TI 2002-102/3
Tinbergen Institute Discussion Paper

Speed Choice, Car Following Theory and Congestion Tolling

Jan Rouwendal

Faculty of Economics and Business Administration, Vrije Universiteit Amsterdam, Wageningen University, and Tinbergen Institute

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Keizersgracht 482
1017 EG Amsterdam
The Netherlands
Tel.: +31.(0)20.5513500
Fax: +31.(0)20.5513555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31.(0)10.4088900
Fax: +31.(0)10.4089031

Speed Choice, Car Following Theory and Congestion Tolling

Jan Rouwendal
Free University, Wageningen University, Tinbergen Institute

This version: September 25, 2002

Key words: Car following, congestion tolling, traffic speed, traffic heterogeneity.
JEL codes: R41, R48, D62.

Abstract
This paper provides a link between car following theory and the economic theory of road congestion by means of a theory of speed choice. According to this theory speed choice is based on a trade-off between the benefits (shorter travel time) and cost (higher accident risk) of driving faster. Accident risk is related to the distance to the 'leader' and by elaborating this relationship a number of car-following models can be derived from this theory of speed choice. Wit homogeneous traffic, steady state analysis leads to a model that generalizes the conventional Pigou-Knight analysis: it has an endogenous speed choice curve and requires the incorporation of accident risk in the value of travel time. A further generalization of this model to steady states with heterogeneous traffic is possible and leads to the conclusion that first best tolls will in general require differentiation over groups of drivers. Finally a general bottleneck model is discussed that contains Vickrey's (1969) and Verhoef's (2002) versions as special cases. This results in a clarification of Verhoef's finding that implementation of the optimal Vickrey toll can result in a deterioration of welfare.

Acknowledgement
Erik Verhoef's useful comments on two earlier versions are gratefully acknowledged.

Jan Rouwendal, Department of Economics, Free University, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, tel 31-20-4446093, fax 31-20-4446004,email jrouwendal@feweb.vu.nl.

1 Introduction

Traffic problems are analyzed from various perspectives. Economists are mainly interested in the external effects involved, such as congestion, and have developed models that bring out the issues involved as clearly as possible. Traffic engineers have been interested in, for instance, the stability of traffic flows and have developed models that concentrate on phenomena such as shock waves and car following behavior. As a result, various branches of literature have developed separately. It is usually difficult to connect research results that originate from different branches since each has developed its own set of models based on different simplifying assumptions. One may nevertheless, conjecture that the various schools of thought might benefit from each other's results. The present paper attempts to substantiate this conjecture by means of an integration of car following theory, which is an important element of the traffic engineering literature, with the Pigou-Knight and bottleneck models of traffic congestion, which are important elements of the economic analysis of traffic problems. The paper intends to make three contributions. The first is to provide a theory of speed choice from which a number of car following models can be derived, thereby providing economic content to this model. The second is to use this economic interpretation of car following theory to develop a generalization of the Pigou-Knight model to heterogeneous traffic. The third is to clarify the relation between Verhoef's (2002) recent bottleneck model based on simple car following theory and the bottleneck model of Vickrey and Arnott, de Palma and Lindsey.

The link between the car following model and the economic models is provided by an economic theory of speed choice. According to this theory, a driver chooses his speed by trading of the benefits of a higher speed (reduced travel time) against the cost (higher accident risk). The accident risk is related to traffic speed and density.

In the Pigou-Knight model traffic density determines traffic speed by means of an aggregate relationship (the speed flow curve or related concepts such as the fundamental diagram) that is often interpreted as a technological one, even though one would expect that a strong behavioral element is involved. The theory of speed choice makes this behavioral content explicit at the individual level. For the individual driver aggregate traffic density and speed are less relevant, but the distance to his leader and his speed can be regarded as appropriate equivalents. Since car following theory uses these variables to explain acceleration behavior, this leads one to conjecture that such models can also be connected to the theory of speed choice. This connection will be formally established below for a number of car-following models.

The static Pigou-Knight model of traffic congestion refers to the stationary state behavior of a population of homogeneous drivers using the same road. This stationary state can be derived from a model in which individual drivers behave in accordance with the speed choice theory. The model with speed choice is richer than the original Pigou-Knight model since the speed-flow relationship is now endogenous. It will be shown below that with an intuitive specification of the cost function the extended model leads to the conventional conclusions with respect to optimal tolling if the cost of accident risk is incorporated in the value of travel time.

The Pigou-Knight model with homogeneous drivers is often used to model actual situations in which vehicles are heterogeneous. Sometimes this heterogeneity is taken into account by using passenger car equivalents in order to homogenize the driver population. The integrated model developed in the present paper allows for an investigation of the adequacy of this conjecture by means of a formal analysis of a model with heterogeneous drivers. In the generalized model the drivers are

differentiated in the values of the parameters of their utility function and therefore in their car-following behavior. In the generalized model first best congestion taxes are in general different for vehicles belonging to different groups. The differentiation of the toll is proportional to the road space occupied by the vehicles, which is the sum of the vehicle length and the space between two subsequent vehicles. The latter is determined by differences in driver behavior that are not necessarily correlated with easily observable characteristics of the vehicles or their drivers. This implies that first best congestion taxes may be difficult to implement in practice.

Verhoef's (2002) analysis of congestion uses a bottleneck model that looks similar to that of Vickrey (1969) and Arnott, de Palma and Lindsey (e.g. 1990) in many respects. Its main novel element is the use of simple car following theory in order to model the behavior of drivers in the queue and in the bottleneck in a more explicit way. His results are in some respects strikingly different from theirs. In particular, he shows that in the context of his model the first best tolling scheme proposed by Arnott, de Palma and Lindsey results in a deterioration of welfare. In order to clarify this difference, which is apparently due to the use of car following theory, the present paper formulates a more general model that is identical to Verhoef's model for one particular value of a parameter. When this parameter approximates a particular limiting value the model exhibits a number of characteristics of the alternative model. It is then shown that in Verhoef's model the flow of cars through the bottleneck is less than the maximum possible flow if the congestion toll removes all queuing in front of the bottleneck, as is optimal in the alternative model. Removing all congestion in Verhoef's model implies therefore that scheduling delay cost will be higher than they are when there is a queue in front of the bottleneck. This is a consequence of his use of the simple car following theory.

The paper is organized as follows. Section 2 contains a brief discussion of car following theory, concentrating on the so called GM-model that covers a number of other specifications as special cases. Section 3 exposes the theory of speed choice. It starts with a general discussion and concentrates on a specific example of the travel cost function. In section 4 the two theories are integrated by using the distance to the 'leader' as the relevant measure of traffic density. It is shown that the speed choice theory leads to a relation between speed and distance to the leader and that differentiation of this relationship leads to the car-following relationship of the GM-model if the sensitivity coefficient of the latter can be interpreted as the first derivative of optimal speed with respect to the distance to the leader. Some other car-following theories that use the relation between distance to the leader and speed as a primitive concept, and for these theories integration with the theory of speed choice is trivial.

Section 5 starts with the derivation of the travel cost function under steady-state conditions. This relationship can be derived from the model on the basis of the optimizing behavior with respect to speed choice. In order to analyze the congestion externality, the model is extended with a demand curve. If the specific form of the travel cost function is valid, maximization of consumer surplus gives the conventional Pigouvian toll provided that the cost associated with accident risk is incorporated in the value of travel time. The model can therefore be regarded as an extension of the conventional Pigou-Knight model that takes the cost function as a primitive concept. Further generalization is possible by allowing for traffic heterogeneity and leads to the result that the composition of traffic will in general be important for the speed-flow relationship and that first best tolls are differentiated.

Section 6 deals with a generalization of Verhoef's (2002) model that allows us to study the reasons behind the substantial difference between his conclusions with respect to optimal tolling and those reached by Vickrey and Arnott, de Palma and Lindsey. The generalization is a more flexible formulation of the simple car-following theory that treats one of its parameters as a variable whose value may change. Section 7 concludes.

2 Car following theory

The intuitive idea that forms the basis of car following theory is that drivers react to the behavior of the vehicle immediately in front of them so as to avoid accidents. This idea was not elaborated by means of a model in which costs and benefits associated with a particular choice of their speed are traded-off against each other leading to a decision to accelerate or decelerate, as an economist would be inclined to do. Instead, it was used as a motivation for finding a descriptive device for driver behavior. In this section we discuss the main components of an important example of this theory, viz. the GM model.[1]

Car following theory[2] deals with the behavior of a series of drivers 1,2,…,n,… on a road. The location of the *n*-th driver at time *t* is denoted as $x_n(t)$. The speed of this driver is denoted as $\dot{x}_n(t) = dx_n / dt$, and the change in speed (acceleration or deceleration) as $\ddot{x}_n(t) = d^2x / dt^2$. The behavioral equations formulated in car following theory concentrate on the relationship between changes in the speed of the *n*-th driver and the difference between the speed of the *n*-th driver and his leader, the (*n*-1)-th driver. The two are related by a sensitivity coefficient λ:

$$\ddot{x}_n(t) = \lambda \left(\dot{x}_{n-1}(t) - \dot{x}_n(t) \right) \qquad (1)$$

This equation says that there will be no acceleration or deceleration if a vehicle proceeds with the same speed as its leader. Speed differences lead to changes in speed that are proportional to these differences with a sensitivity coefficient λ. This sensitivity coefficient is not a constant, but depends on the speed of the *n*-th driver and on the distance between the *n*-th and (*n*-1)-th driver. According to the general formulation of the GM model (see Gazis et al., 1961) , this sensitivity coefficient can be specified as:

$$\lambda = \lambda_0 \frac{\left(\dot{x}_n(t) \right)^m}{\left(x_{n-1}(t) - x_n(t) \right)^l} \qquad (2)$$

This equation says that drivers react stronger to speed differences with their leader if their own speed is higher and if the distance to the leader is shorter. Intuition suggests that the parameters *m* and *l* should both be nonnegative, but in applications sometimes other values are used as will be noted below.

