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

Fosgerau, Mogens

Technical University of Denmark, Centre for Transport Studies, Sweden

2010

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

Mogens Fosgerau

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

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.

Key words: Bottleneck model, Random capacity, Congestion, Nash Equilibrium, Loop

JEL codes: D8, R41

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.

Email address: mf@transport.dtu.dk (Mogens Fosgerau)

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 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 below 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 the appendix 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 one minute 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 (2009). There is a distinct morning peak in the data, which appears as the counter-clockwise loop on figure A.1.

FIGURE 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 on figure A.2.

FIGURE 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 the standard deviation relative to the mean delay. The bottleneck model with random capacity and demand, introduced in the next Section, represents these features in a simplified way.

3. Model formulation

Consider a bottleneck with a capacity of $\frac{1}{s}$ users per time unit. The number of users is normalized to 1, such that s becomes the time to serve all users. This means that variation in s encompasses variation in bottleneck capacity as well as in demand. Assume that the service time is random with s having a density ϕ and corresponding cumulative distribution Φ and assume also that the service time is bounded from above by $\overline{s}$ such that the queue cannot last indefinitely long. One realisation of s is drawn each period, such that the ratio of capacity to demand is constant in each period.

Users arrive during some interval of time, which we may take to be simply $[0, 1]$. Users arrive at the time varying rate $\rho(t) \geq 0$ such that the total number of users having arrived at time t is $R(t) = \int_0^t \rho(t')dt'$ and $R(1) = 1$. We assume that R is concave or equivalently that ρ is decreasing. Concavity of R results in Nash equilibrium under certain assumptions on user preferences (Arnott et al., 1999).

Conditional on the capacity, the length of the queue at time t is

$$Q(t) = \max\left(\left(R(t) - \frac{t}{s}\right), 0\right). \tag{1}$$

This is the number of users that have arrived but not yet been served. The queue at time 0 when the first user arrives is zero. We assume a nonzero probability of a queue arising. We introduce a function $s(t)$ to indicate the service time at which the queue is exactly gone at time t. From (1) we have $s(t) = \frac{t}{R(t)}$. It is useful to establish a few properties of this function, namely

$$s(0) = \lim_{t\to 0} s(t) = \frac{1}{\rho(0)}, s(1) = 1, \frac{\partial s}{\partial t}(t) = \frac{1 - s(t)\rho(t)}{R(t)} > 0,$$

where the latter inequality follows by concavity of the cumulative arrival rate R.

The waiting time in the queue for users arriving at time t is $q(t)$ given by

$$q(t) = sR(t) - t$$

when $s \geq s(t)$ with $q(t) = 0$ otherwise. Now we may investigate the expected time in the queue given arrival at time t. Let $\mu(t)$ denote the expected queue

time at time t. Then

$$\mu(t) = E(q(t)) = \int_{s(t)}^{\overline{s}} (sR(t) - t)\phi(s)ds.$$

We see immediately that $\mu(0) = 0$ and that $\mu(t) \to 0$ as $t \to \infty$. We may similarly define the variance of time in the queue given arrival at time t. Denoting the variance by σ (omitting the power for notational simplicity) we have

$$\begin{aligned} \sigma(t) &= Eq^2(t) - \mu^2(t) \\ &= \int_{s(t)}^{\overline{s}} (sR(t) - t)^2\phi(s)ds - \mu^2(t). \end{aligned}$$

Again we see immediately that $\sigma(0) = 0$ and that $\sigma(t) \to 0$ as $t \to \infty$.

The derivatives of μ and σ are given by

$$\begin{aligned} \mu'(t) &= \int_{s(t)}^{\overline{s}} (s\rho(t) - 1)\phi(s)ds \\ \sigma'(t) &= 2\int_{s(t)}^{\overline{s}} (sR(t) - t)(s\rho(t) - 1)\phi(s)ds - 2\mu(t)\mu'(t). \end{aligned}$$

This completes the model formulation

4. Analysis

We are now able to formulate the main proposition relating the derivatives of $\mu(t)$ and $\sigma(t)$ to each other. Recall that $s(t)$ is the service time such that the queue is exactly gone at time t. Then $\Phi(s(t))$ is the probability that the queue is gone at time t.

