number of vehicles because even in a steady state vehicles can both, randomly turn at intersections and either begin or finish their trips along the street itself. Second, these turns and trip ends violate the tenets of VT. And third, route choice may come into play. We address the last two issues first because we can do so deterministically, and then introduce a stochastic refinement to address the first.

3.1. Turns, trip ends and route choice

Here, we first derive an upper bound for the average network flow q conditional on the average network density k and then suggest some conditions under which the bound should be tight. These averages are defined in the sense of Edie (1963), as proposed in Daganzo (2005a, 2007): $q = \sum_i q_i l_i/D$ and $k = \sum_i k_i l_i/D$, where l_i is the length of link i and D the total network length.

Although VT does not hold on links with turns and trip ends, the observer method of Section 3 still yields cuts that bound from above the possible flow-density states on any such link; thus, we can find functions T_i that ensure: $T_i(k_i) \geqslant q_i$. These bounds can then be composed to form a bound for the average network flow: $\sum_i T_i(k_i) l_i/D \geqslant q$. A network level bound for flow conditional on density, $T(k)$, can now be obtained by maximizing the left hand side of this inequality with respect to the k_i subject to a network density constraint; i.e., we have

$$T(k) = \max\left\{ \sum_i T_i(k_i) l_i \Big/ D \,\middle|\, \sum_i k_i l_i \Big/ D = k;\ \ 0 \leq k_i \leq \kappa_i \right\} \geq q. \tag{6}$$

We see by symmetry that if all links are similar the optimum is achieved for $k_i \cong k$.

If an optimum distribution of densities were to arise naturally in some real world networks, and if the T_i bounds were tight, then the LHS of (6) would be a good approximation for q – not just a bound – and an approximate MFD, $q \cong Q(k) \cong T(k)$, could be defined. We conjecture that the following regularity conditions should ensure this outcome: (i) a slow-varying and distributed demand; (ii) a redundant network (e.g., an irregular grid made up of blocks) ensuring that drivers have many route choices and that most links are on many desirable routes; (iii) a homogeneous network with similar links; (iv) links with an approximate FD that is not significantly affected by turning movements when flow is steady.[3]

Conditions (i)–(iii) should create a near-equilibrium as in Wardrop (1952) with similar average speeds on all links; and, since the links are similar, with similar (i.e., optimal) densities too. Condition (iv) implies that the VT method, applied to a single link with many efficient cuts, yields a tight T_i. Thus we conclude that

$$q \cong Q(k) \cong T(k) \cong T_i(k_i), \quad \text{if conditions (i)—(iv) are satisfied.} \tag{7}$$

This result is useful in practice because T_i can be derived easily. Section 4 will put it to the test.

3.2. Stochastic fluctuations

Here we propose a second order approximation to roughly capture the stochastic effects induced by both, turns and trip ends. Experience with simulations and real-life shows that random variations in trip-making can create spatial pockets where the average speed and accumulation are temporarily different from the prevailing average. These localized differences should be temporary in neighborhoods with constant demand due to the effects of route choice. But, despite the stabilizing effect of route choice, both speed and density must be distributed over space at any given time with some dispersion – even if their long term averages are the same everywhere. We now examine how this spatial dispersion in density affects the long term average flow.

Since traffic is granular and random (even in the steady state) the vehicular input and output to any given street or link behaves as a superposition of binomial processes, so that the number of vehicles in it fluctuates from the average as a

[2] The reason is geometric. Consideration shows that for $t_0 \to \infty$ a least cost path with zero average speed (which defines the capacity of our system) can always be constructed by splicing together a subset of our five elementary paths.

[3] This condition should apply if links are sufficiently long because turns only affect the relationship between flow and density locally – mostly at links' ends – and not so much in their middles. It should also apply if one knows a priory that the fraction of turns is small.

random walk. We are interested in the distribution of these fluctuations over space, conditional on the total number of vehicles in the network, n. If the stabilization effects of route choice allow significant excursions from the average on individual links but are strong enough to preclude them in large multi-link regions (which seems reasonable), the distribution of vehicles across links may look as if each vehicle randomly chose without replacement an available position on the network where the number of positions on link i is $N_i \equiv \kappa_i l_i$ and on the network, $N = \sum_i N_i$. In other words, we model the number of vehicles on a link n_i as the outcome of a hypergeometric random variable consisting of n draws from an urn containing N_i good balls and $N - N_i$ bad balls. Thus

$$E(n_i) = nN_i/N \quad \text{and} \quad \text{var}(n_i) \cong nN_i(N-n)/N^2. \tag{8}$$

The formula for the variance is a simplification that holds for $N \gg N_i$. Since $N_i \sim 10^1$ to 10^2 for typical links, the mean of n_i exceeds its standard deviation by a factor of 2 or greater when the network saturation level n/N is close to capacity ($n/N \sim 0.3$). Under these conditions, n_i is approximately normal, and we shall use this approximation in our calculations.