Substitution of (2) in (1) gives:

[1] GM means General Motors, presumably the employer of the developers of this theory.
[2] See e.g. May (1990) for a review and Zhang and Jarret (2001) for a recent contribution.

$$\ddot{x}_n(t) = \lambda_0 \frac{\left(\dot{x}_n(t)\right)^m}{\left(x_{n-1}(t) - x_n(t)\right)^l}\left(\dot{x}_{n-1}(t) - \dot{x}_n(t)\right) \qquad (3)$$

Earlier studies formulated models that are special cases of the GM specification: Chandler et al. (1958) had *m* and *l* equal to 0; Gazis et al. (1959) had *m*=0 and *l*=1; Edie (1961) has *m*=1 and *l*=2.[3]

In car following theory it is usually assumed that drivers react to speed differences with some delay Δt. The reason is that it takes drivers some time to recognize speed differences and react to them. Here we interpret the car following equation as referring to the intentions of the drivers. That is: we interpret the car following equation without delay as reflecting the driver's preferences with respect to reactions to speed differences at given values of the own speed and the distance to the leader.

Attempts to estimate the parameters of car following models seem to be scarce. May and Keller (1967) fitted the model both with integer values of m and l and with non-integer values. They found *m*=1 and *l*=3 the best integer values and *m*=0.8 and *l*=2.8 the best not necessarily integer values. Ozaki (1993) found *m*=-0.2 and *l*=0.2 for acceleration and *m*=0.9 and *l*=1.0 for deceleration. Subramanian (1996) reports *m*=-1.67 and *l*=-0.88 for acceleration and *m*=1.09 and *l*=1.66 for deceleration.[4]

The equations given above reflect the notation used in the literature on car following theory. In what follows we will use a different notation that is more convenient for the present purposes. We concentrate on a single driver or to situations in which all drivers belonging to the same group behave in the same way, so there is no need for using the suffix *n* for individual cars. Instead we use an upper index 0 to refer to variables of the leader. We use the symbol *s* for speed. Moreover, it will be convenient to have a single symbol for the distance with the leader and we will use δ for that purpose. We interpret this distance as the 'nose-to-tail' distance, that is: as the extent of the open space between two cars. Finally, we use *a* for acceleration or deceleration. With this alternative notation, we can rewrite equations (1), (2) and (3) as:

$$a = \lambda\left(s^0 - s\right) \qquad (1')$$

$$\lambda = \lambda_0 \frac{s^m}{\delta^l} \qquad (2')$$

$$a = \lambda_0 \frac{s^m}{\delta^l}\left(s^0 - s\right) \qquad (3')$$

where s^0 denotes the (not necessarily constant) speed of the leader.

[3] A review of car following models is beyond the scope of this paper, see e.g. May (1990).
[4] These results are summarized in Ahmed (1999).

3 Speed choice by minimization of travel cost

This section discusses a theory about speed choice that was introduced in Verhoef and Rouwendal (2001) in order to endogenize speed in models of traffic congestion. The basic idea is that speed is chosen so as to minimize generalized travel costs, which consist of time cost and accident risk. This theory allows one to derive the speed flow curve, which is usually regarded as a technical relationship, on the basis of driver behavior. The model that results from this approach gives an integral treatment of congestion and traffic safety and leads to conclusions that differ from those of the conventional approach. Here we use the theory in a microscopic setting, i.e. we interpret the theory as referring to instantaneous speed choice instead of average speed on a road segment.

Speed choice by minimization of travel costs

Consider a driver at location x who has to determine his speed. The costs c_x of driving at x are a function of the driver's speed s and some measure of traffic density r:

$$c_x = c(s_x, r_x) \tag{4}$$

with c_x a differentiable function defined for nonnegative values of s and r. In what follows we will usually suppress the location index x. However, it should be kept in mind that c refers to the cost at one point in space and that integration over x has to take place in order to arrive at the total cost of a trip.

The specification of the cost function given in (4) is in some respects a restrictive one. For instance, it includes only the driver's own speed and not that of other drivers as an argument. This excludes the possibility that speed differences (which are probably an important element of accident risk) are a determinant of speed choice. The reason for using (4) is a purely pragmatic one: the purposes of the paper can be reached with this specification. The possibility to generalize these results by using alternative specifications of the cost function will be discussed briefly in the concluding section. The function c should be convex in s for every possible value of r. In the next section the appropriate measure of traffic density in the context of the present paper will be considered in some detail. For the moment it suffices to remark that the driver is assumed to take traffic density (however defined) as given. Speed is chosen instantaneously so as to minimize the costs at the given density:

$$s(r) = \arg\min\big(c(s, r)\big) \tag{5}$$

where $s(r)$ denotes the optimal (i.e. cost minimizing) speed, which is a function of traffic density. If speed is positive, $s(r)$ is the solution of the first order condition

$$\frac{\partial c}{\partial s} = 0 \tag{6}$$

Since the function c is convex in s (by assumption) the second-order conditions for a minimum are also fulfilled. We can derive the optimal (=minimal) cost as a function of traffic density by substituting (5) into (4).[5]

[5] The myopic behavior implied by (5) is consistent with minimization of total travel cost $\int c_x dx$ if (a) an individual driver cannot influence the density r_x and (b) there are no binding constraints this minimization due to, for instance, limits on engine capacity (acceleration) and braking power.

It may be noted that (5) implies that we have *derived* a relationship between a driver's speed and the density as perceived by him from an economic theory of speed choice. In section 5 we will show how this relationship can be used to derive an analogous aggregate relationship under stationary state conditions.

We now formulate two requirements that the optimal speed function should obey:

(i) *s(r)* should be increasing in traffic density *r*.
We expect that the optimal speed decreases if traffic density increases. This will be the case if $\partial c/\partial s$ is increasing in *r*. This can be interpreted as saying that it becomes more costly to increase a given speed if traffic density is higher, which seems reasonable to assume.

(ii) the free flow speed s^* is finite
The free flow speed s^* is defined to be the one that would be chosen if the density of traffic approaches 0:

$$s^* = \lim_{r \to 0} s(r) \qquad (7)$$

A specific case
An intuitively appealing specification for the cost function is:

$$c = \frac{vot}{s} + b(r)\, s^2 \qquad (8)$$

where *vot* and *b* are both positive.
This equation states that the costs are determined by travel time and safety considerations. Travel time cost is proportional to the inverse of speed and the cost associated with safety is assumed to be quadratic in speed. The parameter *vot* can be interpreted as the value of time; *b(r)* must be increasing in traffic density.
The relation between speed and density that follows from (5) is:

$$s(r) = \left(\frac{vot}{2b(r)} \right)^{1/3} \qquad (9)$$

It is easy to verify that this speed choice function satisfies requirements (i) and (ii) listed above if *b(r)* is positive for all nonnegative densities.
The speed-density relation $s(r)$ can be empirically observed. This relation can be used to identify the parameter $b(r)$ in (8) as can be seen by rewriting (9) as:

$$b(r) = \frac{1}{2} \frac{vot}{\left(s(r)\right)^3} \qquad (10)$$

Condition (a) is not satisfied in our car following version of the speed choice theory, see the discussion following equation (13) below. Condition (b) seems often satisfied in actual driving situations and will be ignored.

This means that the parameters of the cost function (8) are identified if we know the value of time a and the speed-density relation $s(r)$.
The cost function (8) generalizes of the conventional approach in traffic congestion analysis where travel cost are usually be assumed to be equal to the value of travel time. In this approach, only the first term on the right-hand-side of (8) is relevant. The difference between this traditional approach and the one adopted here can be clarified by substituting (10) into (8). This gives the minimum instantaneous travel cost at density r as:

$$c^{\min}(r) = \frac{3}{2}\frac{vot}{s(r)} \qquad (11)$$

This equation says that the appropriate cost equals 150% of the value of travel time.[6] The traditional approach therefore underestimates the cost of travel time (and congestion) by one third. Another interpretation is that equation (11) says that the appropriate value of travel time is 50% higher than the 'pure' value of time because of the accident risk involved in travelling. If the value of accident risk is incorporated into the value of travel time, the right-hand-side can still be interpreted as the value of travel, and the difference with the conventional approach appears less fundamental. The reason behind the ambiguity implied by the possibility of two different interpretations is that it is not always clear whether empirical estimates of travel time incorporate the accident risk associated with traveling or should be interpreted as a measure of the 'pure' value of time.

4 Speed choice and car following theory

This section is concerned with the question: can the car following theory be derived from the theory of speed choice discussed in the previous section?

Traffic density and the distance with the leading vehicle

The theory of section 2 takes traffic density as one of the determinants of speed choice. Density of traffic is usually measured as the ratio between the number of cars on a road segment and its length. However, it may be noticed that the relevant measure of density from the microscopic viewpoint adopted here is the one experienced by the driver. This suggests that traffic density should refer to a road segment of limited length in front of a driver or that cars that are close to a driver contribute more to traffic density as experienced by him than cars at a larger distance, that cars driving in front of him are more relevant than those driving behind, et cetera. In some circumstances the distance to the vehicle immediately in front of him solely determines a driver's experience of traffic density. This will be especially the case if the driver is unable to look further ahead. Also in other cases the distance to the driver immediately in front may be the major determinant of a driver's experience of traffic density as far as speed choice is concerned.
Moreover, we note that car following theory links the behavior of a driver only with that of his leader and this suggests that we should focus exclusively on the distance with the leader in order to integrate the theory of speed choice with that of car following. We may, for instance, specify:

[6] See Verhoef and Rouwendal (2001) for further discussion of this result and related issues concerning speed policy.

$$r = \frac{1}{\mu + \delta}. \qquad (12)$$

with μ the length of a car. More generally, we may postulate:

$$r = f(\delta) \qquad (12')$$

for some decreasing function *f*.
After substitution of (12') into the function *s(r)* we arrive at a relationship between speed and the distance between two subsequent cars, to be denoted as *s(δ)*. The implied relation between speed and distance to the leader is:

$$s = s(\delta) \qquad (13)$$

A possible objection to the specification of traffic density as perceived by an individual driver is that this driver is in full control of the distance to his leader. It may therefore be questioned whether it is appropriate to assume that a driver takes this distance as given when choosing his speed. The answer to this objection is that when a driver is at *x* and has a distance δ to his leader, he can only control the future value of this distance by means of choosing his present speed. In other words: speed choice is the instrument available to him to adapt the distance to the leader according to his wishes at locations downstream of *x*. When speed at location *x* has to be chosen, the distance to leader is indeed given (i.e. cannot be changed immediately).
If the driver is not (purely) myopic, he may take into account that his present speed choice will influence future values of the headway distance. For instance, instead of behaving purely myopically, the driver may choose a lower speed at the beginning of his trip and a higher speed at the end. The result may be that total travel time is equal to that under purely myopic behavior, while during most of the trip safety costs are lower. The net result may be a decrease in travel cost under the simplified conditions considered here. This behavior can only be successful if the follower can predict the behavior of his leader with considerable accuracy, which is usually not the case in actual situations. Moreover, overtaking cars or cars that enter between a car and its leader will in practice often disturb such behavior. For this reason it will not be considered here.
Equation (13) gives a relationship between speed and distance to the leader, whereas car following theory postulates a relationship between *changes* in speed and speed differences. We have therefore not yet reached the goal of deriving car following theory from the theory of speed choice. Nevertheless, we have taken an important step towards this goal by relating traffic density to the distance between a car and its leader.

