1 THE CONGESTION DILEMMA

1.1 Introduction

With respect to the prevailing congestion problems in the more urbanised regions of the European Union, transportation planners and policymakers are facing a dilemma. Supply-side measures, i.e. increasing the capacities, might shorten the congestion duration, especially if bottlenecks can be removed. However, the addition of extra lanes in motorway bottlenecks is generally very costly, not only in terms of investment costs, but also in terms of space consumption and environmental effects. The strategy to build to provide, that is, adjusting the capacities to the expected peak demand, is sub-optimal from an economic point of view. Economists have shown that users tend to respond to congestion in a way that is not optimal from a system point of view (see for instance Walters, 1968). In many cases, there is a mismatch between the congestion costs experienced by an individual user and the congestion costs caused by the same user, and demand-side measures might help to stimulate 'social' behaviour. The problem is, however, that it is hard to convince the public that measures like congestion pricing are really effective in preventing congestion. The allegation that congestion pricing would only result in paid queuing instead of free queuing has successfully been used in a campaign against the introduction of road pricing in the Netherlands.

The objections against congestion pricing, especially the doubts about the effectiveness of this solution, are less irrational than some advocates of the pricing mechanism try to let us believe. An example is the mismatch between the willingness-to-pay of a potential road user and the utility of the planned trip, which results primarily from income differences. A related problem is the redistribution of welfare. Many (potential) road users are 'direct losers'; that is, their direct transportation costs are higher, though this might be (partially) compensated by a form of tax refunds (Thomson, 1998). Even more convincing, however, is the argument that a knowledge gap prevents even the most flexible congestion-pricing scheme to be optimal. In the case of simultaneous pricing, the congestion costs can be calculated exactly, but the users cannot base their choices on the exact costs (Thomson, 1998). Hence, the tolls will have to be based on demand estimations, and should not vary too much. This might result in under-utilisation of the road if the tolls are too high, or, more likely, congestion if the tolls are not high enough. Moreover, since users are not capable of perfectly performing the utility-maximisation task assumed in classical economic theory, road pricing will never be a perfect way to avoid congestion.
1.2 Departure time choice: a prisoner's dilemma

Contrary to what might be concluded from the previous section, structural congestion is not an inevitable phenomenon. Structural congestion is congestion caused by factors such as recurring peak demands, contrary to incidental congestion caused by unpredictable factors such as adverse weather conditions and traffic accidents. In air traffic and in rail traffic, structural congestion can be avoided by limiting the number of scheduled trips. If the number of requested flights on an airway corridor is greater than its capacity, some flights will have to be rescheduled. For safety reasons, queuing in the air should be avoided. Theoretically, the same principle could be used to eliminate structural congestion on motorways. If every user would reserve the planned path on the road network in advance, congestion could be prevented by limiting the number of users per road segment per period of time. At first sight, we would conclude that this would shorten the trip durations, but at the expense of departure time adjustments. It can be reasoned, however, that it is theoretically possible to speed up all congested trips in a single-bottleneck case by maintaining the same arrival times, and choosing later departure times.

The departure time choice process can be seen as a kind of prisoner's dilemma. If we assume that none of the users has to pass more than one bottleneck, it would be possible for each user to choose a departure time such that the arrival time is the same, but without any queuing delays. However, this situation is not a user equilibrium, since users are not any more rewarded for shifting their departure times to the period before or after the peak. The result of this prisoner's dilemma is that the departure time choice equilibrium resulting from individual optimisation behaviour is not optimal, even from an individual point of view. Section 3 will elaborate on the departure time choice process in a specific situation that is described in section 2. By assuming a simplified situation, it is possible to analyse the difference between user-equilibrium departure time choice and system-optimum departure time choice. Section 4 elaborates on the dilemma mentioned in this paragraph.