If the local fluctuations in density persist for times substantially longer than a signal cycle they should affect link flows so that $q_i \cong Q_i(n_i/l_i)$; and if the network is homogeneous so that $Q_i \cong Q$, the average network flow becomes $q \cong E[Q(n_i/l)]$ and we have

$$q \cong E[Q(n_i/l)] \cong E[T(n_i/l)]. \tag{9}$$

Note that (7) predicts $q \cong Q(k) = Q[E(n_i/l)]$, since $E[n_i/l] = n/D = k$; and that $E[Q(n_i/l)] \leqslant Q[E(n_i/l)]$ because Q is concave. Thus, (9) predicts lower values than (7): granularity reduces network flow.

4. Applications

4.1. The study sites

We now use (7)–(9) to estimate an MFD for two study sites that only roughly meet the regularity conditions of Section 3.1. The first site is simulated and the second real. The first site provides a controlled test that isolates the errors of the proposed approximation. The second site illustrates how the method may work in a real-world application where the input data include some error. For more information about the study sites and the experiments see Geroliminis and Daganzo (2007, 2008).

The first test site is a 5 km^2 area of Downtown San Francisco (Financial District and South of Market Area), including about 100 intersections with link lengths varying from 100 to 400 m. Traffic signals are pre-timed with a common cycle. Network geometry and traffic flow data were available from previous studies.

The second site is a 10 km^2 part of downtown Yokohama. It includes streets with various numbers of lanes and closely spaced signalized intersections (100–300 m). Major intersections are centrally controlled by actuated traffic signals that effectively become pre-timed (with a common cycle) during the rush.

4.2. Results

Although both sites are somewhat heterogeneous we treat them as if they could be decomposed into sets of homogeneous 1-lane links, similar within each city; e.g., by visualizing multi-lane links as side-by-side juxtapositions of 1-lane links. Therefore we use a single typical 1-lane link to define $T_i(k_i) \cong T(k)$ with the formulae of Appendix B, and then use (9). This is a very rough approximation, but it is simple. Only the following information, which is summarized in Table 1, is needed: (i) network variables, D (network length in lane-km) and l (average link length); (ii) link variables (for 1-lane), u_f, κ, $s = q_m$ and $w = u_f/(\kappa u_f/q - 1)$; and (iii) intersection variables, δ, C and G. Table 1 also includes the two derived variables ($u_{\gamma_{max}}$ and γ_{max}) from Appendix B.

Table 1
Estimated parameters for the two sites

	Site 1 (SF)	Site 2 (Y)
D (km)	76.2	157.0
l (m)	122.9	154.0
u_f (m/s)	13.4	13.9
κ (vh/m)	0.13	0.14
$s = q_m$ (vh/s)	0.5	0.5
w (m/s)	5.4	5.0
δ (s)	2.6	0 (peak)
G (s)	21	49
C (s)	60	130
$u_{\gamma_{max}}$ (m/s)	7.0	8.4 (off-peak)
γ_{max}	4	8 (off-peak)

The SF parameters were chosen to match those of the simulations. For Yokohama, real-world data were used but detailed signal setting information was not available. Parameters D and l were estimated from road maps; C, κ and q_m were reported by local experts (Kuwahara, 2007); u_f was available from vehicle GPS data; and the ratio G/C (and therefore G) from the relation $G/C = q_B/q_m$, since q_B was available from detector data. Since the Yokohama site uses traffic responsive control during the off-peak periods when traffic is light, but reverts to fixed cycles when traffic is heavy, δ was not clearly defined. Thus, we used a value consistent with the $u_{\gamma_{max}}$ and γ_{max} measured from taxi GPS data when traffic is light, and assumed that signals operate with zero offsets the rest of the time.

