Denos C. Gazis

Traffic Flow and Control: Theory and Applications

The car increases man's mobility, until all decide to exercise this mobility simultaneously in space and time; then we must call traffic science to the rescue

Man does not live by bread alone—in modern days he also needs a car. Much as we malign the car as a killer and polluter of our environment, we love it dearly, and the love will endure because it is based on mutual dependence and self-sacrifice. If the car is here to stay, so is the need to understand the process which converts traffic into the menace it sometimes is. But it is only during the last decade or so that the scientific method has been applied on a large scale in developing a traffic science.

This new science has addressed questions related to understanding traffic processes and to optimizing these processes through proper design and control. The former questions could be described as basic research and the latter as applied research, with no pejoration or praise implied by either term. In fact, the distinction is by no means sharp and is made here only for reasons of convenience. I shall discuss here some examples of both basic and applied research, with one notable omission: I shall not discuss applications of network theory to transportation planning, which is a major subject outside the scope of this article. My discussion will be in the nature of sampling rather than thorough reviewing, with a definite bias toward topics with which I have been personally involved. Any implication that these topics are among the most important ones is, of course, intentional.

Dr. Denos C. Gazis is Director of the General Sciences Department at the IBM Watson Research Center in Yorktown Heights, New York. Born in Greece, he was educated at the Technical University of Athens, Stanford University, and Columbia University, receiving a Ph.D. in engineering mechanics from Columbia in 1957. He has worked in the fields of mechanics, applied mathematics, and operations research, and has written on a wide spectrum of subjects ranging from solid state physics to traffic theory and computer models of urban housing. Before joining IBM, he worked at the General Motors Research Laboratories in Warren, Michigan. He moved to IBM in 1961 and was named to his present position in 1971. He has lectured on traffic theory at Yale University and has organized summer schools on the subject on behalf of the Operations Research Society of America and in collaboration with Yale, Purdue, and George Washington universities. In 1959 he was corecipient of the Lanchester Prize of Operations Research. Address: IBM Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598.

Traffic flow theories

One way of describing traffic is by considering it as a fluid continuum flowing on our highways. Those who doubt that traffic behaves like a fluid may be convinced by observing Figure 1: not only does traffic flow like a fluid but sometimes it appears to leak and flood an entire area. One of the earliest papers describing traffic as a fluid was that of Lighthill and Whitham (*1*), who used kinematical concepts to describe waves in traffic. The basic premises of their model are (1) that traffic is conserved—i.e. any net increase in density at a point is exactly accounted for by a net inflow of traffic; and (2) that there exists a one-to-one relationship between speed and density along a highway—cars driving at any given (constant) speed fall into this prescribed density, after a few transitory maneuvers which may be neglected for the purposes of investigating "kinematical waves."

Of the two premises, the second is the least satisfactory, on two grounds. First, as Lighthill and Whitham themselves recognized, the speed-versus-density relationship may be more complex than just an algebraic relationship. For example, the speed may depend not only on the density itself but also on the time and/or space derivative of the density. Second, transient motion of the vehicles becomes important when changes of the traffic conditions are appreciable and frequent. For these reasons, the Lighthill-Whitham model should be viewed only as a good first approximation, as was apparently intended by its authors and overlooked by many discussants.

Mathematically, the Lighthill-Whitham model states that the density, k, and flow, q, satisfy the relationship

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = 0 \qquad (1)$$

at any time t and for any point x on a highway. Equation (1) expresses the principle of conservation of cars. In addition, q and k are assumed to satisfy a relationship

$$q = q(k) \qquad (2)$$

From these assumptions, it follows that

$$\frac{\partial k}{\partial t} + V\frac{\partial k}{\partial x} = 0, \qquad V = \frac{\partial q}{\partial k} \qquad (3)$$

which has the solution

$$k = F(x - Vt) \qquad (4)$$

where F is an arbitrary function.

Equation (4) implies that inhomogeneities, such as changes in concentration of cars, propagate along a stream of cars with constant speed V with respect to a stationary observer. For example, a small change of speed propagates with speed $\partial q/\partial k$, which is positive or negative depending on whether the concentra-

Figure 1. Leakage of the traffic fluid in Tokyo (reprinted by permission, Mainichi Shimbun, © Time-Life, Inc.).

tion is below or above an optimum concentration corresponding to maximum flow (Fig. 2). Furthermore, the theory predicts the existence of shock waves when the density of cars is higher in the direction of traffic movement. That is, a shock propagates along a stream of traffic, and cars change speeds abruptly as they pass through the front. This somewhat startling feature, which is not actually observed in real life, stems from the fact that the theory neglects the detailed maneuvers of cars in changing speeds.

The Lighthill-Whitham theory, although not a perfect one, is still the best available description of kinematical waves in traffic, particularly if the intensity of perturbations is not too large. It was used by Lighthill and Whitham to provide a fair description of the behavior of traffic in front of bottlenecks and the periodic disturbances caused by traffic lights.

Let us now look at another class of models of traffic flow, the car-following models. These models, which represent the behavior of individual cars as they fight for survival and a place in a line of cars moving along a highway, were first considered by Reuschel (*2*) and Pipes (*3*) in the early 1950s. They were developed extensively, and checked against experimental measurements, by the General Motors school of traffic theorists, which was formed and led by Herman and has included Montroll, Potts, Rothery, and this author among others. The basic results of the car-following theory, described in a series of papers (*4–9*), are the following.

Every driver who finds himself in a single-lane traffic situation is assumed to react mainly to a stimulus from his immediate environment according to the relationship

$$(\text{Reaction})_{t+T} = \lambda(\text{Stimulus})_t \quad (5)$$

where λ is a sensitivity coefficient and T a reaction time-lag, the combined effect of the sluggishness of the driver and his car. It is reasonable to consider as reaction the acceleration of the car, over which the driver has direct control through the brake and gas pedal. The stimulus was assumed to be a function of the position of the car, the position of its neighbors, and the time-derivatives of these positions. It was conjectured, and verified experimentally, that the strongest stimulus was the relative speed of the car with respect to the car in front. If one further assumes that the sensitivity, λ, is constant, he obtains the "linear car-following model"

$$\frac{d^2x_n(t+T)}{dt^2} = \lambda\left[\frac{dx_{n-1}(t)}{dt} - \frac{dx_n(t)}{dt}\right] \quad (6)$$

in which n denotes the position of a car in a line of cars and x_n the position of the nth car along a highway. The model of equation (6) was used to investigate the following questions: (a) Assuming a pair of cars, one following the other, under what circumstances would a maneuver cause a collision? (b) Assuming a long line of cars, under what circumstances does a perturbation by the car in front get amplified as it propagates down the line of cars?

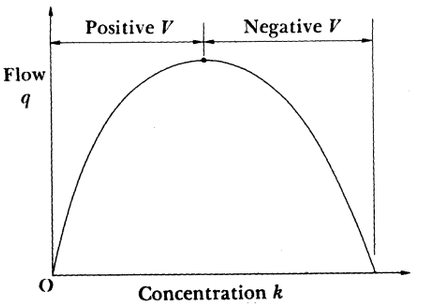


Figure 2. Diagram showing the general character of the flow-versus-concentration relationship assumed by Lighthill and Whitham, and the region of positive and negative speed of perturbations, V.