1.3 Slot reservation systems

The implementation of slot reservation on (urban) motorways would enable the network managers to limit the maximum number of users per period of time, which would help to prevent congestion a priori; thus, the gap between user-equilibrium and system-optimum departure time choice could be bridged. Wong (1997) has presented an outline of a slot reservation system for road transport, with the emphasis on the information flows.

Passenger seat reservation systems are commonly used in long-distance public transport and a slot reservation system on a freeway could function similarly. An analogy with seat reservation systems is the possibility to vary the reservation tariff with the demand and the option to distinguish between different qualities of slots. A major difference is that a seat is a well-defined object that is clearly identifiable, but a slot is harder to define and harder to identify in practice. A major problem is how to indicate to the users whether they are admitted to the system, how to show the users what are the 'edges' of a slot (in time and space), and how to prevent 'violations' of the slot regime. Since freeways have only a limited number of access points, this problem will be easier to solve for freeways than for ordinary roads.
1.4 Problem definition

The aim of this paper is to provide theoretical insights that are necessary to estimate the potential benefits of a freeway slot reservation system. Only if there are convincing indications that slot reservation would lead to considerably less congestion costs, there is sufficient justification to study the technical aspects of freeway slot reservation systems, as well as the costs involved. This paper is confined to the problem how to estimate the potential benefits of slot reservation, compared to a user-equilibrium situation.

Intuitively, there are a few reasons why user-equilibrium traffic patterns might not be optimal. Three main considerations will be discussed in this paper:
- Different users have different preferences with respect to congestion delays versus departure time rescheduling.
- The delay experienced by a user is not necessarily the same as the delay caused by the same user.
- Users do not have perfect knowledge about future delays and thus are not able to predict their congestion disutilities perfectly.

A comparison of the queuing costs and the rescheduling costs in the user-equilibrium and in the system-optimum situation is presented in the third section. The analysis is based on the first two main considerations mentioned in the last paragraph; the last consideration is only assessed in the final part of the second section.

2 OPTIMISATION OF DEPARTURE TIMES: BASIC ASSUMPTIONS

This section presents a comparative analysis of the opportunity costs of user-equilibrium and system-optimum departure time choices in a specific social and spatial context. The spatial context is a situation with two towns connected by a freeway. The capacity of the freeway is determined by a bottleneck, which is located somewhere between the two towns. There are no alternative routes (see figure 2.1).

![Figure 2.1: situation](image)

The social context is that there are a number of users who want to combine an activity in the first town with an activity in the second town. All users are assumed to maximise a deterministic utility function, or, equivalently, to minimise a deterministic cost function. In this paper, costs are not only monetary costs, but include all kinds of disutilities, especially activity durations. Furthermore, it is assumed that the number of users is given, and that they don't have alternatives except departure time choice.

2.1 Behavioural assumptions

The analysis is based on a number of simplifying assumptions. Generally, the standard assumptions of economic theory apply for this analysis. The user equilibrium analysed in this paper assumes perfect knowledge and rational behaviour of all users.
Of course, this is a too optimistic view of what happens in the real world, but it can be helpful to identify the dominant incentives to shift departure times. Moreover, it is assumed that the freeway slot reservation system is able to attain a good approximation of the system optimum, which is defined as the situation with minimum total costs, which consists of delay costs and rescheduling costs.

The utility of both activities is assumed to be a (piecewise) linear function of the starting time or the ending time. This corresponds with the assumption that the value-of-time is a (piecewise) constant function of time. It is assumed that the values-of-time of the origin and destination activities are positive. In the analysis presented in this paper, the value of time of the origin activity changes at \( t_0 \) from \( \alpha_{ante} \) to \( \alpha_{post} \), while the value of time of the destination activity remains \( \alpha_{new} \). This corresponds with the situation that a high-utility activity (for instance: work) has a preferred ending time \( t_0 \). However, it is also possible to do a similar analysis when \( t_0 \) is the starting time of a high-utility activity; then the key variable would be arrival time instead of departure time. All users are assumed to have the same \( t_0 \), but the values of time of the origin and destination activities can be different.