Compatibility of the speed choice model and the GM car following model
We shall now clarify the relation between car following theory as discussed in section 2, and equation (13). In order to do so, we first note that:

$$\frac{d\delta}{dt} = s^0 - s \qquad (14)$$

This equation says that the headway changes if speeds differ. If we differentiate both sides of (13) with respect to time and substitute (14), we then find:

$$a = \frac{ds}{d\delta}\left[s^0 - s\right]. \qquad (15)$$

This is equal to the equation of the standard car following framework (3') with a sensitivity coefficient that equals $ds/d\delta$. Note that the interpretation of this derivative as a sensitivity coefficient makes perfect sense from the viewpoint of the theory of speed choice discussed in the previous section. The partial derivative $ds/d\delta$ is a measure of the sensitivity of the optimal speed to changes in the headway distance. Changes in the headway distance imply larger changes of speed if this partial derivative is larger. Delays in this reaction are costly in terms of travel cost and it seems therefore natural that the drivers react stronger to a given change in the headway distance if the derivative is larger. This implies that the sensitivity coefficient in the car following equation is larger as is the case if both variables are equal.
In order to demonstrate formally the compatibility of the speed choice theory outlined earlier with the standard car following theory discussed in section 2 we have to show that the derivative $ds/d\delta$ can indeed be equal to the sensitivity coefficient λ as specified in equation (2') which corresponds to the GM model that incorporates a number of other versions of car following theory as special cases. Substituting $ds/d\delta$ for λ in (10'), we have:

$$\frac{ds}{d\delta} = \lambda_0 \frac{s^m}{\delta^l} \qquad (16)$$

This differential equation is separable and can be rewritten as:

$$\frac{ds}{s^m} = \lambda_0 \frac{d\delta}{\delta^l} \qquad (16')$$

It is solved (integrating both sides of the equation separately) by the following functions:

$$s = \begin{cases} \left(\lambda_0 \dfrac{m-1}{l-1}\delta^{-(l-1)} + c\right)^{-\frac{1}{m-1}} & l \neq 1, m \neq 1 \\ \exp\left(\lambda_0 / \left((l-1)\delta^{l-1}\right) + c\right) & l \neq 1, m = 1 \\ \left(\lambda_0 (m-1)\ln(\delta) + c\right)^{-\frac{1}{m-1}} & l = 1, m \neq 1 \\ \delta^{\lambda_0} c,\ c > 0 & l = 1, m = 1 \end{cases} \qquad (17)$$

where c is an integration constant.
The solution is only meaningful if it satisfies the two requirements for a speed choice function listed in section 3. For the indicator of traffic density used here, these requirements imply that optimal speed should be increasing in the distance to the leader and approach a finite 'free flow' speed when this distance becomes infinitely

large. It is easy to see that the special cases for which *m* or *l* or both are equal to 1 will violate at least one of these requirements. We are therefore left with the generic case in which both parameters differ from 1. The first derivative can be computed as:

$$\frac{ds}{d\delta} = s\frac{\lambda_0 \delta^{-l}}{\lambda_0 \frac{m-1}{l-1}\delta^{-(l-1)} + c} \qquad (18)$$

We distinguish three cases.
(i) $m>1$, $l>1$
It is easily verified that first derivative is always positive if c>0. Moreover, in this case the derivative approaches the value 0 if δ gets large. For a large headway distance speed can be shown to approach a limit equal to $s^*=1/c^{1/(m-1)}$. We have therefore found one case in which the two requirements are fulfilled. If *c* is negative the derivative may be negative.
(ii) $m<1$, $l<1$
In this case the derivative is also always positive for *c* positive , but it does not approach the value 0 if the headway distance gets large. For $c<0$ the derivative may be negative.
(iii) $m<1$, $l>1$ or or $m>1$, $l<1$
The denominator on the right-hand-side of (18) has to be positive. This implies that the headway distance has to be smaller than some finite positive value if *c* is positive (if *c* is negative this is never the case). When approaching that value both speed and its first derivative increase without bound.
We have to conclude that only the first case satisfies our requirements when *c* is positive. It may be noted that empirical studies mentioned in section 2 have not always found values for *l* and *m* that are larger than 1. However, note also that the specification of the sensitivity parameters is less appealing to intuition with *l* or *m* smaller than 1.

Other approaches to car following
In the literature on car following other models than the GM model discussed in section 2 are sometimes proposed. Newell (1960) suggested an approach in which the speed-headway relation is taken as the primitive concept and studied the following specification (given in the present notation) in detail:

$$s(r) = s^* - s^* \exp\left(-\lambda_0 \delta / s^*\right) \qquad (19)$$

where s* is the maximum (free flow) speed. It is immediately clear that this car following theory also fits within the speed choice framework proposed here.
The same conclusion applies to the 'simple car following model' recently proposed by Verhoef (2001) in order to study the stability of equilibria in the standard static models of traffic congestion. Verhoef derives the speed-headway relation cf. equation (13) above from the 'fundamental diagram'. He refers to this relation as 'simple car-following theory' because classical car following theory as discussed in section 2 of the present paper starts from changes in speed. However, as Verhoef (2001) notes, equation (13) does not necessarily differ from the standard theory. For the purpose of numerical illustration Verhoef uses the following specification (adapted to the present notation):

$$s(\delta) = \begin{cases} 0 & \text{if } (\delta+\mu) \le 5 \\ 33\frac{1}{3} - \dfrac{33\frac{1}{3}}{(100-5)^5}(100-(\delta+\mu))^5 & \text{if } 5 < (\delta+\mu) \le 100 \\ 33\frac{1}{3} & \text{if } (\delta+\mu) > 100 \end{cases} \qquad (20)$$

This specifications (19) and (20) used by Newell and Verhoef cannot be derived formally from the GM model of car following theory, discussed in the previous sections. However, they can be approximated closely by the formula given in the first line of the right hand side of (17) with *m*, *l*>1. Indeed this appears to be true for any specification of simple car following theory, i.e. for any specification of a relationship between headway distance and speed that is derived from the fundamental diagram.

5 Implications for the static model of congestion tolling

In this section the implications of the integration of car following and speed choice theory for congestion tolling will be considered. In order to do this, we impose conditions that bring the model as close as possible to the static model of road congestion developed on the basis of the insights provided in the first half of the twentieth century by Pigou and Knight. This model is widely used in transportation economics as an appropriate analytical tool for studying congestion problems. It seems therefore of some interest to show how this standard economic model can be related to one of the standard models of traffic engineering, viz. car following theory. This is discussed in subsection *a*. In the rest of the section an extension of the model to situations in which traffic is heterogeneous is provided. Since the Pigou-Knight model assumes homogenous traffic, application this model to real world congestion problems presupposes that actual traffic heterogeneity that is present can be conveniently ignored. This is only the case if this heterogeneity does not have consequences for the value of the optimal toll. Arnott and Kraus (1999) have derived conditions under which uniform congestion tolls are sufficient for reaching first-best optimality and it would, of course, be a nice result if it could be formally shown that these conditions are relevant for the extension of the Pigou-Knight model to heterogeneous traffic. This question is considered in detail below.

a) Homogeneous traffic

One implication of the synthesis between car following theory and the economic theory of speed choice outlined above is that it gives us a somewhat different look at the standard analysis of congestion in the economic literature. The major difference is that the trip cost function becomes an endogenous part of the model.

In order to facilitate comparison with the Pigou-Knight type of analysis, we now consider the implications of this synthesis for congestion analysis in a stationary state. In order to do so, we concentrate on a homogeneous road segment in a stationary state. A stationary state is here defined by the following properties: (i) every *t* seconds a car enters the road and (ii) each car has a constant speed *s* on the whole road segment. Since the headway distance determines speed, it must also be constant. Moreover, speed must be equal for all drivers, since speed differences would imply changing headway distances and therefore non-constant speeds.