Proposition 1 $\sigma'(t) > 2\mu'(t)\mu(t)\frac{\Phi(s(t))}{1-\Phi(s(t)).}$

Proof is given in the Appendix. This proposition has the following immediate corollary.

Corollary 1 $\mu'(t) \geq 0 \Rightarrow \sigma'(t) > 0$ *or equivalently* $\sigma'(t) \leq 0 \Rightarrow \mu'(t) < 0$.

We shall need another small proposition, stated without proof. Define first $\overline{t}$ by $s(\overline{t}) = \overline{s}$ as the time at which the queue is always gone.

Proposition 2 $\mu'(0) > 0, \sigma'(0) = 0, t > \bar{t} \Rightarrow \mu(t) = \sigma(t) = 0.$

These results give a lot of information about the possible shape of the plot of $\sigma(t)$ against $\mu(t)$ The plot begins at $(\mu(0), \sigma(0)) = (0, 0)$ and moves East and then North-East. Whenever μ reaches a maximum, local or global, σ will still be increasing. Specifically, it is ruled out that σ peaks at the same time as μ and hence a loop will always be formed. When, later, σ reaches a local or global maximum, μ will be decreasing. When t is large enough that the queue is always gone, both μ and σ are back to zero.

A final proposition gives a lower bound for the variance of delay as a function of the mean delay. Proof is given in the Appendix.

Proposition 3 $\sigma(t) > \mu^2(t)\frac{\Phi(s(t)}{1-\Phi(s(t)}.$

The proposition has a nice interpretation. The variance of delay σ at time t is larger than μ^2 times a function of t. This function is 0 at $t = 0$ and increases towards infinity, which is reached at $t = \bar{t}$. So when t is large, either σ is very large or μ is very small.

5. Concluding remarks

This paper has shown how the bottleneck model with random capacity and demand implies the loop that has been observed empirically a number of times in plots of the variance of delay against mean delay for some congested facilities. The results in the paper give some restrictions on the possible shape of such plots during a demand peak in the random capacity and demand bottleneck model. In particular, there will always be at least one counter-clockwise loop. The possibility of more loops is, however, not ruled out, although that would be an aesthetically pleasing result. It is possible for the curve $(\mu(t), \sigma(t))$ to move to the North-East, back towards the origin and North-East again, after turning clockwise some to the West of the first extreme point to the East. It seems not to be possible to impose simple restrictions on the theoretical model that rules out such behaviour.

The correspondence between the empirical finding of counter-clockwise loops like those shown in the paper and the random capacity and demand bottleneck model adds to the credibility of the model as a description of real congestion phenomena. It is, however, a restriction that the model assumes that the ratio of capacity to demand is fixed during each period. It is not known to which degree the present results are robust with respect to random variation in capacity and demand within periods.

6. Acknowledgement

Financial support from the Danish Social Science Research Council is gratefully acknowledged.

Appendix A. Proofs

Introduce first some notation in order to keep equations short and easier to follow.

$$\Phi_t = \Phi(s(t)), \Gamma_t = \int_{s(t)}^{\overline{s}} s\phi(s)ds, \Psi_t = \int_{s(t)}^{\overline{s}} s^2\phi(s)ds$$

$$\mu_t = \mu(t), \mu'_t = \mu'(t), R_t = R(t), \rho_t = \rho(t)$$

Subscripts t will be dropped when it is possible without ambiguity. It will be useful that Jensen's inequality implies the following lemma, stated without proof.

Lemma 4

$$\Psi > \frac{\Gamma^2}{1-\Phi}.$$

Appendix A.1. Proof of Proposition 1

Rewrite the derivative of μ as $\mu' = \rho\Gamma - (1-\Phi)$. Then the derivative of σ may be rewritten as follows.

$$\begin{aligned}\frac{\sigma'}{2} &= \int_{s(t)}^{\overline{s}} (sR_t - t)(s\rho_t - 1)\phi(s)ds - \mu\mu' \\ &= R\rho\Psi - R\Gamma - t\mu' - \mu\mu'.\end{aligned}$$