\[
\begin{array}{c|c|c}
\text{origin} & \alpha_{ante} & \text{value-of-time} \\
\hline
\phi & \alpha_{new} & \text{destination} \\
\hline
\nu & \alpha_{post} & \gamma \\
\end{array}
\]

\textbf{Figure 2.2: timing utilities}

The symbols \( \nu \) and \( \gamma \) (figure 2.2) have been introduced to represent the marginal opportunity costs of delays and departure time adjustments respectively. Marginal opportunity costs are the derivatives to time of the costs of performing a sub-optimal activity. The reference is the situation where the departure time is \( t_0 \) and where there are no congestion delays:

\[
\alpha_{travel} < \alpha_{post} < \alpha_{new} < \alpha_{ante}
\]

\[
\nu = \alpha_{new} - \alpha_{travel} \quad \text{marginal opportunity costs of delays}
\]

\[
\phi = \alpha_{ante} - \alpha_{new} \quad \text{marginal opportunity costs of earlier departures}
\]

\[
\gamma = \alpha_{new} - \alpha_{post} \quad \text{marginal opportunity costs of later departures}
\]

\textbf{2.2 Queuing model}

\textbf{Figure 2.3: speed-flow relationship}

All cars have to pass one single bottleneck and all origins and all destinations have the same distance from the bottleneck. If all users would depart at their preferred departure time \( t_0 \), queuing would inevitably occur. In this analysis, a deterministic
queuing model is applied instead of a performance function, which, for instance, is used by Henderson (1981) and more recently by Wang (1996). The main weakness of the latter approach is that the capacity of the road is not finite, and that delays are independent of the preceding traffic volumes. In the deterministic queuing model, all vehicles that cannot directly pass the bottleneck because of capacity constraints are stored in a queue. A disadvantage of the deterministic queuing model is, however, that delays caused by nearly congestive traffic are neglected, because it is based on the assumption that there are no delays if the traffic demand does not exceed the bottleneck capacity.

3 USER-EQUILIBRIUM DEPARTURE TIME CHOICE

This section presents an analysis of user-equilibrium departure time choices. Differences in preferences between users are incorporated in the model by distinguishing between different user classes. Koolstra (1999) discusses this model in further detail.

3.1 No schedule adjustment

If all potential users depart at approximately the same time, a queue will be formed with a maximum delay of \( Q/C \) hours, where \( Q \) denotes the instantaneous demand (number of vehicles) and \( C \) the capacity (vehicles per hour). Figure 3.1 provides a graphical example of the situation without schedule adjustment. The horizontal axis represents departure time, and the vertical axis represents both queuing delays and vehicle intensities divided by the capacity. In the case presented in figure 2, the instantaneous demand is two times the hourly capacity. Thus, the resulting maximum delay is 2 hours, and the mean delay is 1 hour. It is obvious that there is a strong incentive to choose an earlier departure time, even if departure time adjustments are valued the same as queuing delays.

3.2 User equilibrium

If users are able to leave earlier, they can significantly reduce their travel costs, especially if we assume that the costs of leaving 10 minutes earlier are considerably less than the costs of having to wait 10 minutes in a queue. As long as the costs of departing earlier are less than the queuing costs, there is still an incentive to choose an
earlier departure time. The rescheduling incentives are assumed to disappear if and only if Wardrop's first principle is met, which is the definition of the user equilibrium. In Wardrop's original formulation, this principle requires that the journey times on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route (Wardrop 1952, p. 345). If 'route' is replaced by 'departure time', the same principle can be used to define the scheduling equilibrium.