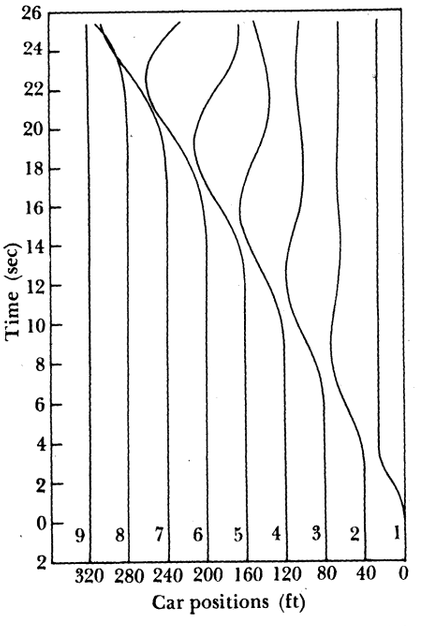


Figure 3. Numerical solutions of Eq. (6) assuming a platoon of identical cars driven in a manner corresponding to asymptotic instability.

The first question refers to the "local stability" and the second to the "asymptotic stability" of traffic. By application of transform techniques on Equation (6) it was found that the key to the answer to both questions was the quantity λT, the product of the sensitivity and the time lag. Specifically, local stability changes as λT moves through two critical values, $1/e \approx 0.367$ and $\pi/2 \approx 1.57$. When $\lambda T \leq 1/e$, any perturbation between two cars decays exponentially; when $1/e < \lambda T \leq \pi/2$, the perturbation has a damped oscillatory character; when $\lambda T > \pi/2$, the perturbation has an amplified oscillatory character, indicating local instability. With regard to asymptotic stability, there is another critical value of λT, equal to 1/2. A "signal" generated by the leader of a long platoon of cars characterized by an identical value of λT is amplified or attenuated as it propagates down the line of cars depending on whether λT is greater than or equal to 1/2.

Figure 3 shows a mathematical experiment done by Herman and his colleagues at General Motors. Using the model of Equation (6), they computed the trajectories of a platoon of cars which are forced to adjust to a perturbation introduced by their leader. While they are all driving at constant speed v, the leader slows down for about a second (say, to observe a passing attractive distraction), and then accelerates back to his original speed. All trajectories are plotted with respect to a coordinate system moving at constant speed v. We see that all cars fall back, with respect to their position before the perturbation, by a constant amount, but they do so after varying degrees of adjustment. The parameters λ and T corresponding to Figure 3 have been adjusted to correspond to asymptotic instability; namely, each driver amplifies the signal before passing it along. The result is a collision between the seventh and eighth car in line.

(Figure 3 has an interesting history. It was seen by an acquaintance of mine who later confided to me that she got into the habit of observing her position in a platoon and leaving the highway if she happened to be seventh in line, because she could

not stand the suspense. Being seventh in line is not, of course, particularly ominous. One or more people can get into trouble regardless of their position in line, as shown in Fig. 4.)

The linear car-following model yields a fair approximation of the behavior of cars undergoing relatively small deviations from a state of flow at constant speed. A number of improvements to the model of Equation (6) have been suggested for the purpose of extending its range of applicability. The first of these suggestions (*6*) was that the sensitivity might be inversely proportional to the distance between leader and follower: the farther a driver gets from the car in front the less strongly he reacts to stimuli.

The motivation for this model was an attempt to obtain a good phenomenological relationship between speed and flow, starting from basic considerations of the behavior of individual cars. The argument goes as follows: If a long line of cars driving at constant speed is characterized by a car-following model, then speed changes result in changes of distance which can be obtained from the car-following equation by a simple integration. The integration constants are then fixed by requiring that the relationship thus obtained be valid over a range of variables including certain end conditions.

The "inverse-spacing" model yielded a phenomenological flow-versus-concentration relationship which had been earlier obtained by Greenberg (*10*) starting with a continuum hydrodynamic model. Later, Edie (*11*) suggested that a different model might describe traffic better at low densities. He proposed a model in which the sensitivity was proportional to the speed of a car and inversely proportional to the square of the distance. A general class of nonlinear car-following models was then suggested by Gazis et al. (*12*), in which the sensitivity λ is given by

$$\lambda = \frac{\left[\dfrac{dx_n(t+T)}{dt}\right]^l}{[x_{n-1}(t) - x_n(t)]^m} \qquad (7)$$

This model includes as special cases the inverse-spacing model ($l = 0$, $m = 1$), and Edie's model ($l = 1$, $m = 2$). An extensive check of several combinations of values of l and m was carried out by us but failed to show the clear superiority of any combination over others. It should be emphasized at this point that all models of the type of Equation (7) have some obvious limitations. They neglect certain important constraints imposed by the mechanical limitations of the car and the overall driving environment. For example, the acceleration capability of a car decays rapidly with speed, and the speed itself may be limited by law or desire. For all these reasons, any attempts to obtain yet a better fit of experimental data by another set of values of l and m would probably be misguided.

Another interesting approach to the study of traffic flow is that of Prigogine and his colleagues (*13–15*). This approach is in the tradition of the kinetic theory of gases—namely, finding properties of a fluid in the large

Figure 4. An experimental confirmation of asymptotic instability.

by examining the statistics of motion of constituent particles. Prigogine describes a traffic fluid by a probability density for the speed of an individual car, $f(x,v,t)$, which may vary as a function of the time, t, and the coordinate, x, along the highway. This density is assumed to satisfy the equation

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = \left(\frac{\partial f}{\partial t}\right)_{\text{relaxation}} + \left(\frac{\partial f}{\partial t}\right)_{\text{interaction}} \quad (8)$$

The first term of the right-hand element of Equation (8) stems from the fact that $f(x,v,t)$ differs from some desired speed distribution $f^\circ(v)$. The second term corresponds to the slowing down of a fast car by a slow one. The exact form of these terms was selected for mathematical convenience and plausibility. The specific form suggested by Prigogine and his colleagues is

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = \frac{f - f^\circ}{\tau} + (1 - p)\,k(\bar{v} - v)f \quad (9)$$

where τ is a characteristic relaxation time, p is the probability of a car passing another one, $\bar{v}$ is the average speed of traffic, and k is the concentration of cars.

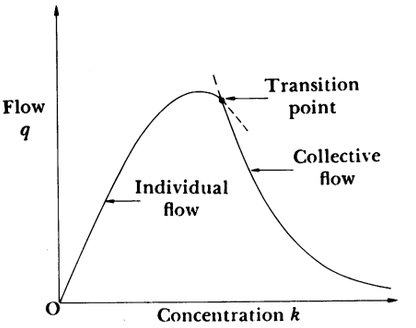


Figure 5. Flow-versus-concentration relationship obtained by means of the Prigogine model of traffic flow on a multilane highway.

