Predicting queue variability to enable analysis of overload risk

Nicholas B. Taylor[1]

Centre for Transport Studies, University College London, Gower Street, London WC1E 6BT

Abstract

Predicting the risk of traffic demands and delays exceeding critical limits at road junctions, airports, hospitals, etc, requires knowing how both mean and variance of queue size vary over time. Microscopic simulation can explore variability but is computationally demanding and gives only sample results. A computationally efficient approximation to the mean is used in many modelling tools, but only empirical extensions for variance in particular situations have been available. The paper derives theoretical formulae for time-dependent and equilibrium variance, believed to be novel and to apply generally to queues covered by the Pollaczek-Khinchin mean formula, and offering possible structural insights. These are applied in an extended approximation giving mutually consistent mean and variance estimates with improved accuracy. Tests on oversaturated peak demand cases are compared with Markov probabilistic simulation, demonstrating accuracy (R^2>0.99) for typical random, priority-like (M/M/1) and traffic-signal-like (M/D/1) queues. Implications for risk analysis, planning and policy are considered.

Keywords: traffic, modelling, queue, variance, uncertainty, overload, risk

1. Introduction

Realistic modelling of traffic, whether of people, vehicles or other discrete units, needs to account for time variation of demand and transient overloading of capacity, for which steady-state equilibrium theory is inadequate and can be misleading. In modelling road traffic in particular, but also relevant to other services with time-variable demands such as airports and other borders, and hospitals including A&E/ER units, there is a need for a computationally efficient general method for predicting time-dependent queue behaviour. This becomes acute where there is risk of overloading with disproportionate consequences like gridlock, flights missed or travel disrupted, and emergency patients left on trolleys.

The paper derives theoretical formulae for time-dependent and equilibrium queue mean and variance, as initially proposed by the author in Taylor (2005, 2014) respectively, which are believed to be novel, and are compatible with and as general as the Pollaczek-Khinchin mean formula. It uses these to construct an extended computationally efficient method for estimating mutually consistent mean and variance as functions of time. Results are verified by Markov-chain simulations, programmed by the author assisted by N H Spencer at TRL, using test cases including oversaturated peak and random demand profiles, for two common and generic queue processes with random arrivals: priority-like M/M/1 with random service; and traffic signal-like M/D/1 with uniform service[2]. Finally, some implications for analysis of risk, planning and policy are discussed.

2. Brief history of queuing theory and modelling

Queuing theory dates back to Erlang's (1909) work on telegraphy, and many standard works on the subject are available, but their treatment of transient behaviour tends to be

[1] Corresponding author email: nicholas.b.taylor@alumni.ucl.ac.uk

[2] In Kendall's (1951) nomenclature, 'M' = Markovian (memoryless), 'D' = 'Deterministic', and 'G' = 'General'.

restricted to particular cases. Microscopic 'Monte-Carlo' simulation can in principle represent any random process, but simulating thousands of events may be computationally demanding, and as queues tend to be highly variable even under identical average conditions, many randomised simulation runs are usually needed to get a spread of representative results. Exact formulae like those of Morse (1955, 1958), Clarke (1956), Sharma (1990), Heidemann (1994) and Griffiths *et al* (2005) can be used to develop dynamic probability distributions for some queue processes, but may require evaluating complicated functions and sums of series. Markov-chain simulation, which develops a distribution in small steps using state transition or recurrence relationships, can handle a wider range of processes, see Olszewski (1990), Viti and van Zuylen (2010), but is still computationally demanding so most suitable for benchmarking.

Most queue processes conserve 'customers', giving rise to a 'deterministic' time-dependent formula for mean queue size. However, this cannot predict the process-specific equilibrium mean size to which some queues eventually converge under steady-state conditions. An expression for this is credited independently to Félix Pollaczek, Aleksandr Khinchin (P-K) and Harald Cramér in 1930-32. Equilibrium mean formulae exist for several processes, but actual size at any time is highly variable. Equilibrium probability distributions of queue size are sometimes adopted for convenience, but simulations show that distributions in realistic cases differ substantially from equilibrium, see e.g. Whitt (1983), Kimber and Daly (1986). Soros (1987) points out that: "*The concept of an equilibrium is very useful. It allows us to focus on the final outcome rather than on the process that leads up to it. But … equilibrium has rarely been observed in real life*".

An heuristic 'coordinate-transformed' or 'sheared' approximation to mean queue development, sometimes ascribed to P D Whiting at TRL, and described by Kimber and Hollis (1979), has been used extensively in traffic modelling and junction design software as described by Robertson (1969), Leonard *et al* (1978, 1989), van Vliet (1982), Semmens (1985), Taylor (2003), Binning (2004), TRL (2015), and is cited in design literature such as HCM (2010/2016). Variations have been described by Newell (1971), Doherty (1977), Catling (1977) and Akçelik (1980). The resulting method is computationally efficient, rational and seamless through saturation, However, it cannot account for day-to-day variability enabling reliable estimation of the risk of overloading, and must therefore be considered incomplete. Semi-empirical methods for estimating queue variance are described by Kimber, Daly *et al* (1986), Fu and Hellinger (2000), and Viti and van Zuylen (2010), but are restricted to particular queue processes.

3. Theoretical development of the mean queue

3.1 Time-dependent mean properties

Equation (1) gives the 'deterministic' mean size of a queue $L(t)$ at time t, reflecting conservation of customers/units in $[0,t]$, where demand intensity[3] ρ is the ratio of mean arrival rate λ to service rate or capacity μ, both assumed constant in $[0,t]$.

$$L(t) = L_0 + (\rho - x(t))\mu t \qquad \text{where} \qquad \rho = \frac{\lambda}{\mu}, \quad x(t) = \frac{1}{t}\int_0^t u(y)dy \tag{1}$$

The utilisation of service u can vary, as can its time-average x, but both converge asymptotically to the equilibrium value ρ as $t \to \infty$, provided that $\rho<1$. Differentiating (1) gives the rate of change of the mean queue, equation (2), which relates utilisation of service or the proportion of time service is occupied u to the average probability $\bar{p}_0$ of the

[3] The term 'traffic intensity' is ambiguous in a time-varying context unless the point of measurement is specified.

queue being zero over a service period, subscript *e* indicating an equilibrium value.

$$\dot{L}(t) = (\rho - u(t))\mu = ((1 - \bar{p}_{0e}) - (1 - \bar{p}_0(t)))\mu \qquad (2)$$

Parameters λ, μ, ρ can be generalised to time-dependent functions, but it is convenient to assume they are piecewise constant. Olszewski (1990), Heydecker and Verlander (1998) and Addison and Heydecker (2006) point out that this leads to results that are non-transitive, i.e. depend on how time is 'sliced'. However, time can be resolved in as short periods as desired or indicated by data that are often aggregated in finite time periods. One would expect correct representation of probability distributions to ensure transitivity.

