EURO J Transp Logist (2012) 1:47–65
DOI 10.1007/s13676-012-0003-z

RESEARCH PAPER

The potential of parsimonious models for understanding large scale transportation systems and answering big picture questions

Carlos F. Daganzo · Vikash V. Gayah · Eric J. Gonzales

Received: 26 September 2011 / Accepted: 2 March 2012 / Published online: 4 April 2012

Abstract A model with few variables is said to be parsimonious. If it is also analytically tractable, physically realistic, and conceptually insightful, it is said to be effective. Effective parsimonious models have long been used in fields such as economics and applied physics to describe the aggregate behavior of systems as opposed to the behavior of their individual parts. In transportation, these models are particularly well suited to address big picture questions because they provide insights that might be lost when focusing on details. This paper presents an abbreviated history of effective parsimonious models in the transportation field, classified by sub-area: regional and urban economics, traffic flow, queuing theory, network dynamics, town planning, public transportation, logistics, and infrastructure management. The paper also discusses the benefits of these models—fewer data requirements, reduced computational complexity, improved system representation, insightfulness—and ways of constructing them. Two examples, one from logistics and one from urban transportation, are used to illustrate these points. Finally, the paper discusses ways of expanding the application of effective parsimonious models in the transportation field.

Keywords Effective parsimonious models · Macroscopic modeling · Continuum approximations · Logistics · Urban mobility

C. F. Daganzo
University of California, 416A McLaughlin Hall, Berkeley, CA, USA
e-mail: daganzo@ce.berkeley.edu

V. V. Gayah (✉)
University of California, 416G McLaughlin Hall, Berkeley, CA, USA
e-mail: vikash@berkeley.edu

E. J. Gonzales
Rutgers University, CAIT 120, Piscataway, NJ, USA
e-mail: eric.gonzales@rutgers.edu

Mathematics Subject Classification 41A99

Introduction

Transportation professionals are often tasked with answering big picture questions about large scale systems. There are a number of different approaches for understanding, analyzing, and addressing these problems, ranging from the very detailed to the very coarse. This paper examines the potential of parsimonious models to study large-scale transportation systems. Parsimonious models rely on few assumptions and have relatively few degrees of freedom or input parameters (see, e.g., Gabaix and Laibson 2008). However, to be truly useful parsimonious models should also be tractable, physically realistic, and conceptually insightful in the sense of Gabaix and Laibson (2008). A model is tractable if it is easy to work with (such as an analytical formula), physically realistic if it is consistent with real-world behavior, and conceptually insightful if it reveals fundamental properties of the system of interest. In this paper, parsimonious models that meet these three conditions are called "effective".

Effective parsimonious models, if they exist, are particularly useful to describe the aggregate behavior of large systems with many agents. For this type of model to exist, some aggregate features of the system it represents should exhibit orderly behavior despite any underlying complexity at the individual agent level. This turns out to be the case for many systems. Organized aggregate behavior can be unveiled in two ways: either by analysis (using approximations to smooth out details and obtain simple analytical formulae) or by experimentation (observing relationships between system variables in empirical or simulated data). Big picture questions about large systems that exhibit this organized behavior can often be answered without detailed information about the system and its agents. For example, a freight company can use effective parsimonious models to determine the optimal number of storage facilities to build without knowing the exact locations of customers that need to be served. Similarly, transportation engineers can use these types of models to design network-wide strategies to allocate scarce street space to buses and cars without detailed information about travel demand patterns or bus routes.

Parsimonious models are not the norm in transportation where, perhaps because of advances in computing power, the trend is toward increasingly detailed models that can describe the behavior of individual components very well. Although detailed models may be able to predict how particular components of a system will behave, they are not well suited for searching a large space of policy options or revealing general insights about a system. Detailed models are most useful to evaluate specific policy options when detailed information is available and disaggregate predictions of individual elements are desired. These models can also be used to unveil orderly aggregate behavior that can be described with effective parsimonious models if it exists and just as importantly to identify when it does not.

This paper explores the benefits of using effective parsimonious models to study large transport systems. "Past use of effective parsimonious models" discusses past use of these models in the fields of transportation economics, traffic, public transit,

logistics, and infrastructure management. "Benefits of effective parsimonious models" illustrates the benefits of effective parsimonious models in two domains where these models exist: logistics and urban transportation. Finally, "Conclusion" suggests possible research directions.

Past use of effective parsimonious models

Effective parsimonious models have long been used to study phenomena when the behavior of a population in aggregate is more important than the behavior of each individual agent. Standard economic models typically look at the demand and consumption behavior of groups of people rather than tracking each individual. Similarly, models in applied physics fields such as fluid mechanics and thermodynamics typically deal with groups of particles rather than each individual part in detail. For example, fluids are modeled in a macroscopic way because the relationship between useful aggregate properties such as temperature, pressure, and velocity can often be predicted, whereas the path of each molecule is unknowable. Transportation has much in common with economics and physics because transportation systems typically involve many individual agents and also because model users are primarily interested in the agents' aggregate impact. The recent trend in modeling transportation systems has been to develop increasingly detailed, microscopic, and computationally intensive mathematical programs that take advantage of developments in computing power. However, there is also some history of using effective parsimonious models in many fields related to transportation. These include regional and urban economics, traffic flow, queuing theory, network dynamics, town planning, public transportation, logistics, and infrastructure management.

Urban economics and regional science grew out of the field of economics and asked questions about population groups and towns rather than individuals. Some early models (e.g., Christaller 1933; Beckmann 1968) looked at the impacts of transportation costs on regional development. Starting in the 1960s, researchers began to ask how urban space is organically self-allocated between development and transportation. Alonso (1964), for example, examined the trade-off between land rents and transportation costs in a monocentric city by extending earlier work on agricultural land use (Von Thünen 1826) to urban development. Economic equilibrium models have been extended to (1) recognize the aggregate spatial requirements of transportation infrastructure (Solow and Vickrey 1971), (2) identify the costs and externalities of transportation, including the effects of congestion (Solow 1972, 1973), and (3) describe equilibrium urban land use patterns (Anas et al. 1998).

Parsimonious traffic models also have a long history. They were created to describe the effects of vehicle interactions within a traffic stream. Perhaps the most famous example is Greenshields (1935) which used a conjectured linear relationship between the average speed and density of vehicles traveling on a road to derive an aggregate speed-flow relationship. A related third relationship between flow and density has come to be known as the "fundamental diagram." Parsimonious models of traffic flow have also been created to describe the spatio-temporal dynamics of

traffic on a single road using fluid approximations, e.g., the LWR model of Lighthill and Whitham (1955) and Richards (1956). As demonstrated in Newell (1961), these models are very good aggregate approximations of detailed car-following models in the genre of Chandler et al. (1958).