If we are interested only in solutions of Equation (9) that are independent of time and space, then the left-hand element of the equation is zero, and it may then be solved (*14*, *15*) to yield an equation of state whose general form for small values of the concentration is an approximately linear increase of flow with concentration; that is,

$$q \approx \bar{v}^\circ\, k \quad (10)$$

where $\bar{v}^\circ$ is the average of the desired speed. As k increases, the flow q falls below the straight line (Eq. 10) due to the increasing influence of interactions. In the range of high concentrations, q is independent of f° and depends only on τ and p, according to the equation

$$q = \frac{1}{\tau(1 - p)} \quad (11)$$

The curve (Eq. 11) may be viewed as a universal curve of "collective flow," which is characterized by high densities and very little passing. For any given f°, the flow q rises with k along a curve characteristic of the specific f° until it intersects the curve, then decreases monotonically with k, according to Equation 11 (Fig. 5). One very realistic feature of this theory is that it predicts probable stoppage of some vehicles in the case of collective flow, which is certainly in agreement with the common experience of stop-and-go traffic at high concentrations.

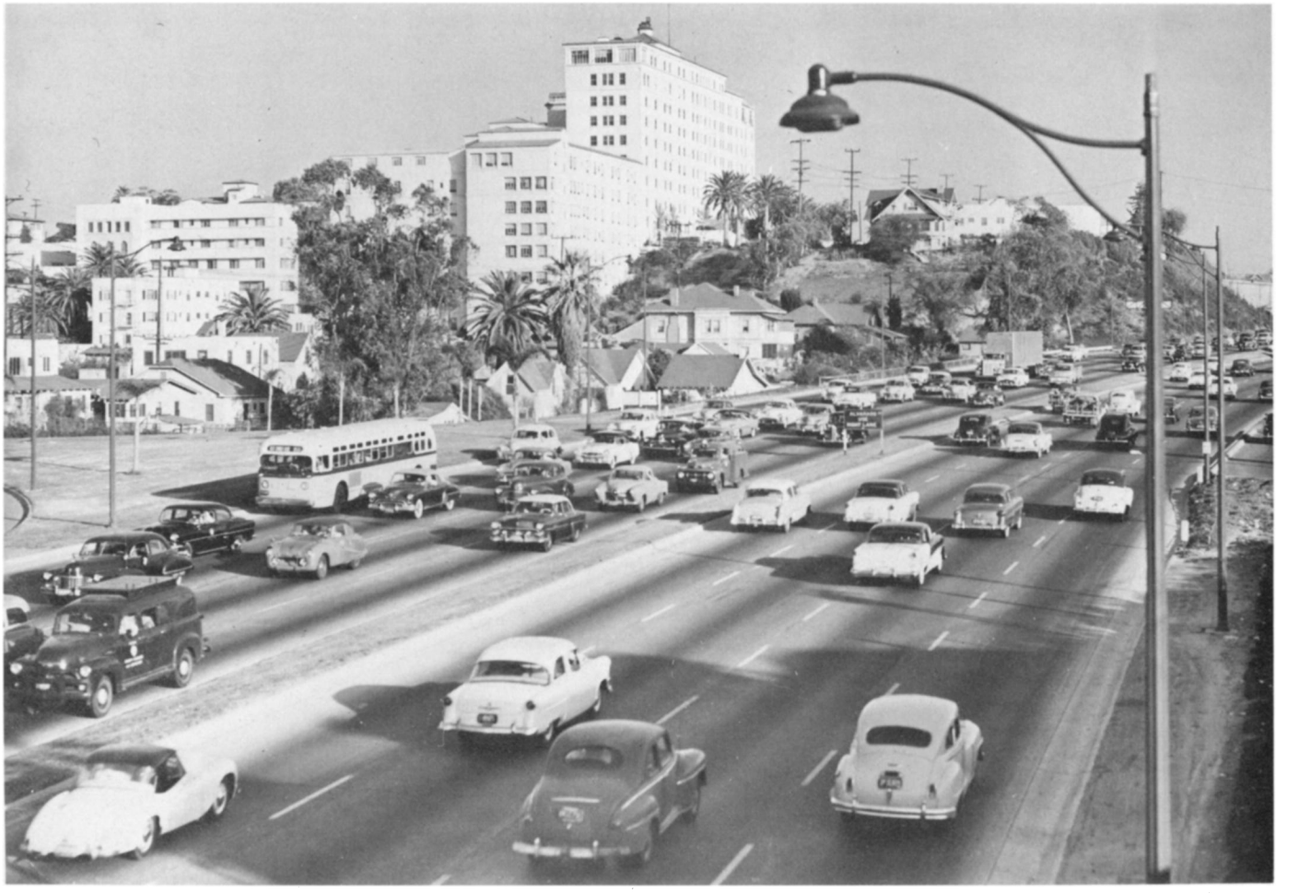


Figure 6. An illustration of light traffic (right lanes) and moderately heavy traffic (left lanes) on a freeway in Hollywood, Calif. (Courtesy of Bettmann Archive, Inc.)

Figure 6 shows two traffic situations which illustrate the main points of Prigogine's theory. In the right lanes we see relatively light traffic, in which passing is easily accomplished and the slowing down of fast cars by slow ones is not too great. The heavier traffic in the left lanes has probably reached the state of "collective flow."

I have given only a brief sketch of the foundations of three basic models of traffic flow. Much work has been added in these subjects over the last ten years. The Lighthill-Whitham model has been scrutinized experimentally, and attempts have been made to increase its realism by including various derivatives in the functional relationship between flow and concentration. Car-following models have been even more prominent in the literature. Memory functions have been introduced to describe the behavior of the driver, and the overall performance of the driver has been increasingly subjected to careful measurements. In another vein, the possibility of eliminating the driver from the picture through automation has also been considered, and the synthesis of appropriate servomechanisms for car-following has been considerably advanced by Cosgriff and his collaborators (*16*). The Prigogine model, brilliant though it is, has not attracted many followers, probably because of its exacting requirements in mathematical sophistication.

All the above models have begun to influence the thinking of the newly trained generation of traffic engineers. Although the application of these models in designing traffic systems is sometimes transcendental, they have been useful in giving us some qualitative, and in part quantitative, understanding of some basic phenomena in traffic.

Conflicts in traffic

My favorite illustration of conflicts in traffic is the story of the truck driver waiting at a YIELD sign behind a timid female driver. After five minutes or so, he stuck his head out and yelled: "Lady, it says YIELD, not SURRENDER!"

A good example of the methodologies used in describing conflicts in traffic is the model proposed by Weiss and his collaborators to describe crossing

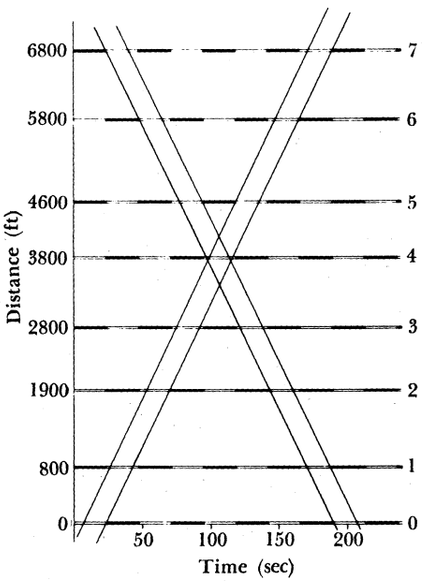


Figure 7. A typical time-span diagram showing the throughband design of progression. A car starting within a throughband can move through all intersections at constant speed without stopping—maybe.

and merging maneuvers at uncontrolled intersections (*17–21*). The simplest case corresponds to a car trying to cross through a single line of traffic. It will be discussed here as an illustration of the methodology of "renewal theory," which was used for its treatment.