3.2 Evaluation of queue moments and characterisation of service processes

Standard works often obtain the equilibrium mean of a specific queue process by substituting the appropriate Laplace transform function of arrival and service processes into the general transform formula due to Pollaczek and Khinchin (e.g. Medhi, 2003). Queue moments can be evaluated directly from recurrence relations if available. The Imbedded Markov (IM) method, which treats a queue as a discrete process and evaluates its state only at the end of each service period, can in principle deliver explicit general formulae for all moments of a family of queue processes. However, as discrete event treatment does not involve time explicitly, issues can arise in deriving continuous time-dependent functions for moments higher than the mean that involve non-linearity. Evaluation at equilibrium is unaffected because it represents an ergodic condition. While all a queue's properties are necessarily implicit in its formulation, in the sense discussed generally by Machta *et al* (2013), an exact procedure giving a *general* formula like (1) cannot also yield the equilibrium result for a *specific* process. Table 1 expresses the rules for obtaining moments of a queue size probability distribution $\{p_i\}$.

Table 1. Rules for evaluating queue moments

Evaluate distribution	**Equilibrium result**	**Time-dependent formula**
1st moment: $\left(\sum ip_i\right)$	P(empty queue) $\bar{p}_{0e}$	Mean $L(t)$
2nd raw moment $\left(\sum i^2 p_i\right)$	Mean L_e	Variance $V(t)$
3rd raw moment $\left(\sum i^3 p_i\right)$	Variance V_e	Skewness (in principle)

A distinction is drawn between queue size including or excluding the customer/unit in service. This does not affect the M/M/1 process whose service is random, but does affect the M/D/1 process where, for example, physical service time at saturation rate at a traffic signal is typically much less than average service time allowing for the red/green cycle. Hence the unit-in-service is excluded from the queue at a signal to first approximation.

3.3 Imbedded Markov derivation of the Pollaczek-Khinchin equilibrium mean

The Imbedded Markov model, including unit-in-service (i.u.i.s), is given by the first of equations (3), see e.g. Bunday (1996) after Kleinrock (1975), being similar to Lindley's (1952) waiting time formula. The second equation applies without unit-in-service (n.u.i.s).

$$q_{n+1} = q_n - U(q_n) + \zeta_{n+1} \text{ (i.u.i.s)}, \qquad \tilde{q}_{n+1} = \tilde{q}_n - U(q_{n+1}) + \zeta_{n+1} \text{ (n.u.i.s) where} \qquad (3)$$

q_n = number of units (customers) in the system at end of the service period of unit n
ζ_n= number of units arriving during the service time of unit n
$U(q_n)$ = 1 if unit remains in system (q_n>1) and will be served in period n+1, otherwise 0
$\tilde{q}_n = q_n - U(q_n)$ represents queue size excluding the unit-in-service (n.u.i.s).

Intermediate working results, noting that ζ_{n+1} is independent of q_n and $U(q_n)$, include:

$$[U(q_n)]^m = U(q_n) \forall m \neq 0,\ q_n U(q_n) = q_n,\ E(\zeta_n) = \rho,\ E(q_n) = L,\ E(U(q_n)) = u,\ E[b] = 1/\mu$$
$$\lambda E[b] = \rho, \quad E(\zeta_{n+1} q_n) = E(\zeta_{n+1}) E(q_n), \quad E(\zeta_{n+1} U(q_n)) = E(\zeta_{n+1}) E(U(q_n)) \qquad (4)$$

Equations (1,3) are essentially equivalent statements of conservation. To obtain the equilibrium mean, the first of equations (3) is squared and relevant equations (4) applied:

$$q_{n+1}^2 = q_n^2 + U(q_n)^2 + \zeta_{n+1}^2 + 2q_n\zeta_{n+1} - 2\zeta_{n+1}U(q_n) - 2q_n U(q_n)$$
$$= q_n^2 + 2(\zeta_{n+1} - 1)q_n - (2\zeta_{n+1} - 1)U(q_n) + \zeta_{n+1}^2 \qquad (5)$$

The expectation of squared arrivals in a service period involves the dispersion index of arrivals I_a (personal communication by B G Heydecker) and the coefficient of variation c_b of the service distribution whose probability generating function is $b(s)$:

$$E(\zeta_{n+1}^2) = \int_0^\infty \left(I_a \lambda s + (\lambda s)^2\right) b(s)\, ds = I_a \lambda E[b] + \lambda^2 \left(\mathrm{var}[b] + E[b]^2\right)$$
$$= I_a \rho + \rho^2 + \lambda^2\, \mathrm{var}[b] = I_a \rho + 2\rho^2 C \quad \text{where} \quad C \equiv \tfrac{1}{2}\left(1 + c_b^2\right) \qquad (6)$$

At steady-state equilibrium the q^2 terms cancel by definition. Writing the equilibrium mean queue $E(q_n)$ as L_e, the expectation of (5) gives the P-K mean formula:

$$2(\rho - 1)L_e - 2\rho^2 + \rho + \left(I_a \rho + 2\rho^2 C\right) = 0 \quad \text{rearranged as} \qquad (7)$$

$$L_e(\rho) = I\rho + \frac{(I_a - 1)\rho}{2(1-\rho)} + \frac{C\rho^2}{1-\rho} \qquad (8)$$

The index I in (8), introduced by Kimber and Hollis (1979) with I Summersgill, represents contribution of the unit-in-service (value 0 or 1). When the second of equations (3) is used, the term $-2\rho^2$ in (7) is replaced by -2ρ, suppressing the term $I\rho$ consistent with the interpretation. Some statistical parameters and equilibrium means are given in Table 2.

Table 2. Statistical parameters and equilibrium means of common queue types

Process	**With u.i.s**	I	I_a	c_b	C	p_{0e} (instantaneous)	**Mean** L_e
M/M/1	Yes	1	1	1	1	$p_{0e} = 1 - \rho$	$L_e = \dfrac{\rho}{1-\rho}$
M/D/1	No	0	1	0	$\frac{1}{2}$	$p_{0e} = e^{\rho}(1-\rho)$	$L_e = \dfrac{\rho^2}{2(1-\rho)}$
M/D/1	Yes	1	1	0	$\frac{1}{2}$	$p_{0e} = 1 - \rho$	$L_e = \rho + \dfrac{\rho^2}{2(1-\rho)}$

The equilibrium probability of zero queue, p_{0e}, refers to the instantaneous probability distribution at the *end* of a service period. The *average* over a service period must be $(1-\rho)$ to ensure steady state, where arrivals and utilisation both equal ρ. The M/D/1 (n.u.i.s) mean matches a stochastic component in Webster and Cobbe's (1966) signal delay formula, although this is supplemented by an empirical term related to capacity in the green period. Some standard works, e.g. Medhi (2003), derive the M/D/1 (i.u.i.s) form of mean expression, but the extra term ρ does not reflect actual delay. Burrow (1987) recommends C=0.6 in place of C=0.5 in practice. These issues are discussed further by Taylor and Heydecker (2014). In time-dependent modelling capacity may change while a customer is in the queue, so individual delays are obtained by piecewise forward-looking

application of Little's (1961) formula, in which delay is given by queue size divided by throughput capacity, so queue size and delay need compatible definitions (Taylor, 2003).