Also noteworthy is the connection between traffic flow theory and models of queuing systems. Moskowitz (1965) showed how vehicle trajectories on a link could be jointly represented as a three-dimensional surface from which cumulative curves of vehicle counts at discrete locations may be extracted to analyze various phenomena. Cumulative curves date back to Edie and Foote (1959). They turned out to be particularly useful to describe queuing phenomena in an aggregate way (Newell 1971), and to express the LWR theory of traffic flow in queuing terms (Newell 1993a, b, c). These queuing models can also be combined with parsimonious economic equilibrium models to derive insights into commuters' arrival and departure times at a road's single bottleneck when the physical extents of the queues are ignored, e.g., as in the original model by Vickrey (1969) and the many works that followed it. Recently, these models have been extended to multi-bottleneck freeways leading to a CBD, accounting for the effect of queue spillovers and merging (Lago 2003).

Models that are parsimonious in one context may not be parsimonious in others. For example, traffic models that parsimoniously describe vehicle interactions along a link are not parsimonious when used to describe vehicle interactions on networks because the resulting models then require require time-dependent origin-destination tables and produce link by link results. Parsimonious models for networks should be even more abstract because they must describe traffic behavior aggregated across many links, e.g., an entire CBD. Early efforts to unveil network-level (i.e., macroscopic parsimonious models), such as Smeed (1963, 1966), used aggregated traffic data to identify macroscopic relationships between road area and the network's capacity to serve traffic. Other macroscopic network models have sought to relate average traffic speeds with network flows based on properties of the network (Thomson 1967; Wardrop 1968; Zahavi 1972a, 1972b). The 'two-fluid' model based on the fraction of vehicles that are moving and stopped (Herman and Prigogine 1979; Herman and Ardekani 1984) also described macroscopic relationships between network-level traffic variables. Godfrey (1969) investigated the relationship between network flow and network vehicle density and appears to be the first to have developed a macroscopic relationship that realistically includes congested conditions in urban networks by allowing for low speeds and high densities for low flows. More recently, Daganzo (2007) proposed an aggregate theory of network dynamics based on conjectured macroscopic relationships between average network density, average network flow, and average network exit rate. These relationships were confirmed to exist on networks with well-connected streets and well-informed drivers (Geroliminis and Daganzo 2008). The relationship between network density and network flow has come to be known as the "macroscopic fundamental diagram" (MFD). The relationship between network density and network exit rate will be called here the "network exit function" (NEF). Daganzo and Geroliminis (2008) stipulated conditions under which a network should exhibit both an MFD and a NEF. This reference also showed how these

relations are theoretically related to network characteristics such as its total length, the number of intersections, and the average trip length.

In addition to describing the behavior of transportation systems, parsimonious models have also been used to optimize their design. For example, there is a history of using these kinds of models to design roadway networks. Several studies, perhaps spurred on by the new town planning movement following World War II, used principles of geometric probability to develop analytical relationships between some of the network's geometric properties, e.g., its arterial and freeway spacings, and its performance (Creighton et al. 1960; Smeed 1963; Tanner 1968; Reynolds 1971; Newell 1980). The insights from these models are useful both for designing and explaining the geometry of road networks.

Effective parsimonious models have also been used to design public transit systems. Holroyd (1967) presented analytical formulae to optimize the bus route spacing and service frequency of a grid network. Subsequently, similar analyses of transit networks have considered different geometry and demand patterns (Byrne 1975; Newell 1979; Wirasinghe and Ghoneim 1981; Daganzo 2010). Parsimonious approaches have even been applied in a multimodal context to design coordinated bus and rail transit systems (Wirasinghe et al. 1977). Estrada et al. (2011) extended the model in Daganzo (2010), showed how to translate the model results into detailed plans, and demonstrated that the aggregate cost predictions of the parsimonious model were realistic.

Aggregate analysis methods have also been used to develop effective parsimonious models to solve optimal facility location (Newell 1973), vehicle routing (Eilon et al. 1971; Daganzo 1984a, b), and other problems arising in logistics, e.g., Larson and Odoni (1981) and Daganzo (1991). The latter presented ways to design logistics systems with aggregate methods and demonstrated the value of macroscopic methods for modeling goods movement and distribution on a large scale. A review of parsimonious models for logistics systems design as of 1996 can be found in Langevin et al. (1996). More recent work in this area has dealt with applications such as vehicle routing problems with service frequency decisions (Francis and Smilowitz 2006) and facility location considering network resiliency (Cui et al. 2010). All the works in this genre are associated with analytical models in contrast to the detailed genre, which uses mathematical programming to generate designs. Hybrid methods that use analytic models to obtain preliminary designs and numerical methods to iron out the final details have also proven useful in both vehicle routing and facility location problems (Robusté et al. 1990; Ouyang and Daganzo 2006).

Effective parsimonious models have also been used for transportation system infrastructure management. Golabi et al. (1982) proposed a pavement management system that groups highways across Arizona into performance states based on the condition of the pavement. By also assigning maintenance activities actions to groups rather than to individual facilities, the system is described so parsimoniously that it can be easily monitored and optimized. Similar aggregation techniques have been applied to bridge management systems (Golabi and Shepard 1997). Others have extended the models to consider uncertainty in measurements and forecasts (Smilowitz and Madanat 2000; Kuhn and Madanat 2005).

The main advantage of parsimony in the above formulations for optimizing facility location, logistics, and infrastructure systems is that they change very difficult problems involving integer variables and uncertainty into much simpler ones with continuous variables and differentiable functions.

Benefits of effective parsimonious models

Effective parsimonious models potentially offer five types of benefits: fewer data requirements, reduced computational complexity, improved system representation, transparency, and insightfulness. The first benefit is obvious since by definition parsimonious models have very few degrees of freedom and require little data. Note that when these data consist of system-wide averages of agent and network characteristics, as is often the case, the amount of data is invariant with system size. The second benefit is also obvious since effective models are analytically tractable. Thus, they are computationally simple. Furthermore, effective parsimonious models can sometimes be combined to represent a large system in terms of linked analytic modules, and in this role they also simplify computation. The third benefit is less obvious. Parsimonious models can sometimes represent a large system more accurately than detailed models because statistical relations needed to capture uncertainty can often be expressed more realistically at the aggregate level—and in some cases can only be formulated at this level. Moreover, due to the law of large numbers, statistical relations become simpler and more robust with increased aggregation. "Logistics: designing a system of warehouses" expands on this idea. The final two benefits follow from the simple nature of the models. Transparent, generic insights are obtained because relationships between important parameters are encapsulated by transparent formulas. This can help improve understanding and lead to the identification of simple and effective policies.