With these data, MFDs were constructed for both cities using the three types of cuts of Section 2.3 for all $\gamma = 1,2,\ldots,\gamma_{max}$. The piece-wise linear curves of Fig. 6 show the result: only non-redundant cuts are shown. The two entries in each box are the value of γ and the observer type (F for forward, B for backward, S for stationary). The smooth grey curve is the granular approximation (9).

Fig. 7a and b compares the speed-based MFDs obtained from the granular approximations in Fig. 6a and b with those empirically observed in Geroliminis and Daganzo (2007, 2008). For the SF site of Fig. 7a, each point represents the city's average speed and accumulation every 5 min. Even though very different spatial and temporal demand patterns were simulated, the city-wide average speeds are consistent and closely predicted.

Fig. 7b includes more error but this should not be surprising because: (i) Yokohama used actuated signals with settings that varied with time; (ii) its network is less homogeneous; and (iii) our input data comes from field observation and expert opinion (not simulation) which may include significant error.

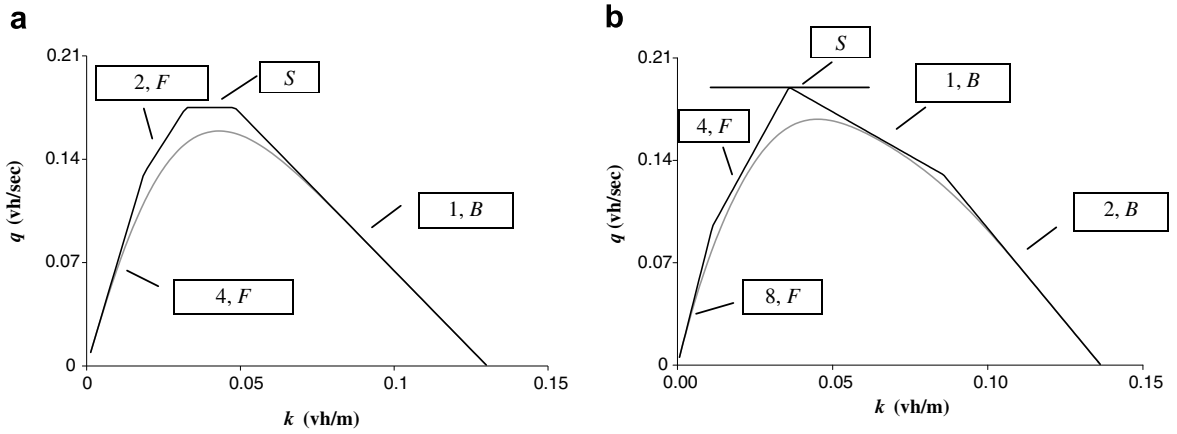

Fig. 6. Theoretical MFD with and without stochastic variations: (a) San Francisco and (b) Yokohama.

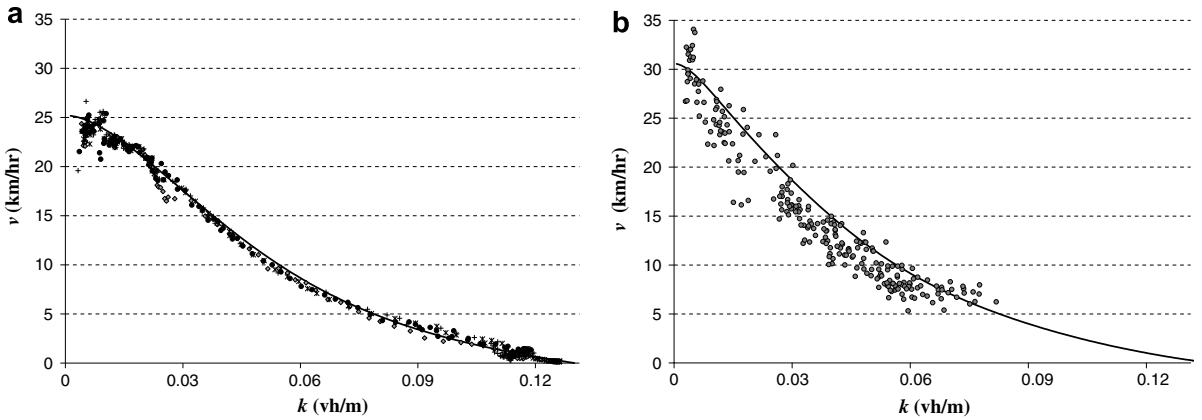

Fig. 7. Estimated MFD: (a) San Francisco and (b) Yokohama.