4. Derivation of theoretical formulae for queue variance

4.1 M/M/1 time-dependent variance from recurrence relations

Full derivations are given here and subsequently for the sake of completeness and verifiability. Time-dependent variance is obtained from recurrence relations as described by Taylor (2005). Addison and Heydecker (2006) derive its time derivative by a similar method. Differential recurrence relations for M/M/1 probabilities at constant arrival and service rates λ and μ, with a 'barrier' at zero queue size, are given by:

$$\dot{p}_0 = \mu p_1 - \lambda p_0, \qquad \dot{p}_i = \mu p_{i+1} - (\mu+\lambda) p_i + \lambda p_{i-1} \quad (i>0) \tag{9}$$

Recalling $\rho = \lambda/\mu$, the rate of change of the raw second moment, M_2, is given by equation (10). Terms in M_2 itself cancel giving equation (11), which integrates to give (12). Here D is the time-averaged mean queue size over the period [0,t], which can be interpreted loosely as delay per unit time. The general behaviour of D is similar to that of L with a time lag, and it converges to the same equilibrium value L_e.

$$\begin{aligned}\dot{M}_2 &= \sum_{i=1}^{\infty} i^2 \dot{p}_i = \mu \sum_{i=1}^{\infty} i^2 p_{i+1} - (\mu+\lambda) \sum_{i=1}^{\infty} i^2 p_i + \lambda \sum_{i=1}^{\infty} i^2 p_{i-1} = \mu \sum_{i=1}^{\infty} (i+1)^2 p_{i+1} \\ &- 2\mu \sum_{i=1}^{\infty} (i+1) p_{i+1} + \mu \sum_{i=1}^{\infty} p_{i+1} - (\mu+\lambda) \sum_{i=1}^{\infty} i^2 p_i + \lambda \sum_{i=1}^{\infty} (i-1)^2 p_{i-1} + 2\lambda \sum_{i=1}^{\infty} (i-1) p_{i-1} + \lambda \sum_{i=1}^{\infty} p_{i-1} \\ &= \mu(M_2 - p_1) - 2\mu(L - p_1) + \mu(1 - p_1 - p_0) - (\mu+\lambda) M_2 + \lambda M_2 + 2\lambda L + \lambda \end{aligned} \tag{10}$$

$$\dot{M}_2 = \left(-2(1-\rho)L + 1 + \rho - p_0\right)\mu = \left(2\rho - 2(1-\rho)L\right)\mu - \dot{L} \quad \text{integrating to:} \tag{11}$$

$$M_2 \equiv V + L^2 = 2\rho\mu t - 2(1-\rho) D \mu t - L + \text{constant} \quad \text{where } D(t) = \frac{1}{t}\int_0^t L(y)\,dy \tag{12}$$

On rearranging (12), inserting the initial state, and substituting L_e for the mean expression in Table 2, the time-dependent variance formula equation (13) is obtained, which in the oversaturated limit (x=1) reduces to the familiar 'deterministic' formula (14).

$$V(t) = V_0 + L_0(L_0+1) + 2(1-\rho)\left(L_e - D(t)\right)\mu t - L(L+1) \tag{13}$$

$$V(t) \to V_0 + (\lambda+\mu)t \qquad \text{in the deterministic limit} \tag{14}$$

4.2 M/D/1 (n.u.i.s) time-dependent variance from recurrence relations

Taylor and Heydecker (2014) derive finite-differential recurrence relations for M/D/1 on a notional uniform service time interval $1/\mu$ and define notional continuous-time derivatives, as in equations (15), leading to rates of change of moments (16,17). Only derivatives of $\{p_i\}$ for i>0 contribute to rates of change, but p_0 itself appears on the RHS.

$$p_i(\mu t + 1) = \sum_{j=0}^{i+1} \frac{\rho^j e^{-\rho}}{j!} p_{i+1-j}(\mu t), \qquad \dot{p}_i \approx \mu \Delta p_i = \mu\left(p_i(\mu t+1) - p_i(\mu t)\right) \tag{15}$$

$$\frac{e^{\rho}}{\mu}\dot{L} = \sum_{i=0}^{\infty}\left[\left(\sum_{j=0}^{\infty}(i+j-1)\frac{\rho^j}{j!} - ie^{\rho}\right)p_i\right] + p_0 \quad \text{so} \quad \dot{L} = \left(\rho - 1 + e^{-\rho} p_0\right)\mu \tag{16}$$

$$\frac{e^{\rho}}{\mu}\dot{M}_2 = \sum_{i=0}^{\infty}\left[\left(\sum_{j=0}^{\infty}(i+j-1)^2\frac{\rho^j}{j!} - i^2e^{\rho}\right)p_i\right] - p_0 \qquad \text{so}$$

$$\dot{M}_2 = \left(\rho^2 + 2\rho L - \rho - 2L + 1 - e^{-\rho}p_0\right)\mu = \left(\rho^2 - 2(1-\rho)L\right)\mu - \dot{L} \qquad (17)$$

Equation (17) differs from the M/M/1 equivalent (11) only in the first term on the RHS, corresponding to the difference between M/M/1 and M/D/1 equilibrium means L_e in Table 2, so integration again gives (13). This together with the symmetrical structure of the formula encourages belief in its generality.

4.3 Derivation of expression for equilibrium variance

In accordance with Table 1, Imbedded Markov evaluation of equilibrium variance requires cubing the first of equations (3), discrete versus continuous time not being an issue here:

$$q_{n+1}^3 = q_n^3 + q_nU(q_n) + q_n\zeta_{n+1}^2 + 2q_n^2\zeta_{n+1} - 2\zeta_{n+1}q_nU(q_n) - 2q_n^2$$
$$- q_n^2U(q_n) - U(q_n)^2 - \zeta_{n+1}^2U(q_n) - 2q_nU(q_n)\zeta_{n+1} + 2\zeta_{n+1}U(q_n)^2 + 2q_nU(q_n)$$
$$+ q_n^2\zeta_{n+1} + U(q_n)\zeta_{n+1} + \zeta_{n+1}^3 + 2q_n\zeta_{n+1}^2 - 2\zeta_{n+1}^2U(q_n) - 2q_n\zeta_{n+1} \qquad \text{so, using (4)}$$
$$q_{n+1}^3 = q_n^3 + 3(\zeta_{n+1}-1)q_n^2 + 3\left(\zeta_{n+1}^2 - 2\zeta_{n+1} + 1\right)q_n - 3\left(\zeta_{n+1}^2 - \zeta_{n+1} + \tfrac{1}{3}\right)U(q_n) + \zeta_{n+1}^3 \qquad (18)$$

Taking expectations at equilibrium, where q^3 terms cancel and u=ρ, after rearranging:

$$3(1-\rho)E(q_n^2) = 3\rho^2 - \rho + 3(1-2\rho)L_e + 3(L_e-\rho)E(\zeta_{n+1}^2) + E(\zeta_{n+1}^3) \qquad (19)$$