By contrast, detailed models require information about each of the agents within a system. The amount of required data increases with the size of the system. For many systems this includes information that is ultimately unknowable, like detailed time-dependent origin-destination data of trips that have not yet occurred on urban traffic networks. By their nature, detailed models require computers and complex numerical methods to be applied. As a result, the models are less transparent, difficult to verify and can only be queried if one has access to the machine. This can be problematic because, even though these models can describe the behavior of individual components very accurately, inaccuracies can still arise due to the underlying assumptions, uncertainty in the data, and errors in data collection or entry. If these pitfalls are not recognized, detailed models might give the user a false sense of precision. Note as well that the stochastic nature of many detailed simulations often makes it difficult to separate noise in the output from the meaningful relationships that are of interest; and that the results of these models are numerical lists, which tend to obscure insight. Still, detailed models, used meticulously, are useful for testing designs and policies with complete and accurate data and for fine-tuning them over small solution spaces.

The remainder of this section illustrates some of the benefits of effective parsimonious models with two examples. "Logistics: designing a system of warehouses" uses the surprising interplay between modeling approximations and uncertain data to derive analytically an effective parsimonious model that can be used to optimize the design of a network of warehouses more easily and more accurately than with a more detailed approach. "Urban mobility: management of an urban network with multiple modes" uses a two-module parsimonious model to show how an urban street network can be optimally configured and managed during a rush-hour to accommodate both private and public transportation. One of the modules is derived analytically, while the other is derived empirically.

Example 1, Logistics: designing a system of warehouses

Before considering the design problem, we first examine the distribution costs from a single warehouse. "Accuracy of continuous approximations in the presence of uncertainty" shows that if the warehouse's shipment sizes are uncertain, more accurate cost estimates are obtained by smoothing the exact, discontinuous relation between shipment cost and shipment size. This demonstrates that accuracy is not necessarily sacrificed (i.e., system representation improves) when approximations are used to derive a parsimonious model. "Continuous approximations for design of large systems" then uses this approximation to derive a multi-warehouse parsimonious model and optimize the configuration of a warehouse system. It also demonstrates how the parsimonious model produces insights while reducing data requirements and computational complexity.

Accuracy of continuous approximations in the presence of uncertainty

The following example is adapted from Daganzo (1987, 1991). Consider a local warehouse that ships items to I customers every day. Individual customers are indexed by i where $i = 1, ..., I$, and are located at a distance d_i (km) from the warehouse. On any given day, n, where $n = 1, ..., N$, customer i requires $v_{i,n}$ truckloads of items. Trucks dispatched from the warehouse visit only one customer per trip; therefore, a whole truck must be dispatched to carry the partial load if $v_{i,n}$ is not an integer. The transportation cost required to serve customer i on day n is:

$$c_{i,n} = \lceil v_{i,n} \rceil (c_s + c_d d_i), \tag{1}$$

where $\lceil v_{i,n} \rceil$ is the smallest integer greater than or equal to $v_{i,n}$, $c_s = 100$ (\$/truck) is the fixed cost to dispatch a single truck and $c_d = 1$ (\$/truck-km) is the variable cost per truck-kilometer traveled.

Customer demand is known at the end of the day when trucks are dispatched and the cost to serve all I customers can be calculated exactly only at this time. Suppose, however, that for other decision-making purposes costs need to be estimated in advance. At this time, an estimate of customer demand, $u_{i,n}$, is available. This estimate differs from the actual demand by a random amount $\varepsilon_{i,n}$ so that for any given demand, $v_{i,n}$, the estimated amount is $u_{i,n} = v_{i,n} + \varepsilon_{i,n}$. We assume that the $\varepsilon_{i,n}$

Table 1 Store location and demand data

i	1	2	3	4	5
d_i (km)	100	200	300	400	500
$v_{i,1}$ (actual truckloads)	5.5	7.2	5.7	2.3	1.8

are independent draws of a normal random variable with mean 0 and variance σ_ε^2. Since the $\varepsilon_{i,n}$ are unknown, the cost to serve each customer can be estimated by replacing the unknown demand in (1) with its estimate:

$$e_{i,n} = \lceil u_{i,n} \rceil (c_s + c_d d_i). \tag{2}$$

Note that $\lceil u_{i,n} \rceil$ is a discontinuous function with discrete jumps. It might be convenient in our application to smooth out this term by approximating it with another that is continuous with respect to $u_{i,n}$. To this end, we choose the linear function that minimizes the maximum error with respect to (2):

$$e'_{i,n} = (u_{i,n} + 0.5)(c_s + c_d d_i). \tag{3}$$

We wish to determine how well (2) and (3) estimate actual transportation costs in the system. To this end, let $c_{i,n} - e_{i,n} \doteq \delta_{i,n}$ and $c_{i,n} - e'_{i,n} \doteq \delta'_{i,n}$, and denote the sum of these two error variables over all i as δ_n and δ'_n, respectively. The accuracy of the cost estimates for day n can be determined by comparing the expected mean square errors M_n of δ_n and δ'_n for a given value of σ_ε.

Consider the data presented in Table 1 for $N = 1$ and $I = 5$ and assume for now that $\sigma_\varepsilon = 0.2$. For this low standard deviation, each $\delta_{i,n}$ is a scaled Bernoulli random variable, so it is easy to calculate its mean and variance. The resulting five means and variances can then be composed to yield $\mathrm{M}_1(\delta_1) = 87,928$ ($\$^2$). We also find that $\delta'_{i,n}$ is a normal random variable, and using the same method we calculate that $\mathrm{M}_1(\delta'_1) = 40,900$ ($\$^2$). Note that the mean square error of the discrete formulation is significantly greater than that of the continuous approximation, even though the former is "exact." The reason for this perhaps counterintuitive result is that when the discrete cost expression makes an error, the error amounts to the cost of a full truckload, whereas the errors in the continuous approximation are always relatively small. It turns out that infrequent large errors are more damaging to accuracy than persistent smaller ones.

The reader may wonder whether this result is due to the collection $\{v_{i,1}\}$ chosen for day 1, but this is not the case. To see this, consider now the accuracy of the two approximations for arbitrary I when days are statistically identical, N is large and $v_{i,n}$ varies with n. We assume that $v_{i,n} = v_i + \gamma_{i,n}$ where only the v_i are known and the $\gamma_{i,n}$ are independent draws of a normal random variable with mean 0 and variance $\sigma_\gamma^2 \gg 1$.