Now use lemma 4 on the first term

$$\begin{aligned}&> R\rho\frac{\Gamma^2}{1-\Phi} - R\Gamma - t\mu' - \mu\mu' \\ &= R\Gamma\left[\frac{\rho\Gamma - (1-\Phi)}{1-\Phi}\right] - t\mu' - \mu\mu'.\end{aligned}$$

At this point it is useful to insert an expression for μ' and collect terms.

$$\begin{aligned}
&= R\Gamma\left[\frac{\mu'}{1-\Phi}\right] - t\mu' - \mu\mu' \\
&= \frac{\mu'}{1-\Phi}\left[R\Gamma - t(1-\Phi) - R\Gamma(1-\Phi) + t(1-\Phi)^2\right] \\
&= \mu'\frac{\Phi}{1-\Phi}\left[R\Gamma - t(1-\Phi)\right] \\
&= \mu'\mu\frac{\Phi}{1-\Phi}.
\end{aligned}$$

□

Appendix A.2. Proof of proposition 3

This proof is again an application of lemma 4, this time to σ rather than to σ'. Otherwise it is straight-forward calculus.

$$\begin{aligned}
\sigma &= R^2(\Psi - \Gamma^2) - t\Phi(R\Gamma - t(1-\Phi)) - R\Gamma t\Phi \\
&> R^2\left(\frac{\Gamma^2}{1-\Phi} - \Gamma^2\right) - t\Phi\mu - R\Gamma t\Phi \\
&= R^2\Gamma^2\frac{\Phi}{1-\Phi} - t\Phi\mu - R\Gamma t\Phi \\
&= \mu^2\frac{\Phi}{1-\Phi} + t^2\Phi(1-\Phi) + \mu t\Phi - R\Gamma t\Phi \\
&= \mu^2\frac{\Phi}{1-\Phi} + t^2\Phi(1-\Phi) - t^2\Phi(1-\Phi) \\
&= \mu^2\frac{\Phi}{1-\Phi}
\end{aligned}$$

as required. □

References

Arnott, R. A., de Palma, A., Lindsey, R., 1993. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. American Economic Review 83 (1), 161–179.

Arnott, R. A., de Palma, A., Lindsey, R., 1999. Information and time-of-usage decisions in the bottleneck model with stochastic capacity and demand. European Economic Review 43 (3), 525–548.

Department for Transport, 2006. Frameworks for Modelling the Variability of Journey Times on the Highway Network. Department for Transport.

Fosgerau, M., Hjorth, K., 2008. The value of travel time variability for a scheduled service. Proceedings of the European Transport Conference.

Fosgerau, M., Karlstrom, A., 2009. The value of reliability. Transportation Research Part B: Methodological In press.

Vickrey, W. S., 1969. Congestion theory and transport investment. American Economic Review 59 (2), 251–261.