The expectation of ζ_{n+1}^3 involves the skewness of the arrival and service distributions. These are not standard P-K parameters but are allowed for here by extra terms, assuming Poisson processes (Willmot, 1986), although the validity of arrivals dispersion I_a appears to require only independence of service, giving equations (20) with ψ=1 by default:

$$E(\zeta_n^3) = M_3(\zeta_n) + 3E(\zeta_n^2)E(\zeta_n) - 2E(\zeta_n)^3 = \int_0^{\infty}\left[J_a\lambda s + 3I_a(\lambda s)^2 + (\lambda s)^3\right]b(s)ds$$
$$= J_a\rho + 6I_aC\rho^2 + 6(C-J_b)\rho^3 \quad \text{where} \quad J_a = I_a^2 + I_a - 1, \quad J_b = \tfrac{1}{3}(1-\psi c_b^3) \qquad (20)$$

Substituting (6,20) into (19) leads to a general expression for equilibrium variance (21), where the unit-in-service parameter I has been introduced semi-empirically by comparing M/D/1 values with and without unit-in-service for various values of ρ.

$$V_e = I\rho((2C-1)\rho+1) + \frac{(I_a-1)\left(L_e + \tfrac{1}{3}(I_a-1)\rho + 2C\rho^2\right)}{(1-\rho)} + \frac{C\rho^2\left(1+\rho+(C-2)\rho^2\right)}{(1-\rho)^2} - \frac{2J_b\rho^3}{(1-\rho)} \qquad (21)$$

Table 2 can now be extended to include equilibrium variance formulae as Table 3.

Table 3. Equilibrium moments of common queue types

Process	I	J_a	J_b	p_{0e}	**Mean** L_e	**Variance** V_e
M/M/1	1	1	0	$p_{0e} = 1-\rho$	$L_e = \frac{\rho}{1-\rho}$	$V_e = \frac{\rho}{(1-\rho)^2}$
M/D/1 (n.u.i.s)	0	1	$\frac{1}{3}$	$p_{0e} = e^{\rho}(1-\rho)$	$L_e = \frac{\rho^2}{2(1-\rho)}$	$V_e = \frac{\rho^2(6-2\rho-\rho^2)}{12(1-\rho)^2}$
M/D/1 (i.u.i.s)	1	1	$\frac{1}{3}$	$p_{0e} = 1-\rho$	$L_e = \rho + \frac{\rho^2}{2(1-\rho)}$	$V_e = \frac{\rho(12-18\rho+10\rho^2-\rho^3)}{12(1-\rho)^2}$

4.4 Generality of variance results and general form of moments

Results obtained from Imbedded Markov should be more general than those from recurrence relations. However, variance obtained from IM exceeds the continuous time result (13) by $2(L\text{-}L_0)\rho$. There seems to be no simple way around this, but it can be resolved by considering the service interval shrunk to infinitesimal size, in effect to continuous time. As the 'unwanted' term $\zeta_{n+1}U(q_n)$ is the only one involving the interaction of two independent variables, it should vanish to first order. However, variability and non-ergodicity make it practically impossible to verify the result exactly by aggregating many simulation runs. Equation (13) resembles (1) in structure, although there is no obvious conserved quantity like number of customers. In both, process dependence is only implicit, through ρ and $L_e(\rho)$. The same result for two different processes supports generality, and a similar formula is obtainable from a simple dispersion model. It is inferred that the mth time-dependent moment, M_m, should satisfy an equation of the following general form, where f_m is some function of moments $\leq m$ and statistical parameters $\{c_i\}$:

$$f_m[(M_m...M_1,\{c_i\})](t) = f_m[(M_m...M_1,\{c_i\})](0) + m(1-\rho)^{m-1}\left(M_{(m-1)e}(\rho) - \overline{M}_{(m-1)}(t)\right)\mu t \qquad (22)$$

Like (1,13) this appears to conserve something at least statistically, and leaves the corresponding equilibrium result undetermined. The factor involving $(1\text{-}\rho)$ ensures that the RHS remains valid and finite when $\rho \geq 1$. Calculating equilibrium variance (21) using approximate skewness parameters (20) is sufficient because, in contrast to the P-K mean whose algebraic *form* matters, only the *numerical value* of the equilibrium variance is needed to constrain queue development. Furthermore, the complexity of explicit time-dependent and equilibrium formulae must rise with increasing m, and this is likely to make obtaining explicit formulae for higher moments difficult or impractical.

4.5 Hysteresis in queue development

Hysteresis in queues is observed in road networks as described by Arup, Bates et al (2004), Addison (2006) and Fosgerau (2008), and shown by simulation in Figure 1. Time lags arise naturally in the relationships between mean queue and demand, and between variance and mean. A queue will continue to grow as long as demand exceeds capacity, even if demand is falling. Likewise, variance can increase even when a queue is decaying. This is evident when the rate of change of variance is expressed in balanced form:

$$\dot{V} = 2\mu(1-\rho)(L_e + .5)\left[1 - \left(\frac{L+.5}{L_e+.5}\right)\left(\frac{1-u}{1-\rho}\right)\right] \qquad (23)$$

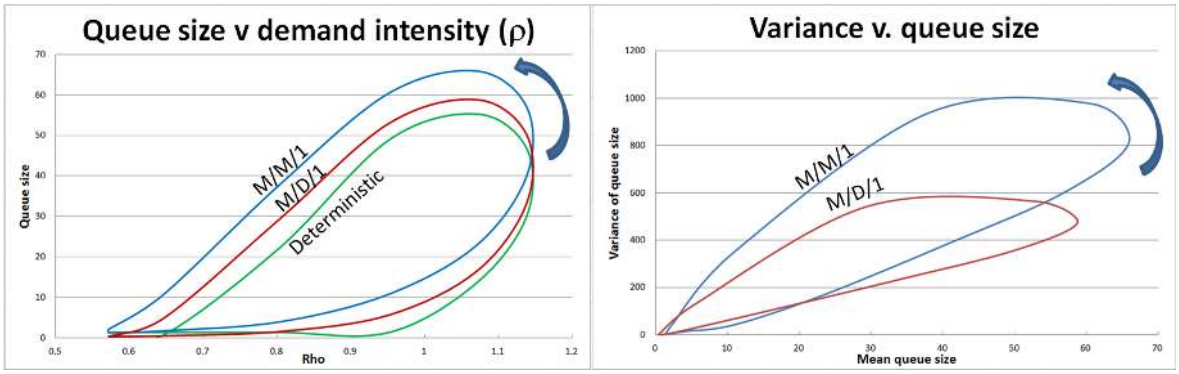

Figure 1. Hysteresis between mean and demand (left) and variance and mean (right)

5. Results needed for estimating probability distributions and model verification

Examples exist of probability distributions with different shapes but the same mean and variance. Taylor and Heydecker (2015) find that a minimum of three moments is needed to represent a queue size probability distribution. However, involving skewness explicitly would be impractical, as argued above. The 'barrier' at zero size induces interdependence between queue moments and the probability of zero queue p_0, which as the complement of utilisation of service is intimately related to the dynamics of the queue and is much easier to work with than skewness. Accordingly, dynamic probability distributions can be estimated from p_0 (styled as the '0th moment'), mean and variance. Estimating them requires an iterative Newton method, although an explicit doubly-nested geometric formula, also the maximum entropy form, can be used in a few cases.