Consider a car arriving at an uncontrolled intersection at time $t = 0$. A single line of traffic is passing in front of it, and the time gaps, Δ_i, between cars are assumed to be independent random variables. The probability density for these gaps is some function $\phi(\Delta)$. We also postulate that a driver "accepts" a gap Δ and crosses the street with probability $\alpha(\Delta)$. We can now write a "renewal equation" which must be satisfied by the probability density function $\Omega(t)$ of the waiting time t, using the following argument: Either the driver (a) accepts the first gap or (b) rejects it and starts all over again, in which case the remainder of his waiting time is given by the same probability density function $\Omega(t)$. These arguments lead to an integral equation for $\Omega(t)$, which has been solved for a variety of combinations of gap distributions, $\phi(\Delta)$, and gap-acceptance functions, $\alpha(\Delta)$.

Several generalizations of the preceding discussion have been given—for example, the possibility that the driver accepts the first gap with probability which differs from $\alpha(\Delta)$, because he may, legally or illegally, keep rolling without coming to a complete stop; or the possibility that the driver checks not only the immediately available gap but also the gap following it and waits for the latter if it is larger, even though the immediate gap is adequate; and the case of crossing an n-lane highway (*22*).

If we now consider the problem of queueing of many cars trying to cross a highway at an uncontrolled intersection, we find that the mathematical problem becomes quite complex. Classical queueing theory is not applicable because the "service time" for a car is neither constant nor a simple function of the position of a car in the queue. Crossings by more than one car may be possible when a large gap appears. In any case, the behavior of the second car in line is not the same as if the car had arrived alone. For all these reasons, analytical results on queueing are scarce, and the best results have been obtained by simulation (*23*).

Investigations of the type just described are useful for deciding on a rational basis when an uncontrolled intersection imposes undue delays to motorists and therefore must be controlled. Furthermore, techniques of probability theory can be and have been used for investigating delays at controlled intersections with an eye toward improvement of that control. A good survey of many papers on the subject of conflicts in traffic has been given by Weiss (*24*).

Control of urban street networks

"Ladies and Gentlemen will order their Coachmen to take up and set down with their Horse Heads to the East River, to avoid confusion." Thus, according to Kane's *Famous First Facts* (*25*), spoke a New York City ordinance of 1791 establishing the first one-way street regulation. The regulation was incidental to performance at the John Street Theater, and there is no record indicating how well it worked. It is fair to say, however, that confusion in New York City has managed not only to survive such assaults but also to expand considerably in scope. Things have changed, of course, in New York City over the last 180 years: we have abandoned horses and moved to a much slower but safer mode of trans-

portation, the car and its ubiquitous companion, the double-parked truck. Confusion reigns eternal, only now it is controlled not only by one-way regulations but also by traffic lights. The lights, together with basic instincts of self-preservation, by and large guarantee that no two cars occupy the same space at the same time.

Traffic control by means of traffic signals dates back to 1914. The first progression system appeared four years later, and the evolution of traffic signals produced two basic principles: (1) synchronization of a group of lights, in order to provide maximum opportunity to as many cars as possible to move through the group of intersections without stopping, and (2) flexible operation of "traffic-activated" signals, in order to reduce delay at critical intersections.

In the late 1950s and early 1960s the digital computer made its entry into the traffic control field, first in Toronto (*26*), Canada, and then in San Jose (*27*), California. The systems developed in these cities, and the ones that followed them in the late 1960s both in this country and abroad, made use of the above two principles in traffic control. We may call the first principle *macrocontrol* because it corresponds to a grand strategy for a large area and its time scale of operation is of the order of many minutes, or even hours. The second principle can be called *microcontrol*, because it affects a small area and its time scale is of the order of seconds.

A typical computerized traffic control system has several macrocontrol schemes residing in memory. They are designed off-line on the basis of statistical data and they are called into play when detection of some gross properties of traffic indicates that current conditions are close to the design conditions. In addition, the system may have one or more microcontrol packages which are applied on a second-by-second basis to a limited number of intersections.

The oldest macrocontrol strategy and still the widest used is the maximum through-band design shown in Figure 7. It consists of offsetting the beginnings of the green phases of successive lights with respect to each other in order to allow as many cars as possible to drive through these lights at some design speed without stopping. It is tacitly assumed that the path of these cars is not blocked by other cars, an assumption which is not always realized in practice. Traffic engineers have used nomograms, graphical techniques, and trial-and-error procedures for maximizing through-bands. In recent years, efficient computer algorithms for this purpose have been given by Morgan and Little (*28*) and by Brooks (*29*).

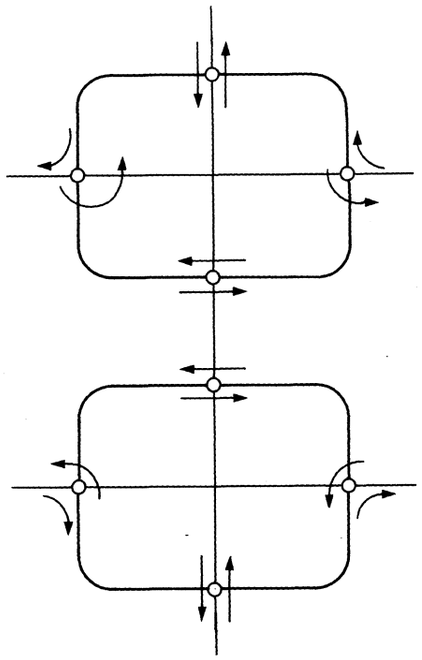


Figure 8. The basic principle of route control showing movements of traffic during two phases of a traffic light.

Starting from a basic progression design, one can make adjustments and obtain others with equal efficiency but different speeds and/or through-bands in the two directions in order to accommodate different traffic demands. Little (*30*) introduced some additional flexibility by allowing small changes of speeds between different pairs of intersections, on the theory that drivers can learn to adapt themselves to such small changes. Allowing some changes could generally result in wider through-bands. At the same time, Little cast the progression-design problem for an arterial network in the framework of mixed-integer linear programming and generalized it to handle any configuration of an arterial network. The applicability of Little's method is only limited by our inability to solve efficiently large mixed-integer linear programming problems.