The stationary state is compatible with the model of individual driver behavior developed above if the headway distance plus the length of the car (again denoted as

μ) is equal to *st*. This means that the stationary state is characterized by equation (13) and:

$$\delta + \mu = st \tag{21}$$

In the above we have referred to speed, cost *et cetera* as instantaneous units referring to a particular location. In order to derive the cost of a trip on the road segment we should integrate over the total length of the road segment. Because the segment is (by assumption) homogeneous, this is easy to do: all that is needed is multiplication of the localized variables by the length of the road segment. For convenience we take the road segment to be of unit length, which implies that we can use our instantaneous variables without further modification as referring to trips as well.
The flow of traffic on the road segment will be denoted as *f* and equals the inverse of *t*:

$$f = \frac{1}{t} \tag{22}$$

Using equation (21), we find:

$$s = (\mu + \lambda)\ f \tag{23}$$

Figure 1 illustrates. We have reversed the axes of the figure in order to make it comparable with Figure 3 below. The straight line in the figure is equation (23). It has slope *f*. The other line pictures the headway-speed relationship, which is here assumed to be concave.[7] The two lines cross each other at two points if the flow is not too large. The equilibrium with the lowest speed corresponds to a state of hypercongestion and has been shown by Verhoef (2001) to be dynamically unstable as a stationary state. Attention may therefore be concentrated on the equilibrium with the highest speed.
It can be readily inferred from the picture (and confirmed by formal analysis) that larger flows can only be accommodated at lower speeds. There is a unique maximum flow that corresponds with the situation in which the straight line (23) just touches the headway-speed relation (13). This maximum flow can be interpreted as the capacity of the road. This derivation makes clear that capacity is not only determined by the technological characteristics of the road, but also by driver behavior.[8]
Equations (23) and (13) determine the speed-flow combinations that are feasible on the road. Since travel time is the inverse of speed, the relation between flow and travel time can easily be derived from this relation. Using the cost function (8) the travel cost in a stationary state can be described as a function of the traffic flow by means of (11). This shows that the relation between generalized travel cost and flow is also derived endogenously in the present model. As noted above, the difference between the standard approach is (which equalizes time cost and travel cost) is that travel costs are 50% higher than the 'pure' time costs, due to the effect of safety costs, when cost function (8) is used.

[7] I.e. speed is a concave function of headway distance. Figure 1 pictures the inverse function, which is convex.
[8] We will return to this point in the next section.

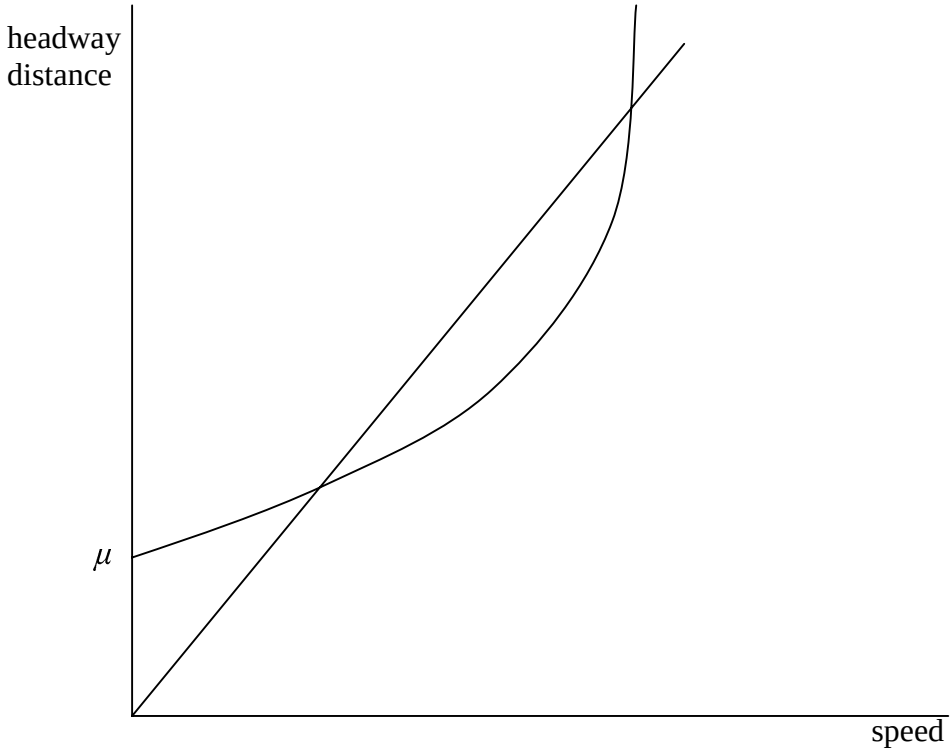

Figure 1 Speed and flow in a stationary state

Once the derivation of the travel cost function is completed, the remaining analysis becomes fairly standard.[9] The demand for trips is given by an inverse demand function and the price of a trip is given by the sum of the travel cost and a possible toll. Maximization of consumer's surplus under the condition that a user equilibrium should obtain allows one to obtain the optimal value of the toll.

b) Steady state traffic with two types of vehicles

Although in economic analyses drivers are often assumed to be homogeneous, it seems unlikely that in reality all drivers have the same values of the parameters of their utility functions. Note that in the model developed in the previous sections these parameters also include those of the car-following relationship. Verhoef et al. (1999) and Rouwendal et al.(2002) have studied the consequences of traffic heterogeneity in models where there are two groups of vehicles that differ in desired free flow speeds.[10] In these papers a simplified relation between speed and headway distance was assumed: drivers always used their desired free flow speed unless the minimum critical distance to the leader was reached and speed equals that of the leader if the headway distance equals its minimum value. In what follows we use the more realistic headway-speed relation based on car following theory derived earlier in this paper and study the stationary state properties of the model.[11]

[9] A formal exposition seems unnecessary. For those who want it, it can be remarked that the appropriate model is a special case (with homogeneous drivers) of the model with to groups of drivers discussed below.

[10] These differences may be caused by the preferences of the drivers or by the characteristics of the vehicles they use or by a combination of the two.

[11] The discussion in the present subsection assumes that the chosen speed will always be increasing in the headway distance, while approaching it s free flow value asymptotically. It is easy to consider also situations in which free flow speed is reached for a finite headway distance, as in eq. 20.

Assume that there are two groups of vehicles with different headway-speed relationships. Figure 2 illustrates this for a case in which for any value of the headway distance, drivers of group 1 choose a lower speed than drivers of group 2, but situations in which the two curves cross each other are also admitted. Both groups use the same road and overtaking is impossible.
We want to use the model with two groups in order to study stationary state traffic with heterogeneous drivers. In order to define a stationary state in the present situation, we have to revise our definition of such a state. The reason is that we can't have a constant time interval *t* between subsequent cars that enter the road *and* the same constant speed for all cars. The stationary state will therefore now be defined as referring to a situation in which all cars have identical constant speed. The headway distance will therefore differ. This means also that the time interval between cars will differ and let t_i refer to the time interval before entrance of a car of type *i* (*i*=1,2). Vehicles of one type drive closer to their leaders than those of the other type. If the headway distance of the two types at the stationary state speed is substantial,

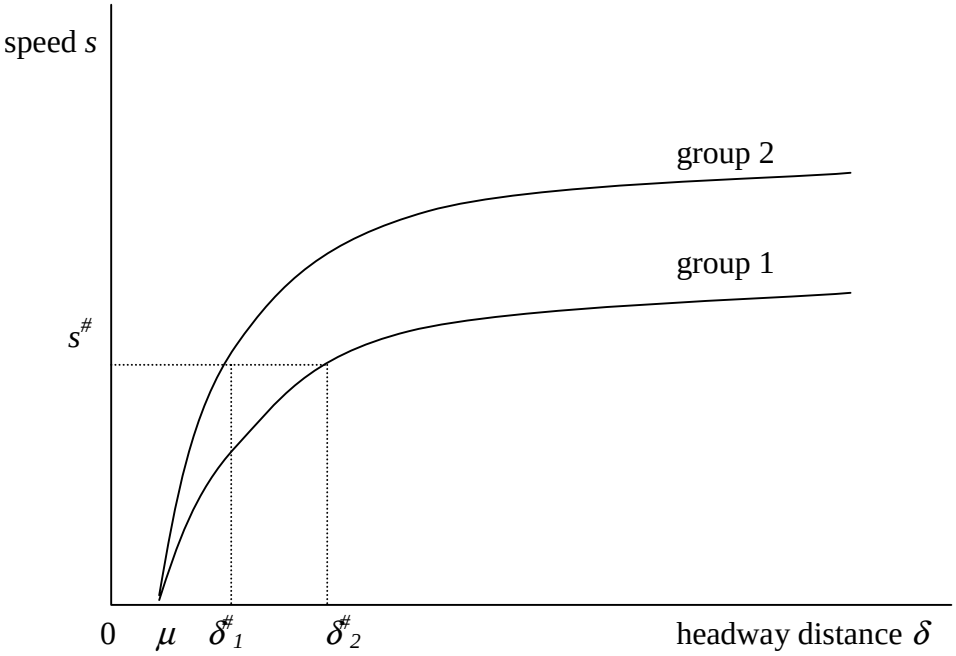

Figure 2 Headway-speed relationships of two groups

stationary state traffic will take the form of single driving vehicles with a large (in a relative sense) headway distance and platoons consisting of a vehicle with a large headway distance followed by one or more vehicles with a shorter headway distance. In order to complete the description of the stationary state, we have to determine the length and frequency of the platoons. In order to do so, we have to specify the mechanism by which cars enter the road.
This mechanism should be compatible with a stationary state. This means that if the the headway distance should be that corresponding to the stationary state speed. Since the cars are of different types this means that for one type this headway distance will be larger than for the other. Since the speed of both types of vehicles is equal, this means that the time that passes before a vehicle of type 1 enters must be different from the time that enters before a vehicle of type 2 enters. We assume that there are

fixed probabilities p_i that the next car will be of type I=1,2. If the vehicles belonging to the first group have the shortest headway distance, there is a probability p_2=(1-p_1) that a vehicle of type 2 that has just entered the road drives single and a probability p_1^n(1-p_1) of a platoon of length n.[12]
Let us now consider the stationary state. In such a state all vehicles have the same speed, but the composition of the flow is determined by the stochastic mechanism just described. It implies that the total number of vehicles of each type that enter during a unit of time is also a random variable, as is the total flow.[13] Let f_1 be the expected flow of vehicles of group 1 and f_2 the expected flow of vehicles of group 2. For given values of these flows the equilibrium speed and headway distances can be determined by the following three relations:

$$(f_1 + f_2)\left(\frac{f_1}{f_1 + f_2}(\delta_1 + \mu_1) + \frac{f_2}{f_1 + f_2}(\delta_2 + \mu_2)\right) = s \qquad (24)$$

$$s_1(\delta_1) = s \qquad (25)$$

$$s_2(\delta_2) = s \qquad (26)$$

The first of these equations (24) says that the product of the total flow and the average road distance occupied by a car should be equal to the common steady state speed s. We can simplify (24) and substitute the inverses of the headway speed relationships (25) and (26) into that equation:

$$(f_1\mu_1 + f_2\mu_2) + \left(f_1 s_1^{-1}(s) + f_2 s_2^{-1}(s)\right) = s \qquad (27)$$