Queue size probability distributions can be developed by Markov simulation of recurrence relations in small time steps, e.g. 0.1-1.0 second, and moments calculated directly from these distributions to verify model estimates. For reassurance, the exact M/M/1 series formula of Morse (1955, 1958) can be used to verify M/M/1 Markov simulations, and Monte-Carlo simulation to check both (Taylor, 2005, 2014). Test cases used consist of 34 transiently oversaturated Gaussian peak profiles of various lengths and intensities (maximum demand intensity ρ=1.1458), originally defined by Kimber *et al* (1986), three persistently oversaturated cases (maximum ρ=1.3), and three randomly generated profiles (maximum ρ=1.4911). These are divided into time slices of a few minutes (4.5-12 minutes for the peak cases) with demand intensity ρ and capacity μ assumed to be piecewise constant. Both M/M/1 and M/D/1 processes are evaluated.

6. Time-dependent approximation to the mean queue

6.1 Coordinate-transformed or sheared method

The series formulae of Morse (1955, 1958) for M/M/1 queue size probabilities show that time development of even the simplest queue is complex, but equations (1,8) constrain the behaviour of the mean. Kimber and Hollis (1979) combine them to create a 'coordinate transformed' model, analogous to a conical volume-delay function (Spiess, 1989). Figure 2 shows graphically how this 'shears' the static equilibrium function (8) (left) into the time-dependent asymptote (1) (right) that rotates anti-clockwise as time proceeds, giving a smooth function that evolves with time (dashed curve). In effect this assumes 'quasi-equilibrium', where the queue at time t is equated to the equilibrium value that *would* result from a traffic intensity equal to x *at the stop line*, so replacing ρ in (8) with x. This results in a quadratic in queue size, equation (24), with solution (25).

$$L_s(t) \equiv L_e(x(t)) \quad <==> \quad FL_s^2 + GL_s + H = 0 \quad \text{whose solution is} \qquad (24)$$

$$L_s(t) = \frac{G + \sqrt{G^2 - 4FH}}{2F} \quad (F \neq 0), \qquad L_s(t) = \frac{H}{G} \quad (F=0,\ G \neq 0) \qquad \text{where}$$

$$F = \mu t - (C - I), \qquad G = (L_0 - I^*)\mu t - 2(C - I)(L_0 + \rho\mu t) - (1-\rho)(\mu t)^2 \qquad (25)$$

$$H = -\left[(C - I)(L_0 + \rho\mu t) + I^*\mu t\right](L_0 + \rho\mu t) \qquad \text{and} \quad I^* \equiv I + \tfrac{1}{2}(I_a - 1)$$

The initial queue L_0 from (1) enters smoothly in (25) but causes an awkward step in L_d in Figure 2. Kimber and Hollis (1979) avoid this by 'divided' or origin-shifted formulae where the time origin is displaced to $-t_0$, obtained by solving for t after inverting (25) with L_0 set to zero, as expressed by (26), a variation of which is applied to a decaying queue.

$$L_t(0, t + t_0) \equiv L_s(L_0, t) \qquad (26)$$

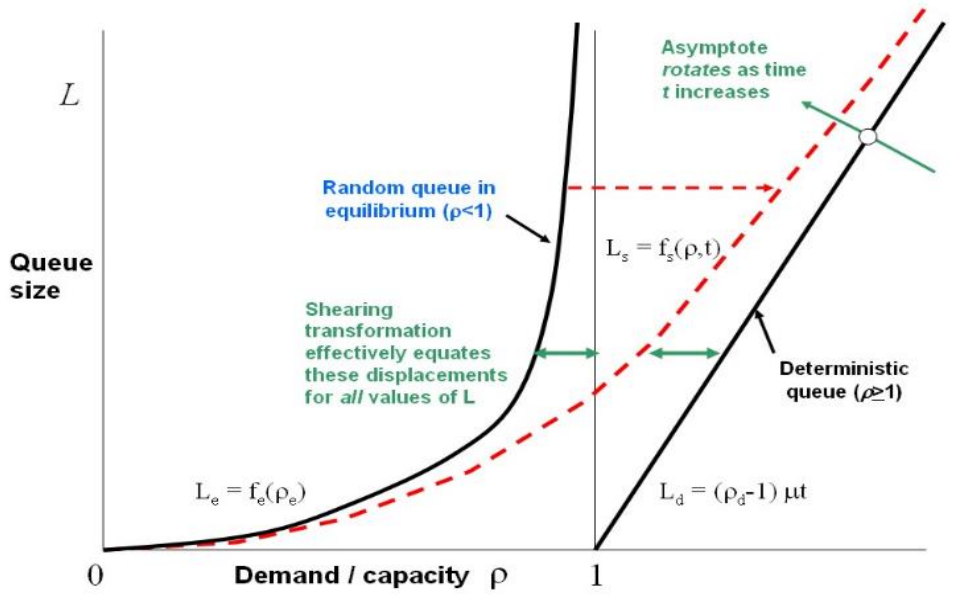

Figure 2. Coordinate-transformed or sheared time-dependent mean queue

The solution has a computationally efficient closed form, free from empirical parameters, with seamless behaviour through saturation, and can accommodate any queue processes represented by the P-K formula, making it convenient for time-dependent network and assignment modelling as referred to earlier.

6.2 Accuracy of the mean approximation and its relevance to variance estimation

Figure 3 plots errors in the mean queue relative to Markov simulation, for undersaturated growth (left), an oversaturated peak case (middle), and variance (right) as discussed below.