5. Discussion

Section 2 of this paper used a moving observer method to show that the average flow-density states of any urban street without significant turning movements must be bounded from above by a concave curve. The section also shows that, under the assumptions of VT, this curve is the locus of the possible (steady) traffic states for the street; i.e., it is its MFD.

Section 3 then showed that the moving observer bounds for the links of a network can always be composed into an upper bound curve for the average flow-density states of the entire network. The section also shows that the average flows on networks satisfying certain regularity conditions should always be close to their bounds, and therefore that these networks should have MFDs close to their bounds. This means that these networks automatically maximize their total flow for any given number of circulating vehicles. This is a good thing from the perspective of a traveler, since as explained in Daganzo (2005a, 2007), maximum flow correlates with maximum system outflow and minimum total delay. Approximate formulae for the MFD of these networks were also presented. Section 4 applied the formulae to the networks of downtown San Francisco and Yokohama with encouraging results.

It should be stressed that the regularity conditions of Section 3 are only sufficient and not necessary for the existence of an MFD. Networks that violate these conditions can also have an MFD, such as the inhomogeneous streets of Section 2, which violate two of the conditions. Thus, an area of research is further understanding when and when not and MFD should be expected. Of particular interest is whether MFDs should be expected for locally heterogeneous but macroscopically regular networks, such as those that accommodate multiple modes with a variety of street types. We suspect that the answer to the existence question is positive for networks that can be partitioned into identical components, say on the scale of several blocks, because the logic behind the regularity conditions of Section 3.1 continues to hold if the word "component" is simply substituted for the word "link" in the phrasing of conditions (ii), (iii) and (iv). Also of interest would be formulae for the MFDs of these networks.

We close with the reminder that MFDs should not be universally expected, despite our findings. In particular, networks with an uneven and inconsistent distribution of congestion may exhibit traffic states that are well below the bound and much too scattered to line along an MFD. An inconsistent distribution of congestion is typical of freeway networks with multiple recurrent and non-recurrent bottlenecks, and of large urban areas with multiple congested sub-centers. Even networks that satisfy our regularity conditions may exhibit significant scatter on their MFDs because of a rapidly changing demand since, as pointed out in Daganzo (1998), minor temporal variations in demand can drastically affect the performance of some congested networks.

Acknowledgements

The authors thank Prof. Masao Kuwahara (University of Tokyo) for his continued support with the Yokohama data, and for his many comments. We also gratefully acknowledge the comments of Profs. Toshio Yoshii (University of Kyoto), Hideyuki Kita (University of Kobe) and Yasuo Asakura (University of Kobe). PhD students V. Gayah and Y. Xuan also provided comments. This research was supported by the UC Berkeley Center of Excellence in Future Urban Transport.

Appendix A. Capacity formulae for some special cases

Here we give capacity formulas for some special cases where the calculations are simple.

A.1. Unsignalized intersections with four-way stops

They can be modeled as signals with very short cycles: letting G_j, $C_j \to 0$ while holding G_j/C_j constant and using a proper value for s_j. Then, $q_{B_j} = s_j \cdot G_j/C_j$.

A.2. Pairs of intersections

There are several cases with simple results.

Case 1: Neighboring unsignalized intersections ($C_j = C_{j-1} \to 0$). In this case shortcuts do not exist and $q_B = \min\{q_{B_j}, q_{B_{j-1}}\}$.
Case 2: Neighboring signalized and unsignalized intersections ($C_j = 0$ and $C_{j-1} > 0$ or vice versa). Assume that $s = q_m$ (zero turns). Let (C, G) be the timing parameters of the signalized intersection and $g = q_{B_j}/q_m$ the equivalent fraction of green for the unsignalized intersection. Then, if $g \geqslant G/C$ (the signal is more restrictive) we have:

$$\text{(Short block)} \quad \kappa l < q_m G : q_B = (l\kappa + g(Gq_m - l\kappa))/C, \tag{A1}$$