All through-band designs have one basic deficiency—namely, they presume relatively empty streets. Everyone is familiar through personal experience with progression designs that get clogged because queues of cars block the path of cars through successive green lights. This situation has been discussed in a previous paper (*31*), which suggested that a platoon of cars should be released so that it would reach the tail end of a platoon in front only when all of that platoon would be in motion. This has also been done, directly or indirectly, in some of the synchronization schemes suggested in recent years. These are the San Jose design by Chang (*32*), the TRANSYT method of Robertson (*33*), and the Combination method of Hillier and Whiting (*34*). These methods all effectively simulate the motion of cars through a network, evaluating an objective function which measures the inconvenience to drivers, and then zeroes in on a good set of offsets, through search techniques over the domain of allowable offsets. Differences between these schemes are in aspects of modeling of car movement, such as platoon dispersion, and the method of search used.

I will only say a little about the subject of microcontrol, which is the least tested. We have borrowed ideas from the design of traffic-actuated signals, such as extending a green phase to accommodate an influx of cars along one of the legs of an intersection. Some good theoretical work has been done by Dunne and Potts (*35*) and by Grafton and Newell (*36*) on the question of optimizing the operation of an isolated, undersaturated intersection. Their results indicate that the "saturation flow algorithm," which calls for switching the light as soon as a stopped queue has been completely served, tends to minimize the overall delay in most cases. Delay minimization is one of the targets of control, and moving traffic without stops is another. For this reason, the saturation flow algorithm is not the best solution except in cases of demand approaching saturation, when maximizing utilization of facilities is particularly important. When demand exceeds the saturation level and the intersection becomes oversaturated, a substantially different problem is presented (see below).

The effectiveness of microcontrol may

increase if it is extended to more than one intersection. This has been the starting point of some recent work by Ross et al. (*37*) in developing control algorithms for systems of intersections comprising a critical one and the four nearest it. The algorithm is based on an extrapolated estimate of delays of traffic within the system and on selection of the best of available alternatives. Using the same principle for a large network had been suggested much earlier by Miller (*38*). Application of such control, however, requires expensive detection instrumentation, and in the minds of most experts the expenditure does not appear to be warranted on the basis of expected improvements. The current trend is to think of improving macrocontrol by adaptively redesigning synchronization schemes on-line, on the basis of measurements of platoon movements. However, no experimentation with such ideas has been done to date.

It is time to raise the question of what may be expected by all these control measures. We have had about ten years of experience with computerized traffic control systems, all of which were installed after series of tests of their capabilities. The most extensive testing of control methodologies, in Glasgow (*39*), showed that modern synchronization schemes, such as TRANSYT, can decrease overall delays by as much as 10 percent. Microcontrol can extend the improvement, largely by improving the utilization of critical intersections. However, we appear to be reaching the limit of potential improvement by clever operation of the traffic lights alone.

There is, however, another tool of traffic control which is as yet largely unexploited. It is the optimum, or at least improved, allocation of traffic facilities through *route control*. In a previous paper (*40*), some simple examples were given of route control through simple systems of intersections near a critical one. The general idea is shown in Figure 8. Splitting a stream of traffic into two, separated in space, allows crossing with another stream without reaching saturation. This may be possible sometimes, provided that there are some underutilized roadways near the point of congestion.

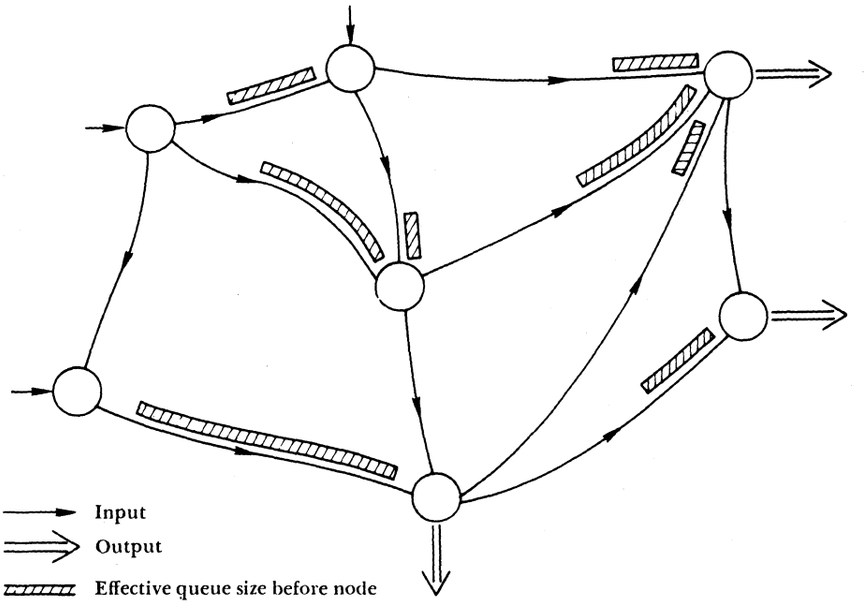


Figure 9. A schematic diagram of a traffic store-and-forward network.

The more general problem of route control is shown schematically in Figure 9, which shows a *traffic store-and-forward* network. Such a network is characterized by a travel time (or travel cost) between nodes, a capacity limit for each arc, and a storage capability at the end of each arc. Traffic moves from node to node at fixed speed and is stored just before the node until it can be dispatched to its destination when space becomes available. In practice, storage may be along the arc, and the delay may be accrued, in part, through the slowing down of the movement of traffic. Two controls may be exercised on such networks: (1) route assignment for each traffic unit from its origin to its destination; and (2) "switching" at the nodes, which determines the rate of service for each queue.

Once a route assignment is selected, a switching strategy can be obtained by using methods of control theory, which minimizes the overall delay to the users of the network. The reader will recognize in this model features of traffic facilities such as freeways, with their systems of ramps and, sometimes, parallel surface streets. What is less obvious is that the *oversaturated intersection*—overloaded during rush hours, by demand that exceeds its capacity—is also a special case of the store-and-forward network (*41–43*). In order to minimize the delay to the users, the allocation of the green time along the various legs of such an intersection must be based on considerations of the queue behavior during the entire rush period, not just at any given instant. In addition, the allocation of green time must satisfy certain constraints: too long green phases mean too long red phases for another direction, and it has been observed that in such cases drivers tend to think that the light has failed and they choose to ignore it. Too short green phases are largely unusable because of the time lost for clearing the intersection, and they present difficulties to pedestrians who are not Olympic sprinters.

Thus, the problem, cast in the framework of control theory, is to optimize an objective function, the delay, by an appropriate choice of a time-dependent control variable, the green phase, which must lie within an appropriate "control domain." The optimum control policy, which can be obtained by techniques of control theory, is known as a "bang-bang" control, an unfortunate term when used in connection with traffic. The term bang-bang means that the optimum control is obtained for the values of the control variable lying along the boundaries of the control domain—in this case, maximum green phase and minimum green phase. During the early part of the rush period, the maximum green phase must be allocated to the stream with the largest saturation flow, and the minimum green phase to the other. At some appropriate time the allocation must be reversed and maintained until the end of the rush period—a pattern

resulting in a simultaneous dissolution of both queues with the minimum overall delay. For more general networks, the analytical difficulties may be avoided by working with difference approximations of derivatives and integrals. The resulting formulation is within the framework of linear programming, and our ability to solve such problems is limited only by their size, which may become quite large for complex networks. It is my belief that the development of a control methodology for traffic store-and-forward networks and the necessary communication equipment holds the greatest promise for improvement of traffic conditions in cities.