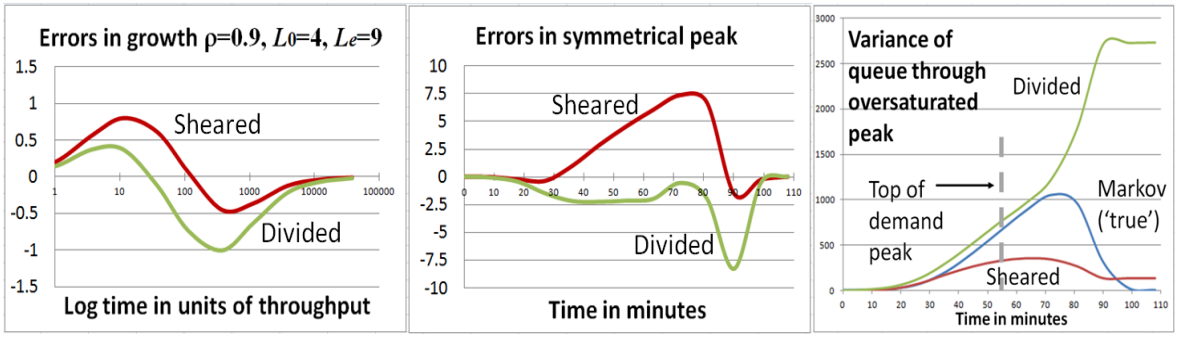

Figure 3. Errors in mean queue and variance estimates compared to Markov simulation

The 'divided' method performs fairly well for oversaturated queue growth, but not decay. As peak queue size is around 65, the errors are moderate in practical terms. It is tempting to ask whether variance can be 'sheared' like the mean. However, Newell (1971) shows this is impractical because the contributions of arrival and service components are additive rather than subtractive, so the graph of time-dependent variance lies outside the equilibrium and deterministic curves. Kimber *et al* (1986) describe a method for estimating variance in oversaturated demand peaks based on regression from simulated Gaussian demand profiles. Fu and Hellinger (2000) calibrate an extension to Webster and Cobbe's (1966) signal delay formula, using parameters regressed from simulation. However, none of these methods can model all processes covered by the P-K formula, or assure consistency between mean and variance estimates. This can be interpreted as any estimate

of the mean queue carrying with it an implicit probability distribution that in general does not reflect real behaviour, reflecting the observation by Olszewski (1990), echoed by Viti and van Zuylen (2004), that queue behaviour is affected by the variance of its distribution.

Thus it is essential to calculate mean and variance in a mutually consistent way. Unfortunately, equation (13) alone is insufficient. The mean solution (25) is naturally well-behaved in the sense of converging to the correct equilibrium mean, but the variance (13) is highly sensitive to D and hence to L, meaning incidentally that evaluating D precisely by integrating (25) according to the definition in (12) would not help. As Figure 3 (right) shows, estimates of variance over a peak, based on either of the sheared forms of the mean $L(t)$, with 'delay' D approximated by $L(t/2)$, are seriously inaccurate or unstable. This implies that an active mechanism is required to ensure consistency of mean and variance.

7. Extending time-dependent approximation to include variance

7.1 Strategy for development of the method

The coordinate-transformed mean queue is the natural starting point because it is quite general, anchored in queuing theory, has convenient computational properties, and is used in many software tools. Objectives are to extend the method so variance is calculated to useful accuracy along with the mean, both being asymptotic to their equilibrium values, and to verify results for M/M/1 and M/D/1 processes against Markov simulation.

As Kimber and Hollis (1979) found with their 'divided queue' approach, the optimum method of approximation can differ between growth and decay regimes. This can be explained by differences in the form and dynamics of the corresponding probability distributions. The present approach exploits the behaviour of the moments in equations (1,2,13) to approximate queue development piecewise through time, mirroring that of the *implicit* queue size probability distributions, as visualised in Figure 4.

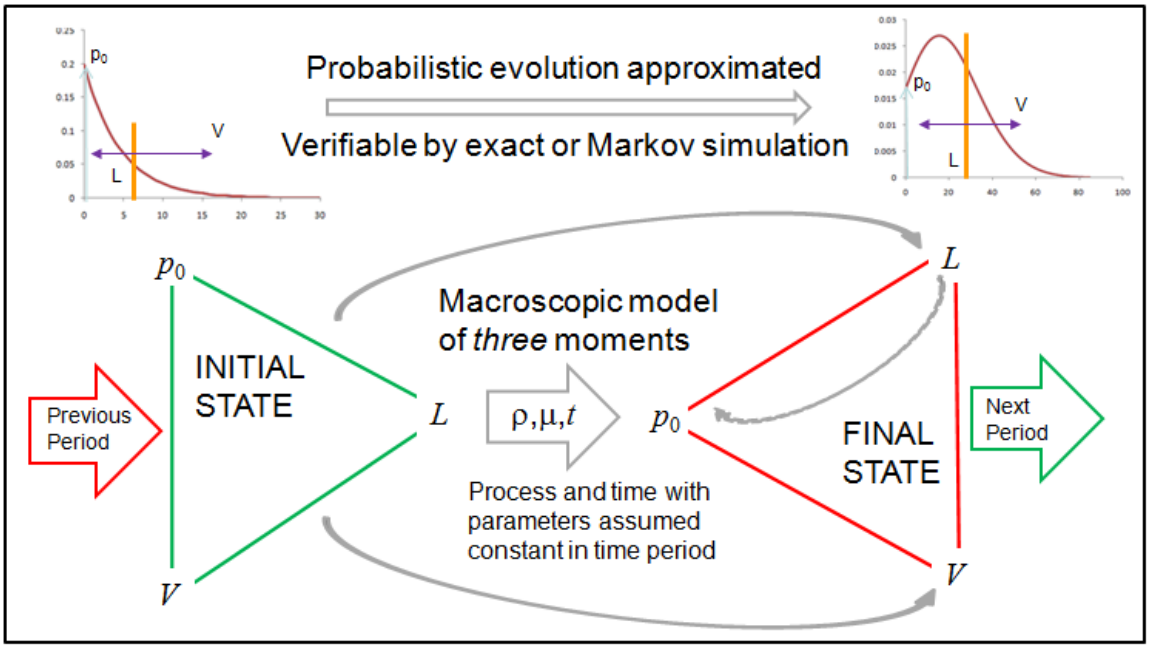

Figure 4. Queue state development over one time period in terms of three moments

Three regimes of queue development are considered separately:

- Undersaturated growth ($\rho<1$), tending towards equilibrium from below
- Oversaturated growth ($\rho\geq1$), with no upper limit
- Decay (undersaturated) ($\rho<1$), tending towards equilibrium from above.

7.2 Undersaturated growth regime

In this regime the 'pressure' of demand intensity is effectively 'resisted' by the degree of saturation, so behaviour near equilibrium is dominated by first-order terms. Because the mean solution (25) is already approximate, variations are allowed if they improve results in a particular context. $D(t)$ can be approximated by $L_s(t/2)$ where the queue is varying linearly or slowly, as in early growth from zero or near equilibrium. However, when substituted into the variance formula (13) with simple initial conditions $L_0=0$, $V_0=0$, $L_s(t)$ turns out to be a better model of D as $t\to\infty$, giving the correct asymptotic rate of change in the mean, although the effect on its absolute value is barely noticeable. This is a hint that transformation of the time variable may be used to correct the mean for consistency with the variance. By Taylor-Maclaurin expansion of the mean formula, expressed in terms of $1/t$, the following asymptotic correction factor to divide into time is obtained, which depends on the *form* of the P-K mean and the extremal *values* of the mean and variance:

$$\omega_\infty = \frac{(1-\rho)(W_e - W_0)}{2(L_e - L_0)(I^* + \rho(2-\rho)(C-I))} \qquad \text{where} \qquad W_x \equiv V_x + L_x(L_x+1) \qquad (27)$$