Under these conditions, the average round up amount in $\lceil v_{i,n} \rceil$ (i.e., the difference between $\lceil v_{i,n} \rceil$ and $v_{i,n}$) is approximately 0.5 so $E_n(\lceil v_{i,n} \rceil) \approx v_i + 0.5$. The $u_{i,n}$ are now $u_{i,n} = v_i + \delta_{i,n} + \varepsilon_{i,n}$, so they too are independent random draws of a normal random variable with mean v_i and large variance, $\sigma_\gamma^2 + \sigma_\varepsilon^2 \gg 1$. As a result, $E_n(\lceil u_{i,n} \rceil) \approx v_i + 0.5$ in this case too. Therefore, $E_n(\delta_{i,n}) = E_n(\delta'_{i,n}) = 0$ ($), and

both formulas provide an unbiased estimate of cost. Additional manipulations can be performed to find that the mean squared error for any location and day is

$$\mathrm{M}_n(\delta_{i,n}) \approx \begin{cases} 0.8\sigma_\varepsilon(c_s + c_d d_i)^2 & \text{for} \sigma_\varepsilon \leq 0.4, \text{ and} \\ (\sigma_\varepsilon^2 + 1/6)(c_s + c_d d_i)^2 & \text{for} \sigma_\varepsilon > 0.4, \end{cases} \tag{4}$$

and:

$$\mathrm{M}_n(\delta'_{i,n}) = (\sigma_\varepsilon^2 + 1/12)(c_s + c_d d_i)^2. \tag{5}$$

Inspection of (4) and (5) reveals that the continuous approximation provides better estimates every day at every location if $\sigma_\varepsilon > 0.123$. Thus, the exact discrete formulation performs better only when errors in the demand estimation are quite small. If significant uncertainty exists, the approximation is more precise. This result illustrates a broader principle: that if there is uncertainty in the input data of an optimization problem, smoothing a discrete objective function may not just simplify the problem formulation but may also improve accuracy. This observation applies both to "one-shot" estimation problems arising in system design problems and to "multi-period" estimation problems arising in dynamic operations and control problems.

Continuous approximations for design of large systems

We now examine the effect of smoothing discrete variables over both time and space and use this approximation to estimate the distribution costs of a system of warehouses serving many customers. We then show how this type of approximation can be used to optimize the design of this system—something which may be very difficult to do without an approximation. The following is based on Newell (1973) and Daganzo and Newell (1986):

For long term planning, neither the v_i nor the d_i may be known; so there is additional uncertainty. It is assumed, however, that the distribution of customers and the spatio-temporal distribution of demand across some service region are known. As before, we deal with uncertainty using expected values.

As an example, assume that customers are randomly (but uniformly) distributed across space, and first consider the distribution cost for a single warehouse serving I customers in a region of area A (km^2). Assume that the warehouse is fairly centered in this region, and that the region is fairly round in shape, so that $E_i(d_i)$ can be expressed as a function of $A, d(A)$. Also assume that as in the previous section $v_{i,n} = v_i + \gamma_{i,n}$, but since the v_i are unknown they are now independent draws from a normal random variable with known mean v and variance $\sigma_v^2 \gg 1$.

The (unknown) actual transportation cost to serve all I customers over N days is then

$$C = \sum_{n=1}^{N} \sum_{i=1}^{I} \lceil v_{i,n} \rceil (c_s + c_d d_i), \tag{6}$$

which for design purposes is approximated by the knowable quantity:

$$E = \mathrm{E}(C) = \sum_{n=1}^{N} \sum_{i=1}^{I} \mathrm{E}_{i,n}(\lceil v_{i,n} \rceil)(c_s + c_d d(A)). \tag{7}$$

As before, $E_{i,n}\lceil v_{i,n} \rceil \approx v + 0.5$ since the $v_{i,n}$ are independent random draws from a normal distribution with mean v and variance $\sigma_v^2 + \sigma_\gamma^2 \gg 1$. And, for example, if distances are given by an L_1 metric and the region is approximately diamond-shaped, then

$$d(A) = \frac{2\sqrt{2}}{6}\sqrt{A}. \tag{8}$$

(For other shapes only the coefficient of (8) changes, but this change is slight if the shape is approximately round.) Thus, (7) reduces to:

$$E \approx IN\left[(v + 0.5)\left(c_s + c_d \frac{2\sqrt{2}}{6}\sqrt{A}\right)\right]. \tag{9}$$

Equation (9) shows that the total transportation cost over N days can be approximated knowing very little data: just the average daily customer demand (v), the total number of customers (I), and the size of the region (A). Equation (9) is an unbiased estimate of cost since $\mathrm{E}(C) = E$ by construction. Additionally, since var(C) is proportional to IN (see 6) and $(E)^2$ is proportional to $(IN)^2$ (see 9), it follows that $\mathrm{var}(C/E) \to 0$ as $(IN) \to \infty$. Thus, $C/E \to 1$ in probability as $IN \to \infty$. This shows that the accuracy of (9) improves with the scale of the system.

Let us now see how to use this approximation to optimize the design of a system of warehouses. Suppose that in order to reduce total costs, a distribution company wishes to build an undetermined number of warehouses, T, in a service region of area R (km^2) containing J customers. Each additional warehouse costs c_T (\$/day) to operate, but additional warehouses also reduce transportation costs because items can be shipped to customers over a shorter distance. We wish to determine the optimal number and location of warehouses that should be built to minimize the sum of transportation and operating costs, Z.

Without approximations, determining the optimal number of warehouses could be a very difficult problem. Even if the location of each customer and its demand were known exactly (i.e., the set $\{d_i, v_{i,n}\}$ was known for all i and n), the cost and optimal warehouse locations would have to be determined for each value of T to determine the optimal number of warehouses, T^*. This could be done by formulating a tessellation problem with discrete locations and solving this problem with a mixed-integer mathematical program. If uncertainty is included, the solution becomes even more difficult. Stochastic programming methods could be used but these tend to limit the solution space and as a result yield sub-optimal solutions. Thus, finding an optimum solution may be quite difficult, perhaps even impossible, when uncertainty is included.