Assume that all users of a certain class have the same cost function, which is a linear function of delay costs and rescheduling costs. The value-of-time parameters $\varphi$ and $\upsilon$ indicate the costs per hour departure time shift and the costs per hour delay respectively, as perceived by the users themselves. The equivalent of Wardrop's first principle is that for each user the sum of queuing costs and rescheduling costs is equal for all used departure times. Users only depart earlier if the additional rescheduling costs are compensated by less queuing costs. For all used departure times, the increase rate of the queuing delay as a function of departure time must equal the gradient of the cost function, which is the ratio of the rescheduling value-of-time $\varphi$ and the delay value-of-time $\upsilon$:

$$\frac{\partial \varphi}{\partial t} = \frac{\varphi_1}{\upsilon_1}$$

where:

- $w$ queuing delay [h]
- $t$ departure time adjustment (relative to $t_0$) [h]
- $\varphi_1$ rescheduling costs of user class $i$ (earlier departure)

### Legend:
- **---** Departures user class 1
- **-** Arrivals (total)
- $\frac{\varphi_1}{\upsilon_1}$ Queue share user class 1

**Figure 3.2: equilibrium for one class, only earlier departures**

Figure 3.2 sketches a situation where the users of one class have reached the equilibrium, provided that they only consider the option to leave earlier. In the figure, the departure rates (before the bottleneck) and the arrival rates (after the bottleneck) are indicated, as well as the resulting queuing function. For each individual, the departure time choice can be considered as a linear programming problem; the goal is minimisation of the cost function and the queuing delay is the main constraint. Only departure times are used that result in minimal costs; these are the departure times where the queuing function equals the minimum iso-cost curve.

Since the queuing delay, as a function of time, grows with the departure rate minus the capacity, the departure rates are:

$$\frac{\partial \varphi}{\partial t} = \frac{\varphi_1}{\upsilon_1} = f_i \lambda_g 1$$

$$f_i \lambda_g 1 + \frac{\varphi_1}{\upsilon_1}$$

where:

- $f_i$ departure rate (relative to the capacity) of user class $i$ [h$^{-1}$]
However, not only earlier departures, but also later departures are possible. The
equilibrium is reached under similar conditions:
\[
\frac{\partial w}{\partial t} = -\frac{\gamma_i}{v_i} = f_i \log_1 \frac{\gamma_i}{v_i}
\]
\[\gamma_i\] rescheduling costs of user class \(i\) (later departure)

The users of class two will also adjust their departure times. Moreover, both earlier
and later departures are possible. Figure 3.3 shows that assuming symmetric departure
time adjustment behaviour (\(\forall i: \varphi_i = \gamma_i\)) a symmetric queuing pattern develops, with \(t_0\)
as the axis of symmetry. The departure rates, however, are not 'mirrored' on \(t_0\), but
show a 180°-rotation symmetry around \((0,1)\). This is contrary to the approach of
Henderson (1981), which results in symmetric departure rates. The users with a
relatively low value of delay compared to the value of departure adjustment stay in
the peak, while the users with a relatively high value of delay tend to choose the
largest departure time adjustments. A difference with the single user-class case is the
generation of a small after-peak, which has its (local) maximum when the queue
dissolves. The symmetry disappears if each user class has different values \(\gamma\) and \(\varphi\).
This analysis assumes that the relationship between \(\gamma\) and \(\varphi\) is the same for all users:
\[\varphi_i = \kappa \cdot \gamma_i \quad \forall i\]

### 3.3 System optimum

The user equilibrium, as described in the previous section, is not optimal if the system
goal is minimisation of the sum of the congestion costs and the rescheduling costs of
all users. The model used in this paper permits that all congestion is eliminated
without an increase in the mean rescheduling costs per user.

![Figure 3.4: optimal schedule](349)
In the system-optimum situation, as shown in figure 3.4, all queuing is replaced by departure time adjustments. Moreover, user class 2 is granted priority over user class 1 if user class 2 has a greater rescheduling value-of-time than user class one. To prove that this is indeed optimal, assume that a user of class 1 changes its departure time with a user from class 2. If 2 has a higher value and has a greater time shift from \( t_0 \), it is clear that the costs are also greater than in the previous situation.