Undershoot of the mean can occur initially if the initial variance is small, modelling which would require reversing the direction of time. This is avoided by defining D by interpolating $L_s(t)$ and $L_s(t/2)$ as in equation (28). The factor $\Omega(t)$ is itself interpolated between its extremal values using an heuristically determined time function (29) which incorporates demand ρ through a 'link function' based on stochastic relaxation time (30).

$$D(t) = (1-\Omega(t))L_s(t) + \Omega(t)L_s(t/2) \qquad \text{where} \qquad (28)$$

$$\Omega(t) = \Omega_0 + (\Omega_\infty - \Omega_0)\left(\frac{t_1}{t_1+1}\right) \qquad \text{and} \qquad (29)$$

$$t_1 = \frac{\mu t - 1}{\mu\tau_{re} - 1} \qquad \text{where} \qquad \tau_{re} = \mu^{-1}\left(1-\sqrt{\rho}\right)^{-2} \text{ (relaxation time)} \qquad (30)$$

The lower limit on time of $1/\mu$ or one service interval is not restrictive. The link function (30) reflects approximate scale-independence in queue behaviour with change of ρ or μ. The initial value Ω_0 is obtained by differentiating (28,29) and letting $t\to 0$, and reflects the difference between the actual initial utilisation determined by the initial state and that 'predicted' by the mean formula (25). Ω_∞ is turns out to be simply related to ω_∞.

$$\Omega_0 = \frac{(\rho - u_0)\mu - 2L_s'(t)\big|_{t\to 0}}{L_s'(t/2)\big|_{t\to 0} - 2L_s'(t)\big|_{t\to 0}} = \frac{\rho + u_0 - 2x_s(0)}{\rho - x_s(0)} \qquad (31)$$

$$\Omega_\infty = \frac{(1-\rho)(W_e - W_0)}{2(L_e - L_0)(I^* + \rho(2-\rho)(C-I))} - 1 \qquad (32)$$

Figure 5 shows that this scheme gives consistent results when the 'true' values of Ω are estimated from Markov simulated queue growth for several values of ρ (left). The fit between simulated and either time- or interpolation-corrected growth is close (right, where the curves are indistinguishable). Taylor and Heydecker (2014) find that a link-function approach based on stochastic relaxation time is also useful when estimating functions to correct signal queue moments for different throughput capacities in green phases. This procedure, while to some extent heuristic, appears to be the simplest way to satisfy the constraints of the initial and asymptotic states, and equations (31,32) together embody all information about them contained in the three moments.

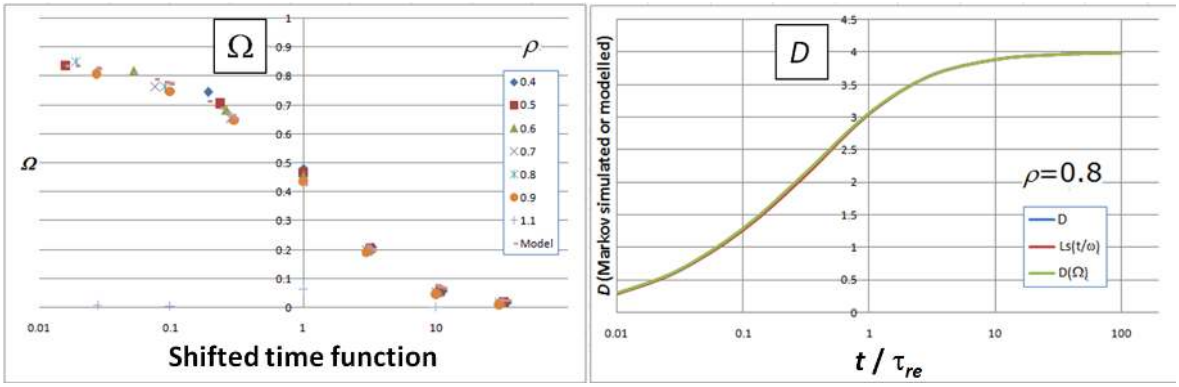

Figure 5. Verification of time transformation and interpolation procedure

The corrected time-dependent mean queue and utilisation are now obtained by successive differentiation of (28). Variance V is then obtained by substituting L and D into (13), its correct convergence to equilibrium being guaranteed by the construction.

$$L(t) = D + \dot{D}t \qquad \text{and} \qquad u(t) = \rho - \dot{L}/\mu \tag{33}$$

Figure 6 (left and middle) compares simulated and approximated development of a queue growing from zero to equilibrium at ρ=0.9 under M/M/1 (upper curves) and M/D/1 (lower curves) processes, plotted on a logarithmic time scale. Theoretical equilibrium means are 9 and 4.05, and variances 90 and 22.8825, respectively. Some overshoot in the variance reflects small errors in D to which the variance formula is highly sensitive at larger values of t. However, this should have little practical significance, especially when expressed in terms of standard deviation. Figure 6 (right) illustrates modelled undershoot of the mean where the initial queue is exact (L_0=9, V_0=0) rather than in equilibrium, with part of the growth trajectory of variance (as standard deviation) also shown.

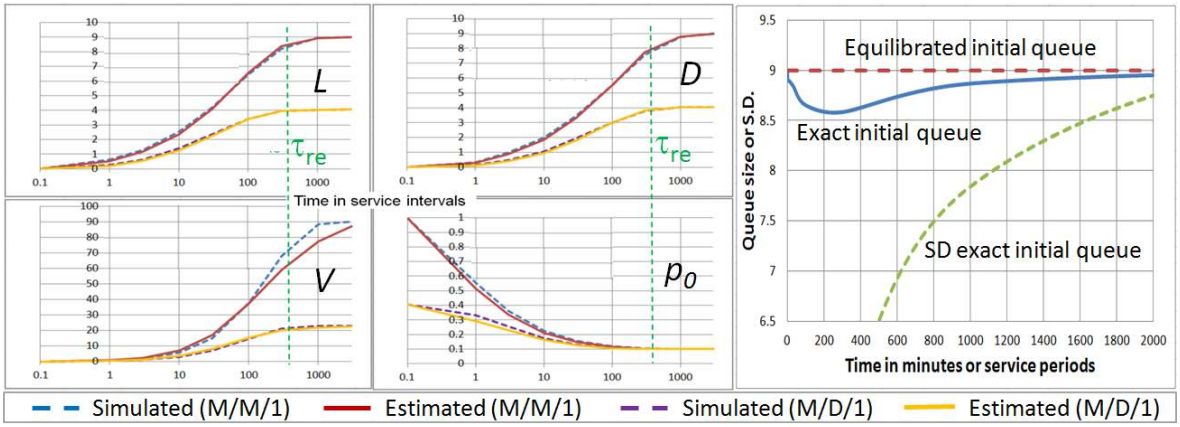

Figure 6. Queue growth at ρ=0.9 (M/M/1 above, M/D/1 below) and undershoot example