On the other hand, (9) can very easily be used to construct a differentiable objective function. If the warehouses are evenly distributed in the service region so that each warehouse serves an approximately round area of size $A \approx R/T$ with

$I \approx J/T$ customers each, the objective function becomes $Z \approx ET + Nc_T T$, where E is given by (9). The result is

$$Z \approx c_s JN(v + 0.5) + JNc_d(v + 0.5)\frac{2\sqrt{2}}{6}\sqrt{\frac{R}{T}} + c_T NT. \tag{10}$$

Note that (10) can be used even before knowing the warehouse locations. In particular, we can find

$$T^* = \left[\frac{c_d J(v + 0.5)\sqrt{2R}}{6c_T}\right]^{2/3}, \text{ and} \tag{11a}$$

$$Z^* = c_s JN(v + 0.5) + N\left(2^{1/3} + 2^{4/3}\right)\left[\frac{c_d J(v + 0.5)\sqrt{Rc_T}}{6}\right]^{2/3}. \tag{11b}$$

Equations (11a) and (11b) yield insights that could not be easily observed with numerical methods, even if these methods could yield the optimal solution. For example, we see at a glance that if freight rates drop by a factor α and the remaining data take any values whatsoever but stay invariant, then the number of warehouses should be reduced by a factor $\alpha^{2/3}$; and this reduction in the number of warehouses cuts the variable part of Z^* by the same factor.

We expect for the same reasons as before that the accuracy of (11b) would increase with the scale of the system. Note as well that for this particular problem, the results in (11a) and (11b) are the same both with and without uncertainty. Problems do exist where the results differ if uncertainty is included, but the same basic ideas apply. Simulations in Cui et al. (2010) confirm that the accuracy of this type of approximation improves with scale in these cases too.

Equation (11a) provides a good approximation of the optimal number of warehouses but does not pinpoint their location. Since these locations should define approximately round service areas of equal size, they can be determined with gradual improvement heuristics based on this property, e.g., with the "flexible disk" method in Ouyang and Daganzo (2006). If appropriate, one may then refine the resulting pattern using the exact formulation and another improvement algorithm.

The ideas of this example can and have been used for other transportation design problems, including many where the demand is not uniform, e.g., considering inventories and inbound costs into the warehouses, hierarchies of warehouses, complex multi-modal transportation systems, and real-time dispatching problems. The common idea is that analytical approximations are used to narrow the range of potential solutions to obtain a tentative design; exact, detailed methods are then used to fine-tune the design.

Example 2, Urban mobility: management of an urban network with multiple modes

Let us now consider an urban mobility problem involving cars and buses. "Minimizing delay to cars during a rush hour" shows how an effective parsimonious model of urban traffic derived from empirical data can be used to

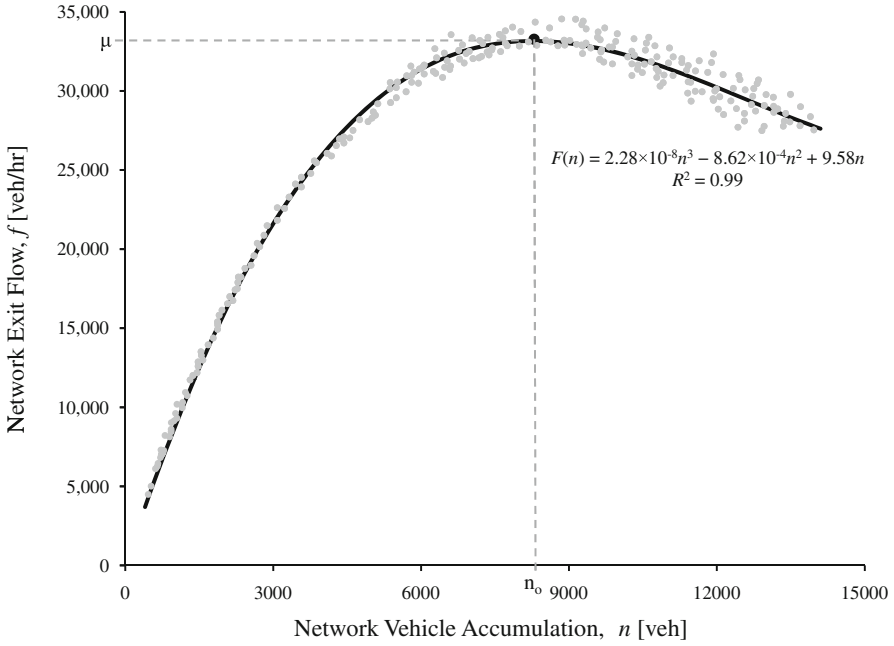

Fig. 1 Network exit function for Yokohama, Japan (source: Geroliminis and Daganzo 2008)

manage an urban network during a rush-hour period when only cars are present. "Optimizing allocation of urban street space to two modes" extends the results to determine how much of the existing car network should be dedicated to transit. This example again demonstrates the benefits of effective parsimonious models in terms of data requirements, computational complexity, and conceptual insightfulness.

Minimizing delay to cars during a rush hour

The following example is based on data for Yokohama, Japan (Geroliminis and Daganzo 2008); see Fig. 1. The street network of Yokohama has a total length of $L = 157$ (lane-km) and has been shown to exhibit both a consistent MFD and constant average trip lengths, $d = 2.3$ (km). Thus, the network has a consistent NEF, $f = F(n)$, where f is the predicted rate at which vehicles complete trips and exit the network, and n is the total number of vehicles on the network. Over the range of observed traffic states, the data of Fig. 1 can be approximated by the following NEF:

$$F(n) = 2.28 \times 10^{-8} n^3 - 8.62 \times 10^{-4} n^2 + 9.58n, \quad \text{for } n \in [0, 14, 100], \qquad (12)$$

which fits the data with $R^2 = 0.99$. Note that $F(n)$ is concave with a maximum exit rate $\mu = 33,168$ (veh/h), achieved for $n_o = 8,271$ (veh).

Equation (12) is specific to Yokohama. A city's NEF depends on network properties such as total network length, block lengths, signal timings, and free-flow

speeds. In particular, if only L' (lane-km) of the total network length L is made available for car use in a way that does not change the MFD significantly, then the new NEF, $F'(n)$, should be related to the old by

$$F'(n) \approx \frac{L'}{L} F\left(\frac{L}{L'} n\right). \tag{13}$$

The new maximum exit rate will then be

$$\mu' = \frac{L'}{L} \mu, \tag{14}$$

and this should occur for an accumulation $n'_o = (L'/L)n_o$.

Assume for now that cars are allowed to use the entire network length ($L' = L$). Consider an idealized rush-hour period in which the network starts out empty, experiences a demand of rate λ_c until N_c cars have arrived, and then the demand rate drops to zero. A control scheme is proposed to limit the rate at which arriving cars are allowed to enter the network to keep it from becoming congested. It is assumed that the control scheme induces cars to wait where they do no harm, e.g., on access streets outside the network or at the trip origins themselves. We wish to determine the optimal cumulative number of cars that should be allowed to enter the network at all times, $G(t)$ or equivalently the time-dependent entry rate, $\dot{G}(t)$. The goal is to minimize the total time that users spend in the system including both time waiting to enter and time being "served" (i.e., traveling) in the network. Since cars exit the network at rate $F(n)$, their average "service time" is $n/F(n)$ if conditions change slowly with time, as per Little's formula.