There are two different substitution effects that improve the performance of the system-optimum schedule compared with the user-equilibrium schedule:

- shift of trips from pre-peak to after-peak period;
- the priority of user groups is based on differences in the absolute value of adjustment time instead of the relative value of adjustment compared to the value of travel time.

The first effect is equal to the total delay costs in the user-equilibrium situation. This implies, that in the situation modelled in this paper, all queuing delay can be eliminated, while the departure time adjustment costs stay the same. It might surprise us that the adjustment costs do not increase, but this is because the elimination of all queuing can simply be achieved by shifting the excess departures in the period before \( t_0 \) to the period after \( t_0 \). The adjustment costs can even be lower, because the sequence of user classes is optimally determined by the rescheduling value-of-time, instead of determined by the ratio between the rescheduling value-of-time and the queuing value-of-time, which happens in the user equilibrium.

Generally, the first effect will be more important than the second effect. If the cost preferences would not differ between user classes, the cost reduction of eliminating the queuing would be 50% and there would be no priority effect. The priority effect can be significant if there are large differences in cost preferences, the queuing elimination effect, however, would also remain significant.

## 4 GENERALISATION TO A VARIABLE DEMAND MODEL

In this section, we discuss generalisations of the conclusions of the previous section. The assumption of the previous section that all users want to depart on the same time is relaxed. It is now assumed that for each user class, the preferred departure time can be described by a bell-shaped symmetric function with mean \( t_0 \).

### 4.1 Robustness of the user-equilibrium departure time adjustment process

It can be assumed that the equilibrium departure intensities \( f(t) \) have the same levels if the assumption that all travellers prefer to depart at \( t_0 \) is relaxed. The main reason is that this equilibrium is the result of individual optimisations of the balance between rescheduling costs and queuing costs. Since this balance depends on the marginal rescheduling costs compared with the marginal queuing costs, the equilibrium levels are basically unaffected by changes in the distribution of preferred departure times. For the individual user in the model, however, it will no longer be the case that both departing earlier and later has equal costs. The preferred departure time of the individual user will determine whether a departure time after \( t_0 \) or a departure time before \( t_0 \) will be chosen.
An important prerequisite for departure time adjustments is that the demand rate has to be less than the equilibrium departure rate before the peak. If, for instance, the demand is 1.4 times the capacity, and the queuing costs per minute are twice the rescheduling costs per minute, there is no incentive for earlier departures. The same is true for after-peak periods if the demand rate after the peak is still higher than the equilibrium departure rate after the peak. Thus, there is a margin around the capacity where there are no incentives for departure time adjustments. The bandwidth of this margin depends on the relative valuation of queuing compared with departure time adjustments. If queuing is valued relatively high, the bandwidth is smaller and the departure time choices are more compliant with the system-optimum departure time choices.

4.2 Robustness of the gap between user-equilibrium and system-optimum

Section 3 revealed that, given the assumptions of section 2, it is theoretically possible to eliminate all congestion costs, without increasing the mean departure time adjustment costs. It is questionable, however, to what extent this conclusion can be generalised. For instance, in the case that the demand levels are too low to result in spontaneous departure time adjustments, but still such that queuing will occur, it is intuitively clear that the benefit of queuing reductions is partly compensated by higher rescheduling costs.

However, given a few assumptions, it is possible to show that each single user could profit from a small departure time adjustment compared with the user equilibrium. Assume that the queuing function is the same as in section 2 and that there is a maximum of 1 bottleneck per route. The assumptions made in section 2 of one origin, one destination, one route, and one preferred departure time, however, can be relaxed. The exit rates of the bottlenecks are per definition equal to or less than the bottleneck capacity. Now assume that all users maintain the same arrival time as in the user equilibrium, and optimise the departure time given the arrival time. Since the exit rates of the bottlenecks cannot exceed the capacity, this means that queuing would not occur in this situation. Thus, all previously congested users can exchange some time spent in their cars for time spent at the location of origin, which is generally valued as a more useful way to spend time.