The function s_1^{-1} gives the headway distance for vehicles of type 1 that corresponds with speed s, and s_2^{-1} has an analogous interpretation. These two inverse functions are defined for nonnegative speeds; they are convex, equal to 0 when s=0 and increasing in s. Figure 3 illustrates the steady state equilibria by picturing the left- and right-hand-sides of (27) as separate lines. The figure shows that there may be two equilibria, just as in the case with homogeneous traffic. One of these can be considered as hypercongested. The dynamic stability of these equilibria is studied in the appendix. It is shown there that the hypercongested stationary state is dynamically unstable. Attention can therefore be confined to the non-hypercongested case.
It may be noted that in the case of heterogeneous traffic, it is impossible to identify the capacity of the road as a unique number of vehicles that may pass through it per unit of time. There may be various combinations of flows f_1 and f_2 for which the line picturing the left-hand-side of (27) just touches that picturing the right-hand-side. For any given value of f_1 that does not exceed the capacity of the road for homogeneous traffic of type 1, one may derive flow f_2 at which the capacity of the road at which

[12] Other mechanisms may also be defined. Deterministic ones (e.g. k_1 cars of type 1 are always followed by k_2 cars of type 2, k_1,k_2>0) and mechanisms in which the probability that the next car is of type 1 depends on the type of its leader are alternative possibilities. The mechanism used here can be interpreted as resulting from a heterogeneous population of drivers who each take their decision to enter the road independently of each other.
[13] Note that the flows of the two types of vehicles are dependent upon each other: if one knows the flow of one type, the flow of the other is also determined. However, the total number of vehicles is *not* a deterministic variable.

capacity is reached. Except in special cases, the total number of vehicles will be different depending on the chosen value for f_1.
The speed of the non-hypercongested steady state speed will in general also depend on the composition of traffic and we may write:

$$s = s(f_1, f_2) \qquad (28')$$

for this speed. If cost function (8) is used, generalized travel cost is inversely proportional to this steady state speed and therefore also dependent on the composition of the total traffic flow.
It should be noted that (28) implies that there is in general no unique relation between steady state speed and total traffic flow f_1+f_2. The exception occurs if:

$$s = s(f_1 + f_2), \qquad (29)$$

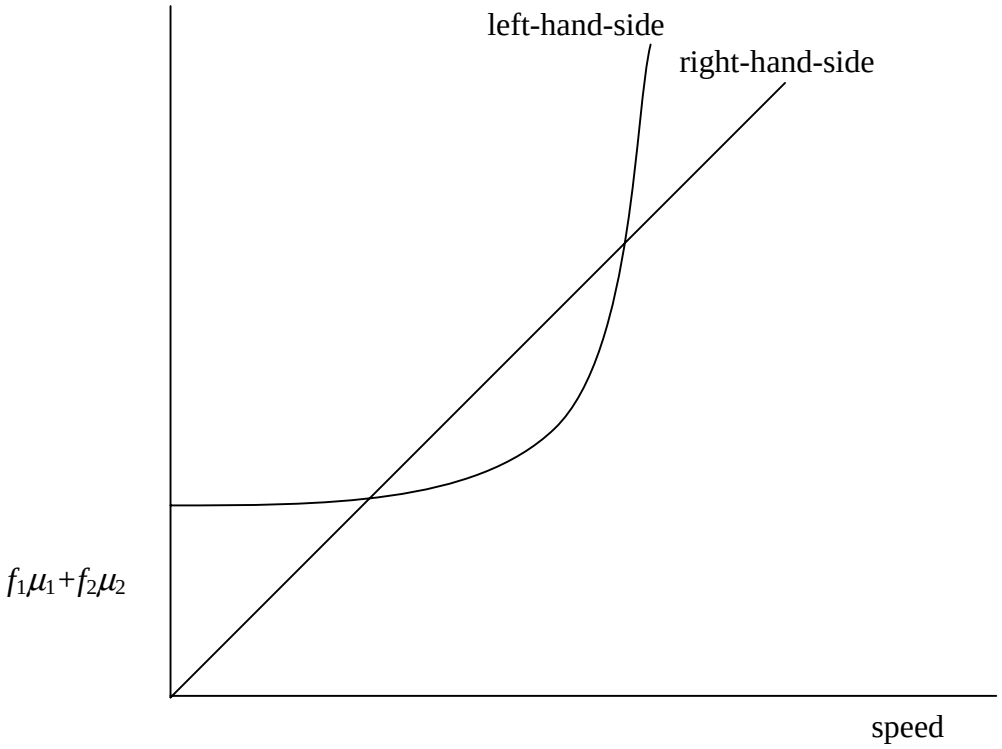

Figure 3 Steady state equilibrium with two groups of vehicles

but it will be shown below that this specification cannot be valid in the model developed here.
Finally, we note that it can be shown that steady state speed is a decreasing function of f_1 and f_2.[14]

[14] See eq. 34 below.

c) Optimal congestion tolling with heterogeneous traffic

The standard Pigou-Knight model of static road congestion assumes homogeneous traffic, but is often tacitly considered being a good approximation to actual situation with heterogeneous traffic. The steady-state model with heterogeneous traffic developed in the previous subsection allows a formal analysis of this conjecture. We maximize the social surplus, i.e. the sum of the consumer's surpluses of the drivers of the two types and the toll revenues under the constraints that a user's equilibrium should be realized. We denote the inverse demand function as *D*, the price of a trip (which equals the sum of travel cost and toll) as *p* and the toll as τ and add the appropriate suffixes. The Lagrangian is:

$$L = \int_0^{f_1} D_1(\varphi_1)d\varphi_1 + \int_0^{f_2} D_2(\varphi_2)d\varphi_2 - c_1 f_1 - c_2 f_2 + \eta_1(p_1 - c_1 - \tau_1) + \eta_2(p_2 - c_2 - \tau_2) \tag{30}$$

with the η's denoting Lagrange multipliers. The first order conditions lead to the following expressions for the optimal tolls:

$$\begin{aligned} \tau_1 &= f_1 \frac{\partial c_1}{\partial f_1} + f_2 \frac{\partial c_2}{\partial f_1} \\ \tau_2 &= f_1 \frac{\partial c_1}{\partial f_2} + f_2 \frac{\partial c_2}{\partial f_2} \end{aligned} \tag{31}$$

This shows that the optimal tolls for the two groups are different unless $\partial c_i/\partial f_1 = \partial c_i/\partial f_2$ for i=1,2. In order to see whether this is the case, we use cost function (8), which leads to the following expression for the optimal cost (cf. eq. 11):

$$c_i = \frac{3}{2}\frac{a_i}{s(f_1, f_2)} \quad i = 1,2. \tag{32}$$

Differentiation shows that:

$$\frac{\partial c_i}{\partial f_j} = -\frac{3}{2}\frac{a_i}{s^2}\frac{\partial s}{\partial f_j} \quad i, j = 1,2.. \tag{33}$$

In order to find the partial derivative of the steady state speed with respect to each flow we fully differentiate (27) and compute the ratio of the two differentials. The result is:

$$\frac{\partial s}{\partial f_j} = -\frac{\mu_j + \delta_j}{1 - f_1 \frac{1}{ds_1/d\delta_1} - f_2 \frac{1}{ds_2/d\delta_2}} \tag{34}$$

The numerator of the right-hand-side is the distance occupied by a vehicle of type *j* in the steady state. It is different for the two types of vehicles and this shows that for the model developed here the steady state relation between flows and speed cannot be

described by (29). Hence there is no unique relation between speed and total traffic flow in the model with heterogeneous drivers.
Returning now to the optimal tolls, we observe, using (33), that the fact that $\partial s/\partial f_j$ depends on j implies that $\partial c_i/\partial f_1 \neq \partial c_i/\partial f_2$. This means that the optimal tolls for the two types are in general different (cf. eq. 31). In the model with heterogeneous drivers developed in this paper, first best tolling requires different treatment of the two types. Uniform tolling is a second best solution.
The model with two types of vehicles developed in this and the previous subsection generalizes the Pigou-Knight model to situations in which traffic is heterogeneous. One important reason for developing this model was to investigate the validity of the common practice to use the Pigou Knight approach as an approximation to situations with heterogeneous traffic. We conclude that in one important respect the approximation is invalid: with heterogeneous traffic uniform tolling can in general only be a second best measure.

c) Discussion
It is easy to generalize the model of the previous two subsections to situations in which there is an arbitrarily large number of vehicle types. Such a generalization seems to be needed in order to capture the diversity of driving styles and vehicles characteristics that interfere with them in actual situations. Newell (2002) has recently stressed the empirical relevance of steady state traffic as would be described by such a model. He proposed a simplified car following theory "wherein, if an n-th vehicle is following an (n-1)th vehicle on a homogeneous highway, the time-space trajectory of the n-th vehicle is essentially the same as the (n-1)th vehicle except for a translation in space and time." (p. 195, abstract), but where the headway distances between the vehicles may be different. After exposing this (essentially steady state) theory he motivates its empirical relevance by a number of references.
The conclusion reached above that optimal congestion tolls should vary with vehicle types implies that the Arnott-Kraus (1999) condition for the feasibility of marginal cost pricing by means of uniform tolling are violated in the model developed above. The reason is that the different types of vehicles contribute to congestion in different ways. The essential difference is the amount of road space $(\mu_j + \delta_j)$they occupy. It should be noted that the amount of space in-between vehicles, δ_j, will probably vary with driving conditions and for this reason a differentiation of the tolls on the basis of observable characteristics (such as vehicle length μ_j) will not be able to solve the problem. Under severe congestion it may be the case that headway distances are more or less equal for all types of vehicles, so that differences in space occupied are almost completely determined by the length of the vehicles, which would justify a passenger car equivalent rule based on observable characteristics. However, such a rule is unlikely to be valid in general.