It is known (Daganzo 2007) that for any cumulative arrival curve, the optimal control strategy is intuitive and quite simple: allow as many cars to enter the network as possible without ever allowing the accumulation to exceed n_o. For our single-peak rush hour, this strategy consists of three stages: (1) at the beginning of the rush, cars are allowed to enter the network as quickly as they arrive, $\dot{G}^*(t) = \lambda_c$, to saturate the network to $n = n_o$; (2) once accumulation reaches n_o, entrance to the network is limited to $\dot{G}^*(t) = \mu$, and since $F(n_o) = \mu$ the accumulation n_o is maintained during the heart of the rush while the system's service rate is maximized; (3) at the end of the rush, when there are not enough cars to maintain the critical accumulation n_o, cars are again allowed to enter the network as they arrive.

An advantage of this strategy is that it only requires one piece of information (n_o), and no modeling. If desired, proxies for n_o that are easy to measure such as a critical average speed or critical average occupancy across a sample of relevant detectors can be used. If a significant number of trips originate outside the central network, the control instrument can be a ring of traffic signals which can be used to restrict entry rate. A drawback of this approach is that it can create queues blocking the roads outside the controlled area. Alternatively, dynamic prices equivalent in magnitude to the queuing cost for each trip can be charged to convert queuing time into productive revenue and incentivize travelers to wait at home.

Figure 2 presents the queuing diagram with the optimal result for our example. The solid black line represents the (given) cumulative arrivals and the dashed black

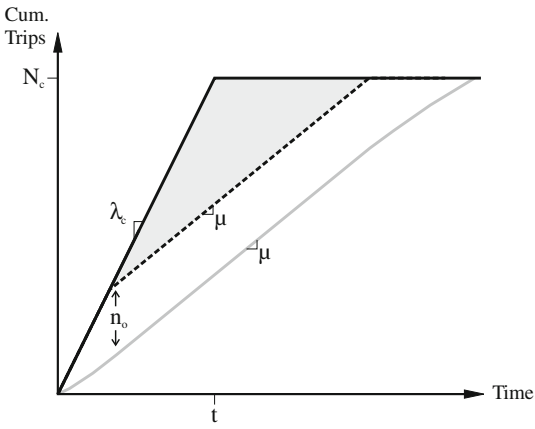

Fig. 2 Queuing diagram for cars during the morning rush hour

line the optimum cumulative entries ($G^*(t)$), and the solid gray line the cumulative exits corresponding to the optimal control.[1] The shaded region in this figure represents the total wait outside the network, and the unshaded region below is the total service time inside the network. Using this queuing diagram, we approximate the total car delay (including wait and service time) as:

$$T_c \approx \frac{N_c^2}{2}\left(\frac{1}{\mu} - \frac{1}{\lambda_c}\right) \text{ for } \lambda > \mu. \tag{15}$$

Note that the total delay is proportional to the square of the number of car users in the rush hour and that it decreases with the service rate, μ.

In this example, a parsimonious model of traffic on an urban network allowed us to identify an efficient control strategy without the extensive data collection, data entry, computation requirements, and possible human errors of microscopic approaches. The part of the Yokohama network under consideration covers an area of $A = 10$ (km^2) and consists of over 1,000 directional links. In order to analyze the effect of a control strategy on the system-wide queuing time and service time with detailed simulation models, each of these links and their intersections would need to be explicitly modeled. This would require data on the properties of each street and traffic signal, as well as detailed time-dependent origin destination information which is expensive and difficult to obtain. Modeling assumptions about the behavior of traffic would also have to be made. It should be clear that a control policy developed on the basis of simulations is not just difficult to optimize, but may not perform as anticipated if the detailed model on which it is based contains significant errors.

[1] The solid gray line in Fig. 2 is slightly curved at the beginning and end of the rush since network accumulations are increasing and decreasing at those times. The curved portions are only an approximation, however, because the NEF model does not describe well situations where the network accumulation changes rapidly. We shall therefore treat the departure curve as if it was linear. The resulting error is not of much consequence when the middle period is long.

Optimizing allocation of urban street space to two modes

Let us now study how much of Yokohama's network should be devoted to transit. To do this macroscopically, focus on the fraction α of the streets dedicated to transit. The more street space that is dedicated for transit, the lower the cost to users of the transit system. However, this lower transit cost comes with a trade-off—higher cost for cars. Thus, the issue is one of balancing costs to users of both modes. The trade-off is explored by linking together two parsimonious models: one analytically constructed for transit and one empirically constructed for cars.

Consider transit users first. We assume that transit is uniformly deployed across the network, exists only on the αL (km) of dedicated lanes, and does not interact with car traffic. If the transit system has a fixed number of users, N_t, and operates with a fixed headway and constant commercial speed, independent of the amount of space provided for transit, then the total costs to transit users is fixed except for the user access cost which depends on the spatial coverage, i.e., on α. Since the total length of the transit network is αL and the system is uniformly distributed across the city, users must travel an average distance of approximately $A/(2\alpha L)$ to access the system on each end of a trip. The average access time for each transit trip is $A/(\alpha L v_w)$, where v_w is an effective average walking speed including delays at crosswalks and recognizing that people dislike walking more than riding. Since each transit user experiences this access time, the α-dependent part of the total transit cost is:

$$T_t(\alpha) = \frac{N_t A}{\alpha L v_w}. \tag{16}$$

Now consider the N_c car users. For simplicity, we define total costs to these users as the total vehicle delay experienced during our idealized rush-hour period. In this bi-modal network, cars can only travel on $L' = (1-\alpha)L$ (km) of the street network. We assume the optimal control strategy is used to limit vehicle entry into the network. Therefore, total car delay can be calculated with (15) after replacing μ with $\mu' = (L'/L)\mu = (1-\alpha)\mu$, as per (14), assuming the optimal control strategy is used to limit vehicle entry into the network. Thus, the total costs to car users when a portion α of streets is dedicated to transit is

$$T_c(\alpha) = \frac{N_c^2}{2}\left(\frac{1}{(1-\alpha)\mu} - \frac{1}{\lambda_c}\right). \tag{17}$$

Note that the transit costs (16) increase with N_t whereas car costs (17) increase with the square of the number of car users. This happens because only cars suffer from congestion.

