Journal of Economic Dynamics & Control 34 (2010) 598–603

Contents lists available at ScienceDirect

Journal of Economic Dynamics & Control

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

# On the relation between the mean and variance of delay in dynamic queues with random capacity and demand

Mogens Fosgerau [a,b,*]

[a] Technical University of Denmark, Knuth-Winterfeldts Allé 116V, 2800 Kgs. Lyngby, Denmark
[b] Centre for Transport Studies, Sweden

ARTICLE INFO

*Article history:*
Received 30 October 2008
Accepted 27 September 2009
Available online 4 January 2010

*JEL classification:*
D8
R41

*Keywords:*
Bottleneck model
Random capacity
Congestion
Nash equilibrium
Loop

ABSTRACT

This paper investigates the distribution of delays during a repeatedly occurring demand peak in a congested facility with random capacity and demand, such as an airport or an urban road. Congestion is described in the form of a dynamic queue using the Vickrey bottleneck model and assuming Nash equilibrium in arrival times. The paper shows that the expected delay and the variance of delay vary differently over time during the peak and must hence be considered separately. The paper gives some characterization of how the expected delay and the variance of delay are related, which explain the looping phenomenon that has now been observed a number of times. Empirical illustration is provided.



## 1. Introduction

Congestion is becoming an increasingly serious problem in many places, including road and rail networks, airports and airspace. We may also think of congestion in computer networks, at ski resorts, McDonalds and many other places. In such places, there is often a pattern where peaks in demand and resulting congestion repeat regularly like the morning rush hour.

Congestion is an inherently dynamic phenomenon, since an arrival at some point in time affects only users arriving later, not those arriving earlier. The dynamics are important for example for the consideration of time-varying tolls designed to internalize the costs of congestion. For such purposes it is necessary to consider the profile of demand and costs over a period of time such as a day.

Congestion does not just cause delays. Congestion also causes delays to become increasingly unpredictable as random variations in capacity and demand become important in facilities operating near capacity. The random variability of delays is a significant part of user costs when congestion is severe and is hence important in its own right.

Economic analysis of congestion generally involves the notion of equilibrium where the profile of demand over a period of time, say a day, is endogenous. The analysis of equilibrium is, however, complicated when allowing for both dynamics and random delay. It is then tempting to ignore the aspect of randomness from analysis of congestion. One could have the

* Correspondence address. Technical University of Denmark, Knuth-Winterfeldts Allé 116V, 2800 Kgs. Lyngby, Denmark. Tel.: +45 45256521.
*E-mail address:* mf@transport.dtu.dk

0165-1889/$ - see front matter 
doi:10.1016/j.jedc.2009.12.002

intuition that randomness does not essentially influence the analysis and that expected delay and delay risk are proportional. This is, however, not the case as the present paper shows.

This paper incorporates all the elements mentioned, dynamic congestion, randomness and equilibrium, and shows how the dynamics of congestion cause the time profile of expected delay and the variance of delay to be different. More specifically, the paper considers a regularly occurring demand peak in a congested facility with random capacity and demand. If the demand peak repeats every day, we may consider the random distribution of delay at different times: The random distribution of delays will be different at, e.g., 8 AM and at 9 AM. Consider then the expected delay and the variance of delay as functions of time. The paper shows that the following implication holds under quite general assumptions: If the variance of delay decreases then also expected delay decreases. This may equivalently be expressed as the logical converse: If the expected delay increases then also the variance of delay increases.

This result has implications for the relationship between the mean and variance of delay, which may be verified against empirical data. At the beginning of the demand peak, both the expected delay and the variance of delay are small and increasing. At some point in time the expected delay reaches a (possibly local) maximum. At this point the variance of delay must be increasing as a consequence of the above result. At some point in time the variance of delay reaches a (possibly local) maximum. At this point the mean delay must be decreasing. Hence a plot of the variance of delay against the expected delay forms a counter-clockwise loop. Such loops have been found repeatedly in empirical data, although this seems to have been reported only a few times in published papers.

Section 2 shows some examples of congested facilities. Plots of the variance of delay against the mean delay at different times of day exhibits the characteristic counter-clockwise loop during demand peaks as predicted by the theoretical analysis in this paper. This paper is the first to provide a theoretical explanation for the looping phenomenon.

The Vickrey (1969) bottleneck model captures many of the essential features of equilibrium demand for a congested facility. The congested facility is described using a bottleneck congestion technology, where queueing users are served at a fixed service rate. A vertical queue builds up when users arrive at a faster rate and dissipates when users arrive at a slower rate. There is a continuum of users assumed to have costs of delay and also scheduling costs such that deviations from their preferred service time are costly. In Nash equilibrium, each user selects his optimal arrival time, conditional on the actions by the other users. With identical users this translates into the condition that user costs are constant over the interval when users arrive and larger outside. The bottleneck model and its implications for congestion pricing have been analyzed extensively (Arnott et al., 1993). In particular, there is always a queue in the Nash equilibrium. In the social optimum there is no queue since users then arrive at a rate equal to the service rate. It is fairly easy to design an optimal time-varying toll to implement the social optimum.

Arnott et al. (1999) consider the case when the ratio of demand to capacity is random. Their main interest is to investigate the effect of imperfect information to users about the random capacity and hence delays, where they find that information is not always welfare improving. The equilibrium arrival rate has not been found in the bottleneck model with random capacity and demand, but Arnott et al. (1999) are able to show that it is concave. This paper will assume a concave arrival rate with this motivation. The optimal time-varying toll has also not been found. It is therefore of interest also from this perspective to establish properties of the delay costs.

Section 2 provides some empirical examples of loops. The theoretical model is formulated in Section 3 and the analysis is carried out in Section 4, with proofs deferred to Appendix A. Section 5 concludes.

## 2. Looping examples

This section presents two examples of plots of the variance of delay against the expected delay in congested facilities. The examples exhibit the characteristic counter-clockwise loop during a demand peak. That is, the variance of delay peaks later than the mean delay in these examples. The choice of examples is motivated by data availability. Other examples have been analyzed by Department for Transport (2006).

The first example concerns a congested urban road in Copenhagen. The data record the travel time on an 11 km stretch of road, using cameras to match licence plates at the entry and the exit. A data point consists of the average travel time recorded during 1 min and the data cover a three month period. The mean travel time is estimated as a function of the time of day using nonparametric kernel regression. The variance is estimated as a function of time of day using nonparametric kernel regression of the squared residuals of the mean regression against the time of day. The estimated standard deviation is then computed as the squareroot of the estimated variance function. More details are available in Fosgerau and Karlstrom (2010). There is a distinct morning peak in the data, which appears as the counter-clockwise loop in Fig. 1.

The second example concerns a congested rail section in Denmark. Each data point consists of the delay of a rail arrival relative to the time table. Data cover all arrivals during a year. Details are available in Fosgerau and Hjorth (2008). These data have two distinct demand peaks, one in the morning and one in the afternoon. Both are clearly visible as counter-clockwise loops in Fig. 2.

Some common features of these examples may be noted. First, both concern queueing systems with a queueing discipline that is roughly first-in-first-out, since possibilities for overtaking are limited. Second, demand peaks occur regularly every weekday. Third, randomness of capacity and demand plays a significant role as evidenced by the ratio of