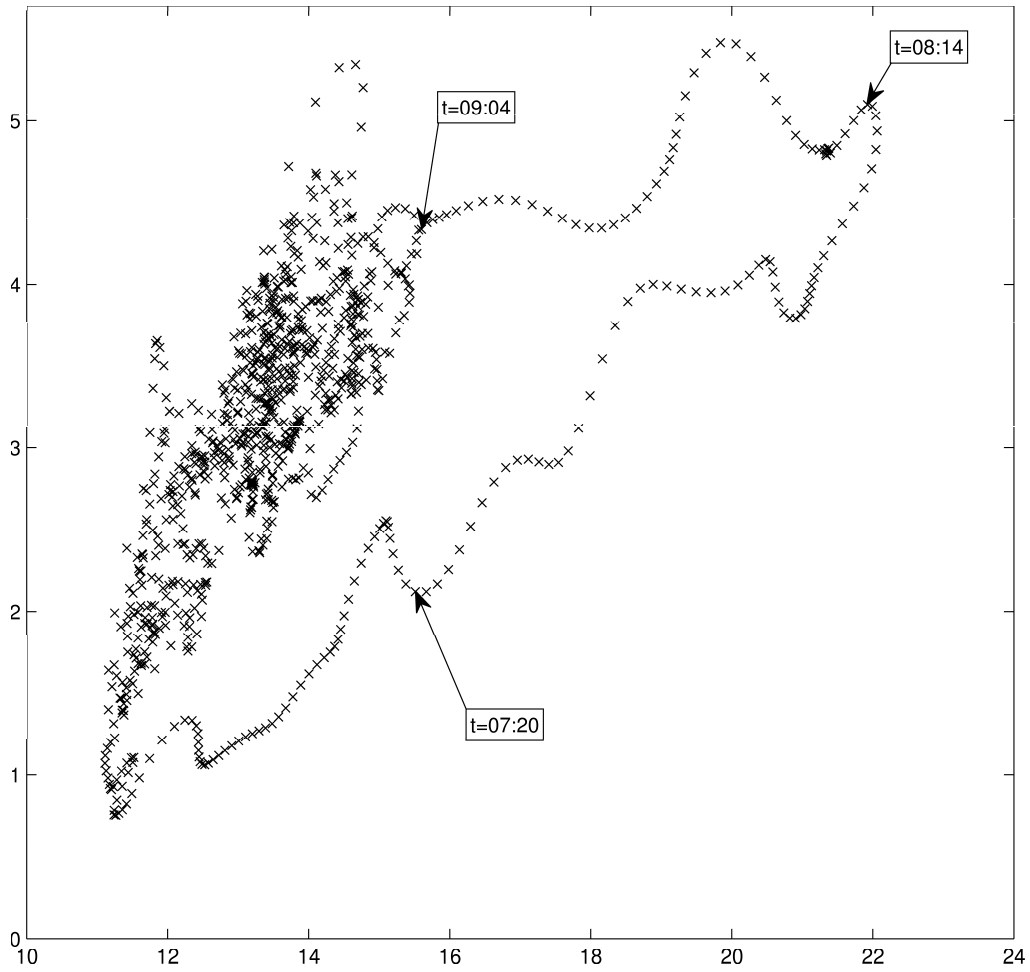

Figure A.1: Mean and standard deviation of travel time (in minutes) for a congested urban road

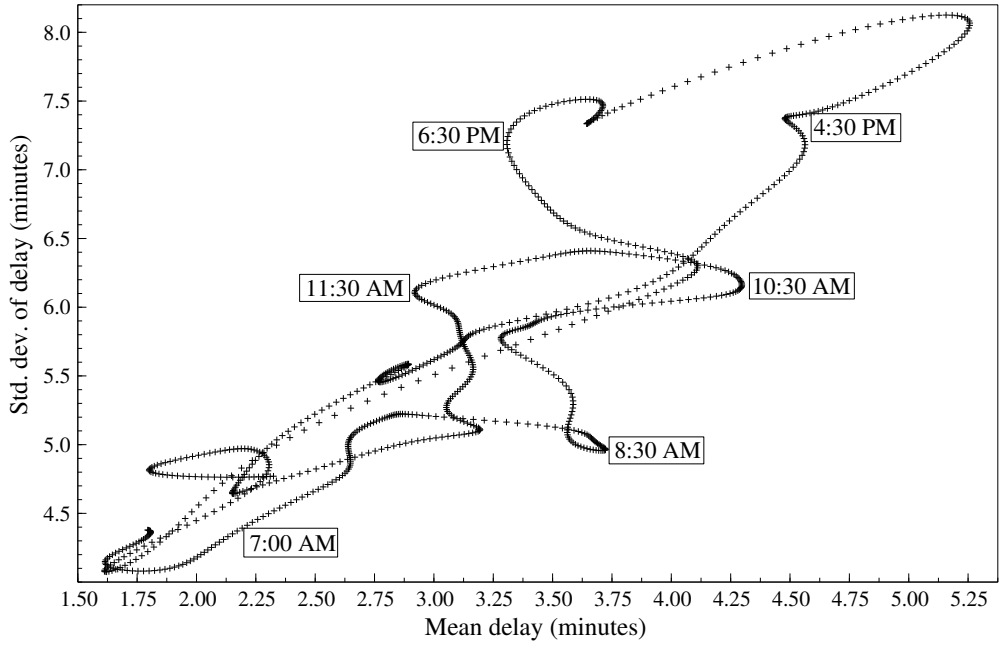

Figure A.2: Mean and standard deviation of travel time (in minutes) for a congested rail section