6 Bottleneck model

The leading economic model that does not refer to a stationary state is the bottleneck model developed by Vickrey (1969) and studied extensively by Arnott-de Palma-Lindsey (henceforth AdPL, see for instance there 1990 article). In a recent paper Verhoef (2002) uses simple car following theory (his terminology refers to the headway-speed relation given in (20)) to construct an alternative version of the bottleneck model. The paper demonstrates some striking results and it is the purpose of the present section to shed some light on one of its substantial findings, viz. that the

tolling scheme proposed by AdPL may result in a welfare *loss* in the car-following version of the model.
In this section a generalization of Verhoef's (2002) model is developed that becomes similar to a Vickrey/AdPL model when a parameter approaches a critical value. We consider a population of *N* identical drivers who have to move from home to work using the same road. The road starts with two lanes, but there is a bottleneck in which only one lane is available. The two-lane road in front of the bottleneck is of sufficient length to allow for a queue to develop. Driver behavior is given by the following generalization of (20):

$$s(\delta)=\begin{cases}0 & \text{if } (\delta+\mu)\le 5\\ 33\frac{1}{3}-\dfrac{33\frac{1}{3}}{(\delta^*-5)^5}(\delta^*-(\delta+\mu))^5 & \text{if } 5<(\delta+\mu)\le\delta^*\\ 33\frac{1}{3} & \text{if } (\delta+\mu)>\delta^*\end{cases} \qquad (35)$$

for some $\delta^*>5$.

It is easy to verify that $s(\delta)$ is smooth and continuous in δ if $\delta>5-\mu$, in particular around the value $\delta^*-\mu$. If δ^* is close to (but larger than) 5, drivers choose free flow speed unless distance to their leader becomes very small, and in that case they have the same speed as their leader.[15] It will be shown below that in this limiting case traffic behaves as in the AdPL bottleneck model. For δ^* equal to 100 the model is identical to that of Verhoef (2002).
If we use the speed choice theory discussed earlier in this paper with cost function (8) the travel cost of driver *n* is:

$$ct(n)=\int_0^X \alpha\frac{1}{s_x(n)}dx \qquad (36)$$

with α equal to 1.5 times the value of time.
Scheduling cost is given by the usual piecewise linear function:

$$cs(n)=\beta\min\{0,t^*-\tau(n)\}+\gamma\min\{0,\tau(n)-t^*\} \qquad (37)$$

where $\tau(n)$ denotes the arrival time of the n-th driver. The total travel cost of the n-th driver, C(n) are therefore equal to:

$$\begin{aligned}C(n)&=ct(n)+cs(n)\\&=\int_0^X \alpha\frac{1}{s_x(n)}dx+\beta\min\{0,t^*-\tau(n)\}+\gamma\min\{0,\tau(n)-t^*\}\end{aligned} \qquad (38)$$

Total trip cost is equal to the sum of the individual travel costs:

[15] If $\delta^*=5+\varepsilon$, and ε is close to zero small differences between the speed of a vehicle and its leader will soon lead to a speed equal to 0 (if speed exceeds that of the leader) or free flow speed (if speed is lower than that of the leader).

$$CT = \sum_{n=1}^{N} C(n) = \sum_{n=1}^{N} ct(n) + \sum_{n=1}^{N} cs(n) \tag{39}$$

In a user equilibrium all drivers have equal trip cost. Equal trip cost for all drivers can only be realized by compensating the inevitable differences in scheduling delay cost by differences in travel cost with the same magnitude, but opposite sign.
We now show that the present model, which is identical to Verhoef's model for if $\delta^*=100$, has the properties that are very similar to those of AdPL's bottleneck model if $\delta^* \downarrow 5$ (see e.g. Arnott, de Palma and Lindsey, 1990 for comparison). In this situation every driver chooses the free flow speed unless if $\delta^*=5$, and then his speed is equal to that of his leader. With this behavior, capacity of the bottleneck is equal to the product of the minimum headway distance at the free flow speed $cap=\delta^* s^*$. If the number of cars that approach the bottleneck during one unit of time exceeds capacity, a queue develops in front of the bottleneck.
All drivers want to arrive as close as possible to t^* and it is therefore efficient to use the full capacity during a time interval containing t^* and chosen so as to minimize total scheduling delay cost. Travel time is equal to the free flow travel time plus waiting time. The latter is the additional travel time caused by the lack of sufficient capacity to let all traffic flow at free flow speed. In the limiting version of the model studied now, this additional travel time is spent in a queue with distance to the leader equal to the minimum headway distance. If $t(n)$ is the departure time of the n-th driver, queuing time is equal to:

$$ct(n) = \alpha\left(ct^* + \left(t(n) - t(1) - \frac{n}{cap} \right) \right) \tag{40}$$

Travel time equals free flow travel time (while capacity is still completely used if $t(n)$-$t(1)$ equals n/cap for all n.
This discussion of the special case of the general model that occurs when $\delta^* \downarrow 5$ shows that the model has characteristics that are similar to Vickrey/AdPL's bottleneck model. There also remains a difference, as Vickrey/AdPL assume that there is no time needed to travel the distance from home to the bottleneck or the tail of the queue. This implies a simplification of the model, which may be interpreted as 'vertical queuing' of cars.[16] It allows one to abstract for driver behavior between home and the tail of the queue and from 'flow congestion' on that part of the trip. Indeed, in important element of Verhoef (2002) is the introduction of this element into the analysis of bottleneck congestion. However, in the limiting case of car following theory considered now this behavior becomes very simple: vehicles then use either free flow speed (on the first uncongested part of the trip) or the speed of their leader (after

[16] Arnott, de Palma and Lindsey (1990) state 'an individual arrives at the bottleneck as soon as he leaves home,' but make clear that they only introduce this unrealistic element in their analysis because they can simplify by putting the fixed component of travel time (corresponding to the free flow travel time in the setting of the present paper) at an arbitrary value. For convenience, they choose this value to be equal to 0, but nothing that is essential would change in their analysis if they had adopted a different (positive) value.

entering the queue).[17] The speed in the queue is constant and determined by the capacity of the bottleneck. The delay caused by the presence of the queue is equal to the length of the queue times the difference between free flow speed and speed in the queue. The number of cars in the queue is equal to the length of the queue divided by the amount of road space used by each car at the speed prevailing in the queue. This means that just as in the Vickrey/AdPL model there will be linear relationships between departure rates, queue length, travel time and clock time in user equilibrium. Since utility is here the same piecewise linear function as in Vickrey/AdPL, the user equilibrium of the model considered here has similar characteristics as theirs. It follows also that a time varying toll that is equal to the monetary value of additional travel time caused by the presence of a queue in the user equilibrium will result in a social optimum.

This comparison leads us to the conclusion that the differences between Verhoef's model and Vickrey/AdPL's model are apparently due to a different value of the parameter δ^*. So what changes in the model when δ^* is larger than 5? The most important difference seems to be that the capacity of the bottleneck is no longer completely used when traffic proceeds at free flow speed. In order to derive the maximum flow of traffic through the bottleneck we return to the model of the previous section with homogeneous traffic. It can be inferred from Figure 1 and the accompanying discussion that the maximum flow is reached when the straight line in Figure 1 touches the bended line, but does not cross it. This gives us two equations. The first says that the combination of speed and headway distance should be on both curves in Figure 1:

$$(\mu+\delta)f = s(\delta) \tag{41}$$

and the second that the slopes of the two curves should be equal:

$$f = \frac{ds(\delta)}{d\delta} \tag{42}$$

Since $ds/d\delta=0$ in the present model whenever free flow speed is used, the second condition immediately shows that capacity corresponds to a situation in which speed is lower than its free flow value. For the present model the relationship between headway-distance and speed is given by (35) and condition (42) becomes:

$$f = \frac{33\,1/3}{(\delta^*-5)^5}5(\delta^*-(\delta+\mu))^4 \tag{43}$$

If $\delta^* \downarrow 5$, the left hand side becomes infinitely large unless $(\delta+\mu)\rightarrow\delta^*$, which implies that for this case the Vickrey/AdPL situation obtains, as expected. However, except for this limiting situation, the present model with car following behavior of the drivers has the possibility that total trip cost can be reduced by letting traffic proceed through the bottleneck at a speed lower than its free flow value. A toll that removes all congestion and enables all drivers to use their free flow speed can therefore in

[17] This behavior is identical to that supposed in the studies of speed differences by Verhoef et al (1999) and Rouwendal et al (2002).

general not be expected to achieve full optimality. Indeed, Verhoef's surprising result is that such a toll will actually result in an *increase* of the total trip cost *CT*.
In Verhoef's model the flow through the bottleneck is maximal when there is a queue in front of it. It is therefore clear that in his model removing all congestion implies higher scheduling costs. In other words: the presence of some congestion has the beneficial effect of inducing drivers to use the full capacity of road infrastructure and is therefore not a pure 'bad' as it is in the Vickrey/AdPL model. In his approximate social optimum there is indeed a significant amount of queuing.
Table 1 shows what happens in the bottleneck if the parameter δ^* approaches the value 5 from above. The first column indicates the minimum distance needed to induce vehicles to drive at the free flow speed of 33,33 meters per second, which is equal to δ^*. The second column indicates the distance between cars ($\delta+\mu$, i.e. length of the car is included) at which the flow through the bottleneck is maximal. The next two columns indicate the free flow speed and the speed at which the flow is maximal. Column 5 indicates the flow that results when all cars use the free flow speed and the minimum headway distance allowing them to do so. Column 6 shows the maximum possible flow. The last column indicates the ratio between the maximum flow at free flow speed (column 5) and the maximum possible flow at any speed (column 6).