$$\text{(Long block)} \quad \kappa l \geqslant q_m G : q_B = q_m G/C. \tag{A2}$$

Case 3: Properly timed signals with a common cycle: If there is a common cycle an offset always exist that guarantees the same system capacity as if $l_j = \infty$, e.g., the offset $\delta = 0$ (This is a well known result and can be verified with VT). Thus, for properly timed signals: $q_B = \min\{q_{B_j}, q_{B_{j-1}}\}$.
Case 4: Improperly timed signals (different cycles): Also of interest is the case, where $C_j \approx C_{j-1} \approx C$ but $C_j \neq C_{j-1}$. In this case, the offsets vary approximately uniformly between 0 and C and we find:

$$q_B \approx \frac{l\kappa}{C} + \frac{1}{2}\left(\frac{Gq_m}{C} - \frac{l\kappa}{C}\right). \tag{A3}$$

In summary, for cases 2, 3 and 4 above, we have:

$$q_B = \frac{l\kappa}{C} + \alpha\left(\frac{Gq_m}{C} - \frac{l\kappa}{C}\right) \quad \text{if } \kappa l \geqslant Gq_m \text{ (short block)} \tag{A4}$$

$$= Gq_m/C \qquad \text{otherwise} \tag{A5}$$

where $\alpha = 1/2$ for improperly timed signals, $\alpha = g \geqslant G/C$ when one of the intersections is unsignalized (but not restrictive) and $\alpha = 1$ if signals have favorable offsets or the block is long. Cases not covered by Eqs. (A1)–(A5) can be evaluated with the VT recipe.

Appendix B. Cuts for homogeneous streets

We consider here a street with uniform block length l, and the same signal settings (offset δ, cycle C, and green phase G) at all its intersections; see Fig. B1.

Shown on the figure is the trajectory of a fast observer that travels at free flow speed, u_f, except when stopped by a red signal. Recall that γ_{max} is the number of links a fast observer traverses between consecutive stops ($\gamma_{max} = 3$ in the figure). We explain below that the delay at each stop to observers of this type is given by

$$d_{\gamma_{max}} = C \cdot \left(\left\lceil \frac{\gamma_{max}(l/u_f - \delta)}{C} \right\rceil - \frac{\gamma_{max}(l/u_f - \delta)}{C}\right); \tag{B1}$$

that the observer's average speed is given by

$$u_{\gamma_{max}} = \frac{\gamma_{max} l}{d_{\gamma_{max}} + \gamma_{max} l/u_f} \tag{B2}$$

and that γ_{max} is given by

$$\gamma_{max} = 1 + \max\{\gamma : \gamma(l/u_f - \delta)/C - \lfloor \gamma(l/u_f - \delta)/C \rfloor \leqslant G/C\}. \tag{B3}$$

B.1. Explanation

In (B1) and (B3), the quantities $\gamma(l/u_f - \delta)/C$ are the observer's arrival times to the γth intersection assuming that time is measured in units of cycles and that the clock of the γth intersection is started δ time units after the clock of the departure intersection. Thus, (B1) merely expresses that the stop at $\gamma = \gamma_{max}$ causes a delay equal to the time left in the arrival cycle. Similarly, the quantity on the LHS of the inequality in (B3) is the fraction of a cycle that has elapsed when the observer arrives at the γth intersection – and the inequality merely indicates that the γth intersection is bypassed if and only if the observer arrives during the intersection's green phase.

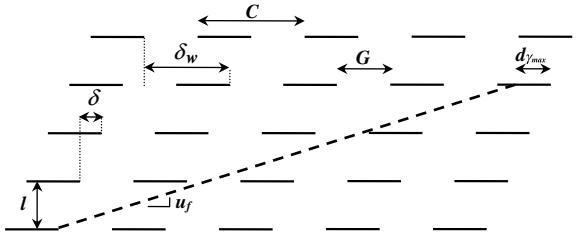

Fig. B1. A time–space diagram for a fast forward-moving observer.