Critical traffic links

It has been said that the Long Island Expressway is the biggest parking lot in the country. While this claim may be disputed by some users of the Los Angeles freeways, it dramatizes a recurrent phenomenon of rush-hour traffic—the congestion of critical traffic links. As we have just seen, judicious allocation of the traffic facilities can alleviate congestion, but it has not yet been implemented on a large scale. Meanwhile, attention has been focused on the critical traffic links, for two reasons: safety and the desire to counteract the "revolving door effect," which makes facilities less efficient when they are jammed.

Particularly striking examples are the cross-Hudson tunnels of New York City, connecting Manhattan and New Jersey, which demonstrate in the extreme most of the problems of getting traffic through critical links. These tunnels, partly because of their geometry, degrade appreciably when the traffic density inside them exceeds about 50 cars per mile. Whereas under good traffic conditions the throughput may range between 1,200 and 1,300 cars per lane per hour, it falls to 1,100 to 1,200 during stop-and-go traffic conditions.

Experiments starting in 1959 showed that the throughput of the Lincoln Tunnel could be increased by regulating the input so as to prevent congestion (*44*). The first experiments involved "open loop" control—namely, metering the traffic to allow a fixed number of cars, usually 22 per lane, to enter the tunnel every minute. This fixed number corresponded to the maximum throughput rate that had been observed at the tunnel for any sustained period. As might be expected, this control worked well part of the time. However, once the tunnel became congested—for any reason—the open-loop control of the input could not bring it back to optimum operating level. Closed-loop experiments were initiated, using specially constructed control gear, leading to the joint study by The Port of New York Authority and IBM begun in the spring of 1966.

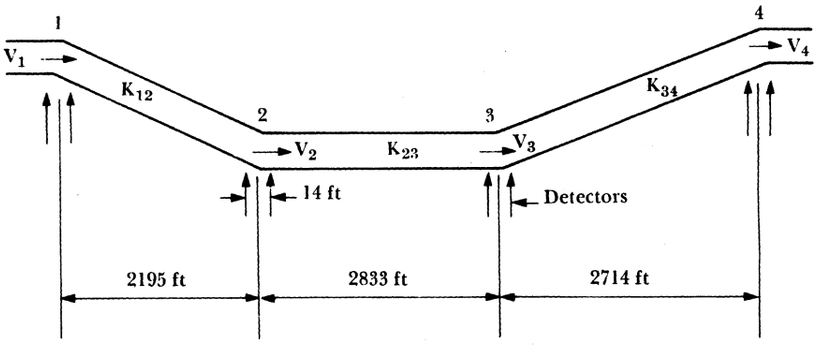


Figure 10. Geometric configuration of the Lincoln Tunnel showing the four locations of the detector pairs.

The study aimed at developing a methodology for the surveillance and control of the Lincoln Tunnel using a digital computer (*45*, *46*), and it went through two phases, first using a 7040 computer for the control of one lane of the south tube of the tunnel, and then using a process control system 1800 for the control of both lanes of this tube. Both computers were located at the IBM Research Center in Yorktown Heights, about forty miles from the tunnel. Figure 10 shows the geometric configuration and Figure 11, the hardware configuration of the system during the second phase of the experiment.

Pairs of photocells, about 14 ft apart, formed observation points, or "traps," at four locations on each of the two lanes of the south tube, as shown in Figure 10. The four traps were located about half a mile apart, at the entrance, foot of the downgrade, foot of the upgrade, and exit of the tunnel. Signals from the photocells were multiplexed and transmitted to Yorktown Heights via a telephone line. There they were "unscrambled" and fed into the 1800 computer via a specially constructed interface. The signals were processed from the tunnel at regular intervals, usually 5 seconds, to produce the following steps.

1. A record was obtained of the speed, length, and time of passage of every vehicle that passed each trap.

2. The exact number of cars between traps was obtained by matching the patterns of the lengths of cars passing through successive traps. This number is equal to the number of cars entering a section between traps, while an identifiable pattern of vehicles moves from the upstream to the downstream trap.

3. On the basis of an ad hoc control algorithm, a decision was made by the computer as to whether or not the input of traffic should be reduced. If such reduction was called for, because the traffic density had increased above optimal level, a signal was transmitted to the tunnel via another telephone line. This signal activated a system of signs and lights which had the overall effect of reducing the input rate by about 30 percent—a reduction sufficient to bring the density down near or at the optimum operating level.

4. Monitoring information was displayed on the on-line printer, and all data was stored on a magnetic disc for future off-line processing.

The control algorithm of step 3 was developed by analyzing data of performance of the tunnel, collected during a "tune-up" phase. The algorithm was based on the premise that the performance of the tunnel can be predicted reasonably well by a function of three "state variables," the vehicle counts for the three sections between traps, for each lane. Measurements during uncon-

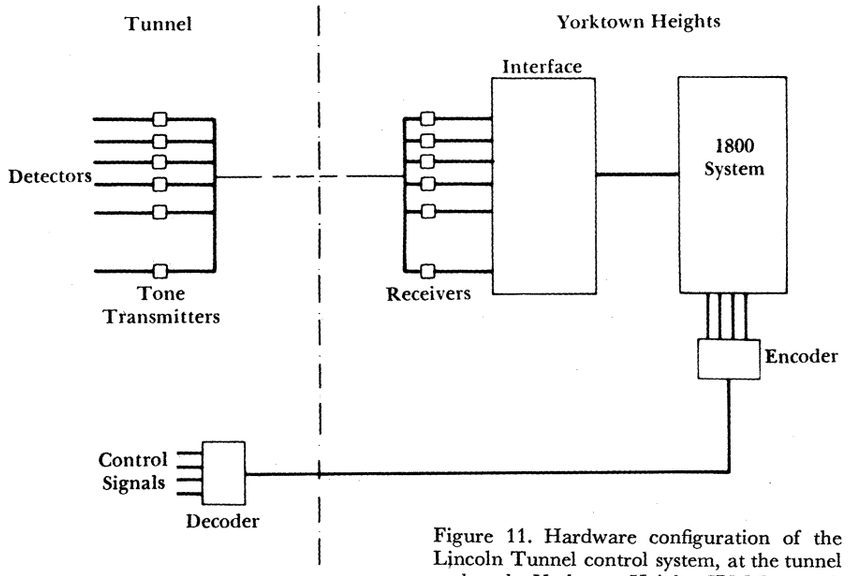


Figure 11. Hardware configuration of the Lincoln Tunnel control system, at the tunnel and at the Yorktown Heights IBM Research Center.

trolled conditions indicated that the premise was correct, in that the probability of congestion, in the absence of input control, increased as the three counts increased.