In short, when choosing their departure times, users are facing a kind of prisoner's dilemma. In the prisoner's dilemma, two prisoners can expect acquittal if they both deny guilt. However, if one is betrayed by the other, he can expect imprisonment for quite a number of years. In that case, the betrayer can expect penalty reduction. In this case, it is clear that both prisoners would be better off by denying guilt in court. However, if they do not know what the other will do, the best strategy is to betray each other, which results in reduced imprisonment for both.

Just as the prisoners in the original problem would both be best off by denying, but will both opt rationally for betraying the other, all travellers would be better off if they would adjust their departure times to the user-equilibrium arrival times. However, since road travellers cannot make such an agreement, the result is congestion. The implementation of a reservation system on motorways, such as proposed by Wong (1997) could artificially prevent congestion by imposing a limit on the maximum number of users per road per period. Since it is impossible to know exactly what
arrival time would have resulted for each individual user in a situation without slot reservation, it is not possible to make the new situation beneficial for each individual, but since capacity elimination is the most important characteristic of the system-optimal departure pattern compared with the user equilibrium, it is probable that slot reservation will be beneficial to a great majority of users.

5 INFORMATIONAL ASPECTS

To interpret the results of the analysis correctly, informational aspects should be taken into account. In the previous analysis, it was assumed that all actors have perfect information. Based on this assumption, the system optimum can be achieved by both a slot reservation system and by a congestion-pricing scheme. In order to be optimal, the slot reservation system should grant absolute priority to users with a higher rescheduling value-of-time, and the congestion price, as a function of time, should also be chosen correctly. In practice, however, there are three major information gaps that preclude optimal traffic management:

- Traffic managers have only limited knowledge about the demand function, and thus are not able to predict beforehand what the optimal tolls should be.
- Traffic managers have only limited knowledge about the differences in value-of-time between users. However, there are methods that stimulate users to reveal their willingness-to-pay, for instance second-price auctions (Lerz, 1996).
- Users do not have perfect knowledge about the traffic supply.

Based on the analysis presented in this paper, it can be assumed that the first problem is generally more serious than the second. Even if all users would be given the same chances to reserve a slot on any time, the slot reservation system could eliminate practically all queuing costs, and thus it would already be quite effective. If the toll levels are chosen incorrectly, however, congestion might still prevail, or the roads might be under-utilised. Moreover, the 'optimal' tolls might be quite high, which triggers the question how the users can be compensated in order to prevent unwanted welfare redistribution effects. Without sufficient compensation, congestion pricing will probably not be accepted by the public, especially if the 'optimal' tolls are relatively high (See Thomson, 1998, for an overview of the practical problems of congestion pricing).

6 CONCLUSIONS

This final section summarises the main conclusions of the analysis presented in this paper. The most important conclusion is that given a number of simplifying assumptions, the queuing costs can be eliminated by using a slot reservation system, without an increase in departure time adjustment costs. There is, however, still another difference between the user-equilibrium schedule and the system-optimum schedule, and that is the sequence of user groups. In the first case, the ratio between the rescheduling values-of-time and the queuing values-of-time determines the sequence, while in the optimal case the sequence is directly determined by the rescheduling values-of-time. However, this effect will generally be less important than the first effect.
The conclusion that all queuing can be eliminated without an increase in the mean rescheduling costs cannot hold under all circumstances. For instance, the peak demand has to exceed a certain level in order to trigger departure time adjustments. However, given some moderated assumptions, it is still possible to prove that for each situation with congestion, there exists a solution without queuing delays, but still with the same arrival times.

From the analysis, we can conclude a freeway slot reservation system is indeed an interesting option, at least theoretically. In practice, slot reservation might be more effective in reducing congestion than road pricing. Slot reservation with variable reservation fees is an option to combine the benefits of both alternatives. However, whether a freeway slot reservation is really feasible in practice will depend on issues such as the technical feasibility, infrastructure requirements of the system, the acceptance by both the public and by policymakers, and, last but not least, the system costs.

References