Table 1 Traffic characteristics at various values of δ^*.

1	2	3	4	5	6	7
Minimum free flow headway distance (m)	Headway distance at maximum flow (m)	Free flow speed (km/hr)	Speed at maximum flow (km/hr)	Maximum flow with free flow speed (veh/hr)	Maximum flow at any speed (veh/hr)	Ratio between maximum flows
100	18,19	120	63,18	1200	3472	0.346
50	13,50	120	77,89	2400	5768	0.416
25	10,14	120	92,78	4800	9155	0.524
12.5	7,66	120	106,52	9600	13912	0.690
6,25	5,68	120	117,64	19200	20710	0.927
5,50	5,32	120	119,17	21818	22421	0.973
5,25	5,17	120	119,64	22857	23133	0.988
5,10	5,07	120	119,88	23529	23622	0.996

Legend. Column 1 gives δ*, column 2 the value of (δ+μ) at the maximum flow, column 3 gives free flow speed (33,33 m/sec) in kilometres per hour, column 4 the speed at the maximum flow, columns 5 and 6 give the ratio between speed and headway distance and column 7 the ratio between the two flows.

The table shows that at the parameter values chosen in Verhoef (2002) the maximum flow through the bottleneck at free flow speed is approximately one third of the 'global' maximum at any speed, which can be interpreted as the capacity of the bottleneck. Scheduling costs will therefore be much higher (approximately three times higher) when a toll is set so as to let traffic proceed at free flow speed. When δ^* has lower values, the difference between the maximum flow that is possible under free flow speed and the capacity of the bottleneck becomes smaller. When δ^* approaches its minimum value 5 the difference between the two flows becomes negligible, as expected. In such circumstances a toll that removes all congestion is compatible with the use of the full capacity of the bottleneck. The flows that are computed for the lowest values of δ^* are of course not realistic (it would be very dangerous to drive120

km/hr in a platoon with a – nose-to-nose - headway distance between the cars of just above 5 meters). However, qualitatively similar results are obtained for higher values of the parameter representing the minimum acceptable headway distance, which now has the value 5.
The essential difference between Verhoef's (2002) and Vickrey/AdPL's models is that the capacity of the bottleneck is not always completely used in the former. What happens before traffic passes the bottleneck is essentially irrelevant as long as its capacity is completely used. If this is the case, scheduling delay cost cannot be minimized further and a social optimum can be realized by 'translating' the delays of the user equilibrium (which is wasted time) into monetary revenues by tolling. In Verhoef's model capacity can only be used completely if there is some congestion present in front of the bottleneck. That makes it more difficult to find the social optimum: the simple rule according to which a time varying toll should be used that just eliminates all congestion in front of the bottleneck is no longer valid.
A new trade-off emerges: eliminating all congestion increases scheduling costs. The Vickrey/AdPL bottleneck model lacks this feature. The analysis of the present section suggests that the reason is that it makes implicit assumptions about driver behavior which may be not as unrealistic as 'vertical queuing' but which are less appealing to intuition than those implied by (simple) car following theory. It would, of course, be desirable to assess the empirical relevance of the alternative assumptions.

7 Conclusion

The main findings of the present paper may be summarized as follows:
1 The GM-model for car-following is, for a subset of the possible parameter values, consistent with an economic model of speed choice in which drivers trade off the shorter travel time against the higher risk associated with a higher speed.
2 Other approaches to car following that take the headway-speed relationship as a primitive are also consistent with this theory.
3 In a steady state with homogeneous traffic this model provides a generalization of the conventional Pigou-Knight analysis under a specific choice of the generalized travel cost function. In contrast with the conventional model, the integration between speed choice and car following behavior implies that the speed flow relation is endogenous. Moreover, the value of travel time should incorporate the cost of accident risk.
4 The model can be generalized to situations with heterogeneous traffic. In this model a uniform congestion toll is in general only second best optimal. First best optimality requires differentiation of the toll on the basis of the road space occupied by the vehicles, which equals the sum of the vehicle length and the 'nose-to-tail' distance with the leader. The latter is determined by the speed-choice/car-following behavior. The result implies that groups with different headway-speed relationships should in general be tolled differently in order to achieve first best optimality. This result was formally derived for a model with two groups, but can be generalized to a model with an arbitrary number of groups.
5 The speed choice – car following model can be used to develop a bottleneck model that encompasses Vickrey/AdPL's and Verhoef's (2002) models as special cases. This model shows that an important difference between the two is that in the latter model scheduling cost can be minimized by allowing some congestion. The reason is that with car-following behavior the full capacity of the bottleneck can only be used when speed is lower than its free flow value, whereas in Vickrey/AdPL's model (which is a limiting case) full capacity is reached with free flow speed.

These results illustrate the potential fruitfulness of integrating elements of the transportation economics and traffic engineering literatures. Other work along these lines may be fruitful. For instance, Newell (1961), argues that the car following theory discussed in that paper (which takes the headway-speed relationship as a primitive) is compatible with the Lighthill-Whitham-Richards theory of shock waves, which is the main element of modern traffic flow theory (see e.g. Daganzo, 1997).

The exposition in the previous section assumes that the relevant measure of traffic density is (the inverse of) the space between to subsequent vehicles. This is of course a simplifying assumption that is useful for the purposes of the present paper, but may not be realistic in all circumstances. It may therefore be noticed that there have been attempts to generalize car-following theory to situations in which driver behavior (notably changes in speed) are related to the speeds of and distances between a number of vehicles (see Bexelius, 1968).

The restriction of the analysis to stationary states in section 5 was convenient for the purpose of analyzing the relationship with the Pigou-Knight model. However, the car following speed choice model is perfectly able to deal with other situations, for instance those in which a road is not homogeneous, for instance because of a bend in the road that induces vehicles to slow down locally.

A potentially fruitful generalization that seems somewhat more difficult to analyze concerns situations in which interaction between drivers in situations that cannot be described as car following. Examples are the decision to overtake, and the decision to let a vehicle enter in from of your car. Both situations effectively refer to the choice of an alternative leader. This means that here we leave the realm of pure car following theory and reach an area that has been relatively unexplored also by traffic flow theorists.

References

Ahmed, K.I. (1999) Modelling Driver's Acceleration and Lane Changing Behavior, PhD Thesis, Department of Civil and Environmental Engineering, MIT, Cambridge (Ma).

Arnott, R., A. de Palma and R. Lindsey (1990) Economics of a Bottleneck *Journal of Urban Economics* **27** 111-130.

Arnott, R. and M. Kraus (1999) When Are Uniform Congestion Charges Consistent with Marginal Cost Pricing? *Journal of Public Economics* **67** 45-67.

Bexelius, S. (1968) An Extended Model of Car-Following *Transportation Research* **2** 13-21.

Chandler, R.E., R. Herman and E.W. Montroll (1958) Traffic Dynamics: Studies in Car Following *Operations Research* **6** 165-184.

Edie (1961) Car Following and Steady State Theory for Noncongested Traffic *Operations Research* **9** 66- .

Daganzo, C.F. (1997) *Fundamentals of Transportation and Traffic Operations* Pergamon

Gazis, R.E., R. Herman and R.B. Potts (1959) Car-Following Theory of Steady-State Flow *Operations Research* **7** 499-505.

Gazis, R.E., R. Herman and R.W. Rothery (1961) Non-linear Follow-the-Leader Models of Traffic Flow *Operations Research* **9** 545-567.

May, A.D. (1990) *Traffic Flow Fundamentals* Prentice Hall, Englewood Cliffs (N.J.)

May and Keller (1967) Non-Integer Car-Following Models *Highway Research Record* **199** 32- .

Newell (1961) Nonlinear Effects in the Dynamics of Car Following *Operations Research* **9** 209-229.

Newell, G.F. (2002) A Simplified Car-Following Theory: A Lower Order Model, *Transportation Research B* **36** 195-205.

Ozaki (1993) Reaction and Anticipation in the Car-FollowingBehavior, pp. 349-366 in: C.F. Daganzo (ed.) *Proceedings of the 12thInternational Symposiumon the Theory of Traffic Flow and Transportation* Elsevier, New York.

Rouwendal, J., E. Verhoef, P. Rietveld and B. Zwart (2002) A Stochastic Model of Speed Differences, forthcoming in: *Journal of Transport Economics and Policy*

Subramanian (1996) Estimation of Car Following Models, Master's Thesis, Department of Civil and Environmental Engineering, MIT, Cambridge (Ma).

Verhoef, E.T. (2001) An Integrated Dynamic Model of Road Traffic Congestion Based on Simple Car-Following Theory: Exploring Hypercongestion *Journal of Urban Economics* **49** 505-542.

Verhoef E.T. (2002) Inside the Queue *Journal of Urban Economics* (forthcoming)

Verhoef and Rouwendal (2001) A Structural Model of Traffic Congestion: Endogenizing Speed Choice, Traffic Safety and Time Losses, Tinbergen Institute Discussion Paper 026/3.

Verhoef, E., P. Rietveld and J. Rouwendal (1999) Speed Differences *Journal of Urban Economics* **45** 533-556.

Zhang, X. and D.F. Jarret (1997) Stability Analysis of the Classical Car-Following Model *Transportation Research B* **31** 441-462.

Appendix

Dynamic Instability of Stationary States with Hypercongestion

The car following equation for the *n*-th car that is our starting point is:

$$s_t(n) = g_n(\delta_t(n))$$

with *g* an increasing function that approaches a finite free flow speed when the headway distance becomes large. The suffix *t* indicates time. The suffix *n* attached to the function *g* indicates that not all cars need to have the same car following relationship. The derivations that follow are consistent with an arbitrary number of groups, where each car that enters has a fixed probability of belonging to each of these groups. The main text of this paper assumes two groups, but the derivations below is consistent with an arbitrarily large number.