For a given number of users, the optimal allocation of space is associated with the value of α that minimizes $T(\alpha) = T_c(\alpha) + T_t(\alpha)$. Manipulation of (17) and (16) yields closed form expressions for the optimal values, α^* and T^*:

$$\alpha^* = \left(1 + \sqrt{\frac{N_c^2 L v_w}{2 N_t A \mu}}\right)^{-1}, \text{ and} \tag{18a}$$

$$T^* = \left(\sqrt{\frac{N_c^2}{2\mu}} + \sqrt{\frac{N_t A}{L v_w}}\right)^2 - \frac{N_c^2}{2\lambda}. \tag{18b}$$

Now suppose that the number of morning trips made by car and transit in Yokohama are the same, $N_c = N_t = 70,000$ (trips), that each car only carries a single passenger and that users access the transit network at an average speed of $v_w = 1.5$ (km/h).[2] Using the Yokohama data, we find that $\alpha^* = 0.167$. This means that about 26.2 lane-km of road space should be allocated for transit, and this would be associated with a total cost of $T^* = 36,476$ (pax-h).

Equations (18a) and (18b) yield space-allocation insights that would be hard to obtain using detailed microscopic tools. For example, using first-order Taylor series approximations of (18a) and (18b) one can find the elasticities of α^* and T^* with respect to N_c and N_t, and evaluate the effect of switching $\varepsilon \ll N_c, N_t$ users from car to transit. For the Yokohama scenario, such a switch would increase the optimal street space devoted to transit $4.68\varepsilon \times 10^{-4}$ (km) and decrease the resulting total user cost by 0.280ε (pax-h). So each individual switch saves 0.28 h of travel to society if it is accompanied by 0.47 m of extra bus-lane length.

This model can also be extended to determine an equilibrium distribution of car and transit trips for the city if the demand for each mode can be approximated by a function of their cost, i.e., if $(N_c, N_t) \approx D(T_c, T_t)$. The reader can verify with this type of macroscopic equilibrium analysis that transit investments in a crowded city benefit everyone, including car users. However, investments in car infrastructure are essentially wasted because they attract people who would otherwise ride transit, and as a result cannot reduce total user cost.

More sophisticated transit cost functions have been developed and can be used in general cases where bus headways and commercial speeds are not fixed; see Gonzales (2011). However, the basic point remains. Effective parsimonious models developed with aggregated data can be used to develop tentative space-allocation policies without considering which specific streets and lanes will be used by the transit vehicles. These details can be formalized once a policy has been chosen. A final detailed evaluation should confirm what the analyst already knows.

Conclusion

There is a history of using effective parsimonious models to describe large systems in the transportation field, especially in disciplines such as economics, planning, traffic, logistics, and urban transportation. By focusing on aggregate behavior and ignoring fine details, analysts in these disciplines have developed models that are tractable and can be used to answer big picture questions. Properly formulated, these

[2] We imagine that N_t includes all transit users for half of the day while N_c includes only the car users for the rush-hour period. This is because the spatial coverage of transit affects transit users during both peak and off-peak periods, while it only affects car users for the periods when the network is congested (i.e., during the rush). Since two rush-hour periods occur every day, we associate half of the daily transit users with each rush.

models can be physically realistic and quite accurate and in some cases, particularly if uncertainty is an issue, even more so than their detailed counterparts. Futhermore, since effective parsimonious models are simple and conceptually insightful, they readily yield optimal designs and policies. Although they are not a substitute for detailed approaches, effective parsimonious models of large systems complement more detailed analysis methods. Large-scale preliminary planning or strategic decisions can be made based on these models, fully recognizing a problem's uncertainty. More detailed numerical techniques can then be used to refine the preliminary strategies into detailed final plans.

For the future, it should be useful to identify through analytical methods and empirical experiments additional transportation systems whose large-scale behavior is sufficiently reproducible to be captured by effective parsimonious models. Not all systems fall in this category. If a system's aggregate behavior is unstable or chaotic, as occurs for congested freeway networks, we may have to accept that some of the system's macroscopic features cannot be predicted. However, even in this case, we may look for policies that make the system predictable and for parsimonious models of the resulting behavior. After all, policies of this type are the ones most likely to be desirable. Efforts should also be made to better understand the acuracy of these models.

Applications that combine parsimonious models of different types also seem worth exploring. Examples are the combinations of: transit and cars in cities as in "Optimizing allocation of urban street space to two modes"; urban logistics and urban traffic; transit scheduling and control; country-wide port and land transport policies; and the distribution and use of new forms of energy for automobility.

Efforts should also be made on the numerical front. Detailed numerical tools typically ignore the reproducible system behavior at the aggregate level. Therefore, research into fine-tuning design algorithms that use the information obtained from parsimonious models to obtain detailed implementable solutions is worthwhile. Some fine-tuning tools of this type have been shown to perform well, as was discussed in the review portion of this paper, but these tools do not exist for every application.

In summary, for many high-level planning, design and management problems, effective parsimonious models based on aggregate values provide a fast and accurate method to search across a wide space of possible solutions. Effective parsimonious models of large systems will be one of the tools necessary to address the emerging big picture problems in the transportation field.

Acknowledgments This research was supported by NSF Grant CMMI-0856193 and the UC Berkeley Center of Excellence for Future Urban Transport. The paper also benefited from the extensive comments of three anonymous reviewers.

References

Alonso W (1964) Location and land use: toward a general theory of land rent. Harvard University Press, Cambridge

Anas A, Arnott R, Small K (1998) Urban spatial structure. J Econ Lit 36(3):1426–1464