Therefore, a boundary was estimated in state space K_{12}, K_{23}, K_{34} (Fig. 10), the crossing of which, through increasing values of K_{ij}, triggered the control signal. A second boundary, closer to zero values of K_{ij}, signaled the removal of the control restraint, for decreasing values of K_{ij}. An additional adaptive feature varied these boundaries as a function of the measured speed at the foot of the upgrade, the most frequently observed bottleneck, and called for earlier application of input control when the observed speed there was unusually low, too low for the observed conditions. This had the overall effect of compensating for unobservable influences, such as conditions outside of the tunnel. The control signal, a combination of an amber light and a sign, "STOP HERE THEN GO," caused a decrease in the input rate of traffic by about 30 percent.

The results of the Lincoln Tunnel experiment indicated that regulation of the input could increase the average throughput by as much as 10 percent. In addition, the speeds through the tunnel were reasonably high, around 30 mph, instead of the usual 10 mph observed during stop-and-go conditions. This fact, together with the overall reduction of traffic densities, not only was more satisfactory to the drivers but also had the salubrious effect of reduced ventilation requirements for the tunnel, as well as possibly reduced breakdowns of cars owing to overheating. The philosophy behind this control system is currently being implemented for all the tubes of the Lincoln and Holland tunnels, using a small digital computer for each tunnel, with a duplicate back-up computer available for increased reliability.

Freeways, another critical traffic link, become congested because the cumulative net input along their length frequently exceeds the capacity of a section of the freeway. There is only one remedy for this problem, and again it is to be found in allocation of resources. Metering of traffic at the entrance points can prevent congestion, for the sake of both safety and improved efficiency, by minimizing the "revolving-door effect," which affects freeways as well as tunnels, although perhaps to a lesser degree.

There is no disagreement concerning the need for input control in freeways, but there are different philosophies concerning the control, among which are the following:

1. The gap-acceptance philosophy promulgated by Drew (*47*). According to this approach, cars are allowed to enter a freeway only when there is a sufficiently long gap in the lane nearest to the entrance ramp. One danger in this policy is that the large gap may be compensated for by a high density wave just in front, in which case the availability of space for the entering car is only illusory.

2. The Wattleworth scheme (*48*), which assumes some known, constant origin-destination requirements of all cars entering the freeway at various entrance ramps. Rates of input are then adjusted so that the capacity constraints on the various freeway sections are satisfied and the overall number of vehicles served per unit of time is maximized. It may be seen that this formulation leads to a linear-programming problem, which can be solved by standard algorithms, such as the Simplex method. A basic defect of the method is that it does not take into account the variation of demands with time. Wattleworth has suggested a partial correction of this deficiency by subdividing the peak period into several periods associated with different values of the various parameters of the problem, such as demands and origin-destination coefficients.

3. The freeway can be viewed as a special case of store-and-forward network. Together with its entrance and exit ramps it forms a tree network, shown with solid lines in Figure 12. If we include available frontage roads (dashed lines), we obtain a more general store-and-forward network. A time-dependent allocation of roadway sections can minimize the aggregate delay to all users of the system during the rush period.

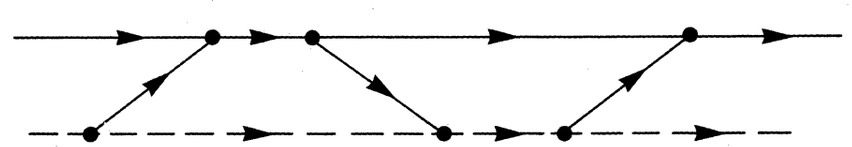


Figure 12. Representation of a freeway and its neighboring surface streets as a network. The solid lines represent the freeway and its ramps, and form a tree-like network.

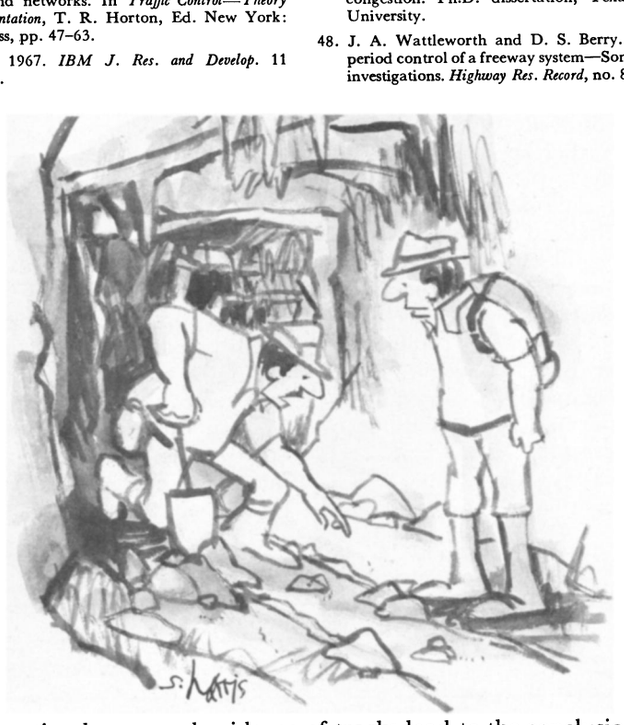
Transportation planning

I have often been asked whether or not we are gaining on traffic with

all our research. Being an optimist I have generally answered affirmatively. I would be first to admit that much of the preceding discussion ignores one basic need—that of judicious transportation planning. I have assumed that traffic in all its messy glory is imposed on us, and we traffic theorists are asked to study it and somehow improve it by manipulating a limited assortment of control devices. However, the fundamental question facing our urban areas today is whether or not all those cars should be there in the first place. The question is loaded with political implications, which is another way of saying that we are as yet unable to decide on an objective function in transportation planning. I feel optimistic about our ability to do even that some day, and then maybe there will be no more traffic problems to solve. However, as a self-appointed congestion theorist, I can only say that I doubt that the day of my unemployment is exactly imminent.