A stationary state arrival mechanism is characterized by the fact that the time interval between the arrivals of the *n*-th and (*n*-1)-th cars is exactly equal to $s^i/\delta(s^i,n)$ where $\delta(s^i,n)$ is the headway distance that makes vehicle *n* drive with speed s^i and s^i is the speed in stationary state *i*. We will distinguish the stationary state mechanism before the change (*i=before*) in which all car drive at the same speed s^{before} and an alternative mechanism (*i=after*) that is relevant after the change. It is assumed that the free flow speed of vehicles belonging to any group exceeds the stationary state speeds to be considered. The last car that enters the road in the stationary state *before* is *n*=0, and therefore $s_t(0)=s^{before}$ for all *t*. The first car that enters after the change in arrival times is *n*=1.

We study what happens if the stationary state arrival mechanism changes suddenly from *before* to *after*. In order to do this we concentrate attention on the first kilometer (or an alternative distance unit) of a road of infinite length where all cars proceed initially at speed s^{before}.

Taking the time derivative of this equation gives:

$$\frac{ds_t(n)}{dt} = \frac{dg_n}{d\delta}(s_t(n-1) - s_t(n)).$$

We can rewrite this as:

$$\frac{ds_t(n)}{dt} + a_t(n)s_t(n) = a_t(n)s_t(n-1)$$

where *a* denotes the derivative of *g* with respect to δ. This equation can be interpreted as a linear differential equation with time-varying coefficients. The general solution to this equation can be written as:

$$s_t(n) = e^{-A_t(n)}\left[s^*(n) + \int_{t(n)}^{t} s_\tau(n-1)a_\tau(n)e^{A_\tau(n)}d\tau\right]. \qquad (*)$$

In this equation $t(n)$ denotes the time at which car *n* enters the road, $s^*(n)$ the speed of the *n*-th car at the time $t(n)$ when it enters the road and:

$$A_t(n) = \int_{t(n)}^{t} a_\tau(n) d\tau \,.$$

For n=1, this results in:

$$s_t(1) = e^{-A_t(n)} \left[s*(1) + s(0) \int_{t(n)}^{t} a_\tau(1) e^{A_\tau(n)} d\tau \right]$$

where the suffix t of the (constant) speed $s_t(0)$ has been suppressed. Now note that:

$$\frac{d}{d\tau} e^{A_t(n)} = a_\tau(n) e^{A_\tau(n)}$$

so that:

$$\int_{t(n)}^{t} a_\tau(n) e^{A_\tau(n)} d\tau = e^{A_t(n)} - 1$$

and use this to write:

$$s_t(1) = s(0) + \left[s*(1) - s(0) \right] e^{-A_t(1)}.$$

Now consider n=2. Substitution of the solution for n=1 in (*) gives:

$$\begin{aligned} s_t(2) = s(0) + & \\ & \left[s*(2) - s(0) \right] e^{-A_t(2)} + \\ & \left[s*(1) - s(0) \right] e^{-A_\tau(2)} \int_{t(2)}^{t} a_\tau(2) e^{A_\tau(2)} e^{-A_\tau(1)} d\tau. \end{aligned}$$

Note that:

$$\begin{aligned} e^{-A_\tau(2)} \int_{t(2)}^{t} a_\tau(2) e^{A_\tau(2)} e^{-A_\tau(1)} d\tau &= e^{-A_\tau(2)} \left[e^{A_\tau(2)} - 1 \right] \frac{\int_{t(2)}^{t} a_\tau(2) e^{A_\tau(2)} e^{-A_\tau(1)} d\tau}{\int_{t(2)}^{t} a_\tau(2) e^{A_\tau(2)} d\tau} \\ &= \left[1 - e^{-A_\tau(2)} \right] e^{-A_{t*(2,1,t)}(1)} \end{aligned}$$

for some t*(2,1,t) between t(2) and t. This means that we can write:

$$\begin{aligned} s_t(2) = s(0) + & \\ & \left[s*(2) - s(0) \right] e^{-A_t(2)} \\ & \left[s*(1) - s(0) \right] \left[1 - e^{-A_\tau(2)} \right] e^{-A_{t*(2,1,t)}(1)} \end{aligned}$$

Next, consider *n*=3. Substitution of the solution for n=2 gives:

$$
\begin{aligned}
s_t(3) = s(0) + \\
&\left[s*(3) - s(0)\right] e^{-A_t(3)} + \\
&\left[s*(2) - s(0)\right]\left[1 - e^{-A_t(3)}\right] e^{-A_{\tau*(3,2,t)}(2)} + \\
&\left[s*(1) - s(0)\right]\left[1 - e^{-A_t(3)}\right] \int_{t(3)}^{t} a_\tau(3) e^{A_\tau(3)} e^{-A_\tau(2)} d\tau
\end{aligned}
$$

where we have defined *t*(3,2,*t*) analogously to *t*(2,1,t). We can now define *t*(3,1,*t*) in analogously in order to simplify the notation of the fourth line. It is not difficult to verify that *t*(*k*,*m*,*t*), *k*>*m* is increasing in *t*. This leads to the following equation:

$$
\begin{aligned}
s_t(3) = s(0) + \\
&\left[s*(3) - s(0)\right] e^{-A_t(3)} + \\
&\left[s*(2) - s(0)\right]\left[1 - e^{-A_t(3)}\right] e^{-A_{\tau(3,2,t)}(2)} + \\
&\left[s*(1) - s(0)\right]\left[1 - e^{-A_t(3)}\right]\left[1 - e^{-A_{\tau*(3,2,t)}(2)}\right] e^{-A_{t(3,2,1)}(1)}
\end{aligned}
$$

Using this approach we can proceed further, but the regularity to be expected is now sufficiently clear.

A more useful form of the above equations is the following:

$$
\begin{aligned}
s_t(1) = \; & s*(1)\, e^{-A_t(1)} + \\
& s(0)\left[1 - e^{-A_t(1)}\right]
\end{aligned}
$$

$$
\begin{aligned}
s_t(2) = \; & s*(2) e^{-A_t(2)} + \\
& s*(1)\left[1 - e^{-A_\tau(2)}\right] e^{-A_{t*(2,1,t)}(1)} + \\
& s(0)\left[1 - e^{-A_t(2)} - \left[1 - e^{-A_\tau(2)}\right] e^{-A_{t*(2,1,t)}(1)}\right]
\end{aligned}
$$

$$
\begin{aligned}
s_t(3) = \; & s*(3)\, e^{-A_t(3)} + \\
& s*(2)\left[1 - e^{-A_t(3)}\right] e^{-A_{\tau(3,2,t)}(2)} + \\
& s*(1)\left[1 - e^{-A_t(3)}\right]\left[1 - e^{-A_{\tau*(3,2,t)}(2)}\right] e^{-A_{t(3,2,1)}(1)} + \\
& s(0)[1 - e^{-A_t(3)} - \left[1 - e^{-A_t(3)}\right] e^{-A_{\tau(3,2,t)}(2)} - \\
& \qquad \left[1 - e^{-A_t(3)}\right]\left[1 - e^{-A_{\tau*(3,2,t)}(2)}\right] e^{-A_{t(3,2,1)}(1)}]
\end{aligned}
$$

or in general:

$$
s_t(n) = \sum_{i=1}^{n} s*(i)\; w(n,i,t) + s(0)\left[1 - \sum_{i=1}^{n} w(n,i,t)\right]
$$

with all w's positive and smaller than 1 and $\Sigma w<1$. It is easy to verify that for any n Σw tends to 0 if $t\rightarrow\infty$. Moreover, it is easy to verify that for any t Σw tends to 1 if $n\rightarrow\infty$. The first observation implies that all cars will ultimately tend to drive at speed $s(0)$. The reason is that any higher speed can only be maintained for a finite time since car zero is driving in front of all cars that follow at constant speed $s(0)$ and cannot be overtaken. The second observation implies that the importance of the first car for the speed of those that follows tends to disappear completely for any finite time.

We now define the excess space of car n at time t as the difference between the actual distance to its leader and the distance $\delta(s^i,n)$ needed to make it choose the speed s^{before}:

$$e_t(n) = \delta_t(n) - \delta(s^{before}, n)$$

Note that for each car that enters the road this distance is positive and at least equal to $\delta(s^{after},n)$- $\delta(s^{before},n)$. For n=1 it is exactly equal to this lower bound, for all cars that enter later it is larger. To see this, consider n=2 and not that car 1 continually consumes some of its excess space since its speed is always larger than s^{before}. The excess space consumed by car 1 is added to that of car 2, since car 1 is the leader of car 2. By continuing this reasoning for n=3,4,.. we see that the excess space of the m-th car that enters is equal to the minimum value $\delta(s^{after},m)$- $\delta(s^{before},m)$ plus the amount that the (m-1)-th car has consumed of its own excess space after it entered the road.

The total amount of excess space on the road increases with a fixed amount with each car that enters. Since speed is determined by headway distance, this implies that the average speed of all cars that entered after car 0 must also increase. Since the speed has an upper bound s^{after}, it follows that the average speed must approach this value.

The excess space available tot cars 1,…,m decreases as soon as m has entered the road. We must therefore conclude that the speed at which cars enter the road will also approach s^{after}. But this requires that their leaders have also has driven at a speed approaching s^{after} during the time interval between the times they entered, and so on. We must therefore conclude that ultimately the new stationary state is approached with cars driving at a constant speed s^{after}.

Now consider a situation in which the time interval before cars that enter the road decreases. If there is a stationary state corresponding to a lower speed, similar arguments as used above can be used to demonstrate convergence to this state. The main difference is that excess space is now negative (it may be called deficit space) but this does not change the argument.
However, if the starting situation is one with hypercongestion, there is no such lower stationary speed. Therefore traffic cannot converge to a positive speed that is lower than s^{before}. However, as deficit space continues to grow, average speed must decrease when cars enter with time intervals determined by the unreachable stationary state. This leads ultimately to a situation in which the headway distance at these arrival times are too small to induce the driver to enter at a positive speed, implying that a queue will develop before the entrance of the road.

In summary: we reach the same conclusions as Verhoef (2001) did for homogeneous traffic: hypercongested stationary states are dynamically unstable whereas non-hypercongested stationary states are stable.