B.2. Formulae for u_γ and f_γ

Consider now a slower observer who stops every γ signals ($\gamma = 1, 2, \ldots, \gamma_{max} - 1$) because of extended red phases, as described in Section 2.3. The speed u_γ and delay d_γ of this observer are still given by (B1) and (B2) after replacing γ_{max} by γ; and the fraction of time f_γ that the observer spends in extended red phases is

$$f_\gamma = \frac{d_\gamma - C + G}{\gamma l/u_f + d_\gamma} \quad \text{for } \gamma = 1, 2, \ldots, \gamma_{max} - 1. \tag{B4}$$

B.3. Formulae for b_γ and w_γ

Eqs. (B1)–(B4) can also be used to evaluate, both, the average speed w_γ and the fraction of time stopped in extended red phases b_γ experienced by the backward-moving observers of "Family 3" in Section 2.3. To find these quantities we only have to reverse the direction of distance in Fig. B1 and recognize that the observers now travel with speed w instead of u_f. Since the direction reversal does not change any street descriptors except the offset, which changes from δ to $\delta_w = C - \delta$ (see figure), the desired results are still given by (B1)–(B4) after substituting w for u_f and δ_w for δ.

References

Ardekani, S., Herman, R., 1987. Urban network-wide traffic variables and their relations. Transportation Science 21 (1), 1–16.

Daganzo, C.F., 1998. Queue spillovers on transportation networks with a route choice. Transportation Science 32 (1), 3–11.

Daganzo, C.F., 2005a. Improving city mobility through gridlock control: an approach and some ideas. Working Paper UCB-ITS-VWP-2005-1, UC Berkeley Center of Excellence on Future Urban Transport. University of California, Berkeley, CA.

Daganzo, C.F., 2005b. A variational formulation of kinematic waves: basic theory and complex boundary conditions. Transportation Research Part B 39 (2), 187–196.

Daganzo, C.F., 2005c. A variational formulation of kinematic waves: solution methods. Transportation Research Part B 39 (10), 934–950.

Daganzo, C.F., 2006. On the variational theory of traffic flow: well-posedness, duality and applications. Networks and Heterogeneous Media 1 (4), 601–619.

Daganzo, C.F., 2007. Urban gridlock: macroscopic modeling and mitigation approaches, Transportation Research Part B 41(1), 49–62. Corrigendum. Transportation Research Part B 41 (3), 379.

Edie, L.C., 1963. Discussion of traffic stream measurements and definitions. In: Almond, J. (Ed.), Proceedings of the 2nd International Symposium on the Theory of Traffic Flow. OECD, Paris, France, pp. 139–154.

Franklin, R.E., 1967. Single-lane traffic flow on circular and straight tracks. In: Edie, L.C., Herman, R., Rothery, R. (Eds.), Proceedings of the 3rd International Symposium on the Theory of Traffic Flow, Vehicular Traffic Science, pp. 42–55.

Gartner, N., Wagner, P., 2004. Analysis of traffic flow characteristics on signalized arterials. In: Transportation Research Record: Journal of the Transportation Research Board, No. 1883, pp. 94–100.

Geroliminis, N., Daganzo, C.F., 2007. Macroscopic modeling of traffic in cities. In: 86th Annual Meeting of the Transportation Research Board, Paper # 07-0413, Washington, DC.

Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transportation Research Part B 42 (9), 759–770.

Godfrey, J.W., 1969. The mechanism of a road network. Traffic Engineering and Control 11 (7), 323–327.

Kuwahara, M., 2007. Private communication.

Mahmassani, H., Williams, J.C., Herman, R., 1987. Performance of urban traffic networks. In: Gartner, N.H., Wilson, N.H.M. (Eds.), Proceedings of the 10th International Symposium on Transportation and Traffic Theory. Elsevier, Amsterdam, The Netherlands.

Mahmassani, H.S., Peeta, S., 1993. Network performance under system optimal and user equilibrium dynamic assignments: implications for ATIS. Transportation Research Record 1408, 83–93.

Olszewski, P., Fan, H.S., Tan, Y.W., 1995. Area-wide traffic speed-flow model for the Singapore CBD. Transportation Research Part A 29 (4), 273–281.

Wardrop, J.G., 1952. Some theoretical aspects of road traffic research. Proceedings of the Institution of Civil Engineers, Part II 1 (2), 325–362 (discussion, 362–378).

Wardrop, J.G., 1965. Experimental speed/flow relations in a single lane. In: Almond, J. (Ed.), Proceedings of the 2nd International Symposium on the Theory of Road Traffic Flow. OECD, Paris, France, pp. 104–119.

Williams, J.C., Mahmassani, H.S., Herman, R., 1987. Urban traffic network flow models. Transportation Research Record 1112, 78–88.