References

1. M. J. Lighthill and G. B. Whitham. 1955. On kinematic waves, 2. A theory of traffic flow on long crowded roads. *Proc. Roy. Soc.* (London) 229 A: 317–45.
2. A. Reuschel. 1950. Fahrzeugbewegungen in der Kolonne bei gliechformig beschleunigtem oder verzogertem Leitfahrzeug. *Zeit. Osterr. Ing. und Archit. Ver.* 95: 59–62, 73–77.
3. L. A. Pipes. 1953. An operational analysis of traffic dynamics. *J. Appl. Phys.* 24: 274–81.
4. R. E. Chandler, R. Herman, and E. W. Montroll. 1958. Traffic dynamics: Studies in car following. *Operations Res.* 6: 165–84.
5. R. Herman, W. Montroll, R. B. Potts, and R. W. Rothery. 1959. Traffic dynamics: Analysis of stability in car following. *Operations Res.* 7: 86–106.
6. D. C. Gazis, R. Herman, and R. B. Potts. 1959. Car-following theory of steady-state traffic flow. *Operations Res.* 7: 499–505.
7. R. Herman and R. B. Potts. 1961. Single-lane traffic theory and experiment. In *Theory of Traffic Flow*, R. Herman, Ed. Amsterdam: Elsevier, pp. 120–46.
8. D. C. Gazis, R. Herman, and R. W Rothery. 1963. Analytical methods in transportation: Mathematical car-following theory of traffic flow. *Proc. Am. Soc. Civil Engrs., Mech. Div.* 6: 29–46.
9. R. Herman and R. W. Rothery. 1959. Microscopic and macroscopic aspects of single lane traffic flow. *J. Operations Res. Soc. Japan* 5: 74–93.
10. H. Greenberg. 1959. An analysis of traffic flow. *Operations Res.* 7: 79–85.
11. L. C. Edie. 1963. Car following and steady-state theory for noncongested traffic. *Operations Res.* 33: 21–27.
12. D. C. Gazis, R. Herman, and R. W. Rothery. 1961. Nonlinear follow-the-leader models of traffic flow. *Operations Res.* 9: 66–76.
13. I. Prigogine. 1961. Boltzmann-like approach to the statistical theory of traffic flow. In *Theory of Traffic Flow*, R. Herman, Ed. Amsterdam: Elsevier, pp. 158–64.
14. I. Prigogine and F. C. Andrews. 1960. A Boltzmann-like approach for traffic flow. *Operations Res.* 8: 789–97.
15. R. L. Anderson, R. Herman, and I. Prigogine. 1962. On the statistical distribution function theory of traffic flow. *Operations Res.* 10: 180–96.
16. R. L. Cosgrift. 1965. Dynamics of Automobile Longitudinal Control Systems for Automobiles. *In* Theory and Design of Longitudinal Control Systems for Automobiles. Communication and Control Systems Laboratory, The Ohio State University, Columbus, Ohio, Report No. EES 202A-8, September 1965, pp. 235–351.
17. G. H. Weiss and A. A. Maradudin. 1962. Some problems in traffic delay. *Operations Res.* 10: 74–104.
18. R. Herman and G. H. Weiss. 1961. Comments on the highway crossing problem. *Operations Res.* 9: 828–40.
19. G. H. Weiss. 1963. A note on highway transparency. *Operations Res.* 11: 460–61.
20. R. Herman and G. H. Weiss. 1963. *J. Res. NBS* 67B: 229.
21. G. H. Weiss. 1964. Effects of a distribution on gap acceptance functions on pedestrian queues. *J. Res. NBS* 68B: 31–33.
22. D. C. Gazis, G. F. Newell, P. Warren, and G. H. Weiss. 1967. The delay problem for crossing an n-lane highway. *Vehicular Traffic Science*. Proc. 3rd Internat'l. Symp. on the Theory of Traffic Flow. New York: Elsevier, pp. 267–79.
23. D. H. Evans, R. Herman and G. H. Weiss. 1964. The highway merging and queueing problem. *Operations Res.* 12: 832–57.
24. G. H. Weiss. 1965. In *Proc. Symp. on Congestion*, W. L. Smith and W. E. Wilkinson, Eds. Chapel Hill: University of North Carolina Press, p. 253.
25. J. N. Kane. 1964. *Famous First Facts*. 3rd ed. New York: The H. W. Wilson Company, p. 626.
26. S. Cass and L. , Casciato. 1960. Centralized traffic signal control by a general purpose computer. *ITE Proceedings*, pp. 203–11.
27. San Jose traffic control project—Final report. IBM Corporation, Data Processing Report, San Jose, Calif., Dec. 1966. See also D.C. Gazis and. O. Bermant. 1966. Dynamic Traffic Control Systems and the San Jose Experiment. Proc. 8th Internat'l. Study Week in Traffic Eng., Barcelona, Spain, vol. 5, pp. 1–9.
28. J. T. Morgan and J. D. C. Little. 1964. Synchronizing traffic signals for maximal bandwidth. *Operations Res.* 12: 896–912.
29. W. D. Brooks. 1964. Vehicular traffic control—Designing arterial progression using a digital computer. IBM Data Processing Div., Kingston, N. Y. (in-house report).
30. J. D. C. Little. 1964. The synchronization of traffic signals by mixed-integer linear programming. *Operations Res.* 14: 568–94.
31. D. C. Gazis. 1965. Traffic control, time-space diagram, and networks. In *Traffic Control—Theory and Instrumentation*, T. R. Horton, Ed. New York: Plenum Press, pp. 47–63.
32. A. Chang. 1967. *IBM J. Res. and Develop.* 11 (4): 436–41.
33. D. I. Robertson. 1969. "TRANSYT" method for area traffic control. *Traffic Engr. & Control* 11: 276–81.
34. J. A. Hillier. 1966. Appendix to Glasgow's experiment in area traffic control. *Traffic Engr. & Control* 7:569–71.
35. M. C. Dunne and R. B. Potts. 1964. Algorithm for traffic control. *Operations Res.* 12: 870–81.
36. R. B. Grafton and G. F. Newell. 1967. Optical policies for the control of an undersaturated intersection. *Vehicular Traffic Science*. Proc. 3rd Internat'l. Symp. on the Theory of Traffic Flow. New York: Elsevier, pp. 239–57.
37. D. W. Ross, R. C. Sandys, J. L. Schlaeffi, and S. H. Hutchins. 1969. Critical subnetwork control—A new approach to urban traffic control. Final Report, Stanford Research Inst., Menlo Park, Calif., Dec. 1969.
38. A. J. Miller. 1965. A computer control system for traffic neworks. In Proc. Second Internat'l. Symp. on Traffic Theory. Paris: Organization for Economic Cooperation and Development, pp. 200–20.
39. J. Holroyd and J. A. Hillier. 1969. Area traffic control in Glasgow—A summary of results from four control schemes. *Traffic Engr. & Control* 11: 220–23.
40. D. C. Gazis and R. B. Potts. 1966. Route control at critical intersections. Proc. 3rd Conf., Australian Road Res. Bd., vol. 3, part 1, pp. 354–63.
41. D. C. Gazis and R. B. Potts. 1965. The oversaturated intersection. Proc. 2nd Internat'l. Symp. on the Theory of Road Traffic Flow. Paris: Organization for Economic Cooperation and Development, pp. 221–37.
42. D. C. Gazis. 1964. Optimum control of a system of oversaturated intersections. *Operations Res.* 12: 815–31.
43. D. C. Gazis. 1965. Spillback from an exit ramp of an expressway. *Highway Research Rec.*, no. 89: 39–46.
44. H. Greenberg and A. Daow. 1960. The control of traffic flow to increase the flow. *Operations Res.* 8: 524–32.
45. B. T. Bennett, D. C. Gazis, L. C. Edie, and R. S. Foote. 1969. Control of the Lincoln Tunnel traffic by an on-line digital computer. Proc. 4th Internat'l. Symp. on Traffic Theory, Bundesminister fur Verkehr, pp. 48–56.
46. D. C. Gazis and R. S. Foote. 1969. Surveillance and control of tunnel traffic by an on-line digital computer. *Transport Science* 3: 255–75.
47. D. R. Drew. 1964. A study of freeway traffic congestion. Ph.D. dissertation, Texas A. & M University.
48. J. A. Wattleworth and D. S. Berry. 1965. Peak-period control of a freeway system—Some theoretical investigations. *Highway Res. Record*, no. 89: 1–25.



"The supporting beams and evidence of tracks lead to the conclusion that the caveman used this tunnel for the express line of his subway system."