Beckmann M (1968) Location theory. Random House, New York

Byrne B (1975) Public transportation line positions and headways for minimum user and system cost in a radial case. Transp Res 9(2):97–102
Chandler R, Herman R, Montroll E (1958) Traffic dynamics: studies in car following. Oper Res 7(1):165–184
Christaller W (1933) Central places in Central Germany (translated by C.W. Baskin)
Creighton R, Hoch I, Schneider M, Joseph H (1960) Estimating efficient spacing for arterials and expressways. Highw Res Board Bull 253:143
Cui T, Ouyang Y, Shen Z (2010) Reliable facility location design under the risk of disruptions. Oper Res 58(4–1):998–1011
Daganzo C (1984a) The length of tours in zones of different types. Transp Res Part B 18(2):135–146
Daganzo C (1984b) The distance traveled to visit N points with a maximum of C stops per vehicle: an analytical model and an application. Transp Sci 18(4):331–350
Daganzo C (1987) Increasing model precision can reduce accuracy. Transp Sci 21(2):100–105
Daganzo C (1991) Logistics systems analysis. Springer, Heidelberg
Daganzo C (2007) Urban gridlock: macroscopic modeling and mitigation approaches. Transp Res Part B 41(1):49–62
Daganzo C (2010) Structure of competitive transit networks. Transp Res Part B 44(4):434–446
Daganzo C, Geroliminis N (2008) An analytical approximation for the macroscopic fundamental diagram of urban traffic. Transp Res Part B 42(9):771–781
Daganzo C, Newell G (1986) Configuration of physical distribution networks. Networks 16(2):113–132
Edie L, Foote R (1959) Experiments on single-lane flow in tunnels. In: Herman R (ed) Theory of traffic flow. Proceedings of symposium on the theory of traffic flow, pp 175–192
Eilon S, Watson-Gandy C, Christofides N (1971) Distribution management: mathematical modeling and practical analysis. Hafner, New York
Estrada M, Roca-Riu M, Badia H, Robusté F, Daganzo C (2011) Design and implementation of efficient transit networks: procedure, case study and validity test. Transp Res Part A 45(9):935–950
Francis P, Smilowitz K (2006) Modeling techniques for periodic vehicle routing problems. Transp Res Part B 40(10):872–884
Gabaix X, Laibson D (2008) The seven properties of good models. In: Caplin A, Schotter A (eds) The foundations of positive and normative economics. Oxford University Press, Oxford, pp 292–299
Geroliminis N, Daganzo C (2008) Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transp Res Part B 42(9):759–770
Godfrey JW (1969) The mechanism of a road network. Traffic Eng Control 11(7):323–327
Golabi K, Kulkarni R, Way G (1982) A statewide pavement management system. Interfaces 12:5–21
Golabi K, Shepard R (1997). Pontis: a system for maintenance optimization and improvement of us bridge networks. Interfaces 27:71–88
Gonzales E (2011) Allocation of space and the costs of multimodal transport in cities. Ph.D. thesis, University of California, Berkeley
Greenshields B (1935) A study in highway capacity. In: Highway Research Board proceedings, vol 14, pp 448–477
Herman R, Ardekani S (1984) Characterizing traffic conditions in urban areas. Transp Sci 18(2):101–139
Herman R, Prigogine I (1979) A two-fluid approach to town traffic. Science 204(4389):148–151
Holroyd E (1967) The optimum bus service: A theoretical model for a large uniform urban area. In: Proceedings of the third international symposium on the theory of traffic flow, New York, pp 308–328
Kuhn K, Madanat S (2005) Model uncertainty and the management of a system of infrastructure facilities. Transp Res Part C Emerg Technol 13(5–6):391–404
Lago A (2003) Spatial models of morning commute consistent with realistic traffic behavior. Ph.D. thesis, University of California, Berkeley
Langevin A, Mbaraga P, Campbell JF (1996) Continuous approximation models in freight distribution: an overview. Transp Res Part B 30(3):163–188
Larson R, Odoni A (1981) Urban operations research. Prentice Hall, Englewood
Lighthill M, Whitham G (1955) On kinematic waves II. A theory of traffic flow on long crowded roads. In: Proceedings of the Royal Society of London A, Mathematical and Physical Sciences
Moskowitz K (1965) Discussion of freeway level of service as influenced by volume and capacity characteristics by D.R. Drew and C.J. Keese. Highw Res Record 99:43–44
Newell G (1961) Non-linear effects in the dynamics of car following. Oper Res 9(2):209–229
Newell G (1971) Applications of queueing theory. Chapman and Hall, London

Newell G (1979) Some issues relating to the optimal design of bus routes. Transp Sci 13(1):20–35
Newell G (1980) Traffic flow on transportation networks. MIT Press, Cambridge
Newell G (1993a) A simplified theory of kinematic waves in highway traffic, part I: general theory. Transp Res Part B Methodol 27(4):281–287
Newell G (1993b) A simplified theory of kinematic waves in highway traffic, Part II: queueing at freeway bottlenecks. Transp Res Part B Methodol 27(4):289–303
Newell G (1993c) A simplified theory of kinematic waves in highway traffic. Part III: multi-destination flows. Transp Res Part B Methodol 27(4):305–313
Newell GF (1973) Scheduling, location, transportation, and continuum mechanics; some simple approximations to optimization problems. SIAM J Appl Math 25(3):346–360
Ouyang Y, Daganzo C (2006) Discretization and validation of the continuum approximation scheme for terminal system design. Transp Sci 40(1):89–98
Reynolds D (1971) Urban layout and transport systems: a theoretical/practical study. Roads and Transportation Association of Canada
Richards P (1956) Shockwaves on the highway. Oper Res 4(1):42–51
Robusté F, Daganzo C, Souleyrette R (1990) Implementing vehicle routing models. Transp Res Part B 24(4):263–286
Smeed R (1963) The road space required for traffic in towns. Town Plan Rev 33(4):279–292
Smeed RJ (1966) Road capacity of city centres. Traffic Eng Control 8(7):455–458
Smilowitz K, Madanat S (2000) Optimal inspection and maintenance policies for infrastructure networks. Comput Aided Civil Infrastruct Eng 15(1):5–13
Solow R (1972) Congestion, density and the use of land in transportation. Swed J Econ 74(1):161–173
Solow R (1973) Congestion cost and the use of land for streets. Bell J Econ Manag Sci 4(2):602–618
Solow R, Vickrey W (1971) Land use in a long narrow city. J Econ Theory 3:430–447
Tanner J (1968) A theoretical model for the design of a motorway system. Transp Res 2(2):123–141
Thomson JM (1967) Speeds and flows of traffic in London: 1. Sunday traffic survey. Traffic Eng Control 8(11):672–676
Vickrey W (1969) Congestion theory and transport investment. Am Econ Rev 59(2):251–260
Von Thünen J (1826) Isolated state (translated by C. M. Wartenberg)
Wardrop JG (1968) Journey speed and flow in central urban areas. Traffic Eng Control 9(11):528–532
Wirasinghe S, Ghoneim N (1981) Spacing of bus stops for many to many travel demand. Transp Sci 15(3):210–221
Wirasinghe S, Hurdle VF, Newell GF (1977) Optimal parameters for a coordinated rail and bus transit system. Transp Sci 11(4):359–374
Zahavi Y (1972a). Traffic performance evaluation of road networks by the α-relationship: part 1. Traffic Eng Control 14(5):228–231
Zahavi Y (1972b) Traffic performance evaluation of road networks by the α-relationship: part 2. Traffic Eng Control 14(6):292–293