7.3 Oversaturated growth regime

After some experimentation, it is considered that the origin-shifted version of the coordinate-transformed queue (26) can be used without correction where the queue is far from equilibrium. Under saturation, where throughput is constant, queue growth is almost linear and the queue integral or delay-per-unit-time on $[t_0,t_0+t]$ can be approximated by:

$$D(t) \approx L_t(0, t_0 + t/2) \tag{34}$$

However, as there is no built-in constraint on the variance, this can be inaccurate under heavy demand. Furthermore, as p_0 diminishes almost to zero its leverage on the form of the

probability distribution decreases. In practice, distributions tend to normal form once p_0 is so small that service is almost continually saturated. The initial probability of zero queue can then be estimated from the initial mean and variance according to the first of equations (35). If p_0 is sufficiently small, 10^{-3} or less, accuracy may be improved by extrapolating mean and variance at time t deterministically according to the last two equations, with the final p_0 value obtained from these moments similarly to the initial value.

$$p_{0N0} = \frac{\exp\left(-\frac{1}{2}L_0^2/V_0\right)}{\sqrt{2\pi V_0}}, \quad L_N = L_0 + (\rho - 1)\mu t, \quad V_N = V_0 + (1-\rho)(2L_e + 1)\mu t \qquad (35)$$

A weighted combination of these methods could be used, but in practice p_0 tends to decline so sharply that it is sufficient to select one or the other. For M/D/1 cases, the modelled average $\bar{p}_0$ must be multiplied by e^{ρ} to give the correct instantaneous value for comparison with that in the Markov simulated distribution (see Tables 2, 3 earlier).

7.4 Decay regime

During decay of a heavy queue, its probability distribution can undergo large and rapid change making approximation difficult. The distribution can be viewed as a linearly weighted superposition of exact queue states that evolve at different rates. High valued states drift down almost linearly while low valued states may be near equilibrium or even growing. Any normal shape produced by a period of oversaturation is quickly lost, and the distribution moves towards equilibrium in a way somewhat resembling the collapse of a viscous mass that collides with and rebounds from the 'barrier' at zero queue size. This can be seen in examples of distributions given by Taylor and Heydecker (2015), where the distribution immediately post-peak develops a 'duck-tail' at small sizes and later becomes 'heavy-tailed' with a bi-modal shape, eventually relaxing to an equilibrium-like shape.

Moments tend to be less volatile, indeed their behaviour can be rather simple as long as service is saturated, but there comes a point when this is no longer true. Kimber and Hollis (1979) assumed linear decay above twice L_e and an inverted form of origin-shifted queue below this. As Figure 3 showed, this clever but structurally questionable manoeuvre can lead to serious error. In fact, the behaviour of variance in decay, as in oversaturated growth, is dominated more by the absolute value of the D function than by its difference from L_e. Together these suggest abandoning the coordinate-transformed model in favour of more direct approximation. Possibly the simplest approximation to the mean queue is an exponential function fitted to initial and equilibrium values, where the timescale factor is also interpolated exponentially to satisfy boundary conditions, equations (36,37).

$$L_m(t) = L_e + (L_0 - L_e)e^{-t/\tau_m(t)} \qquad \text{where} \qquad (36)$$

$$\tau_m(t) = \tau_a + (\tau_i - \tau_a)e^{-t/\min(\tau_i,\tau_a)}, \quad \tau_i = \frac{L_e - L_0}{(\rho - u_0)\mu}, \quad \tau_a = \frac{W_e - W_0}{2(1-\rho)(L_e - L_0)\mu} \qquad (37)$$

With these definitions, differentiating (36) gives an expression (38) for utilisation at small values of t, whose form resembles Shore's (1978) maximum entropy time-dependent solution for M/M/∞, suggesting that this queue model is close to the 'least special' choice.

$$u(t) \approx \rho\left(1 - e^{-t/\tau_i}\right) + u_0 e^{-t/\tau_i} \quad (t \to 0) \qquad (38)$$

Expressions for D_m, V_m are obtained directly from (36) by integration and applying (13), but because the variance result is not explicitly constrained certain safeguards are needed:

- If τ_a is negative, τ_m is set equal to τ_i; and t / $min(\tau_i,\tau_a)$ is limited to avoid overflow
- L_m has a lower limit of the deterministic size of the queue (equation (1) with x=1)
- V_m has a lower limit of $min(V_0,V_e)$.

The exponential approximation is unsatisfactory for growth, for reasons converse to the above, but seems to perform adequately for decay. While it might be improved, modifications requiring complex heuristics or empirical parameters would be undesirable.

8. Results of tests with full demand profiles

Figure 7 compares Markov simulated (dashed) and model-approximated (solid) profiles of an oversaturated peak case (ρ_{max}=1.1384) and a random profile case involving sporadic bursts of heavy demand (ρ_{max}=1.4911). The match between estimated and simulated queue size profiles is so close that they are almost indistinguishable, and the fit between variance profiles may be considered good enough for practical purposes.

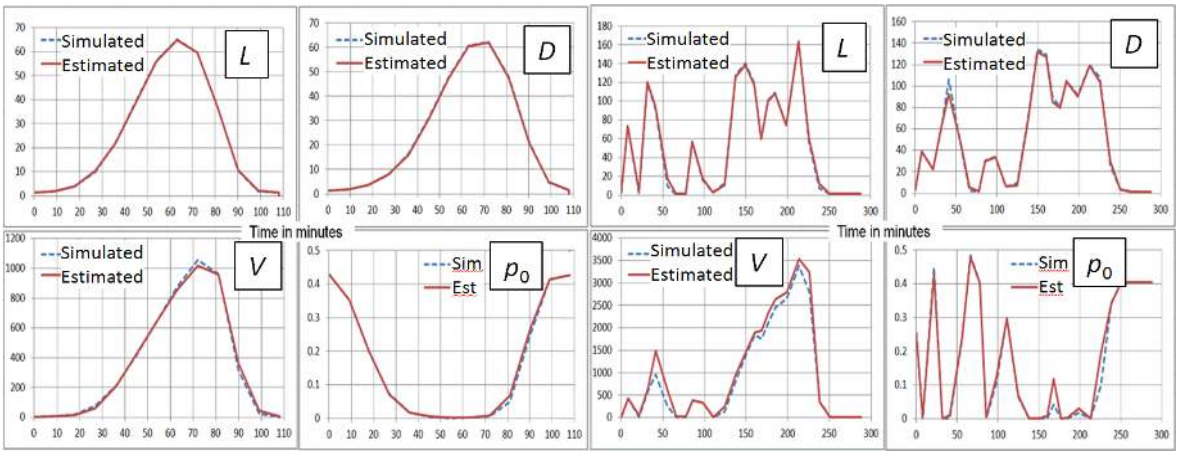

Figure 7. Estimated and simulated profiles for oversaturated peak and random profile

Figure 8 plots results for each separate time slice in all the test cases, standard deviation replacing variance to compress the scale. As absolute magnitudes vary greatly, aggregate error is represented by correlation coefficients, with R^2 exceeding 0.99 for all results except p_0. These results confirm the useful accuracy of the method.

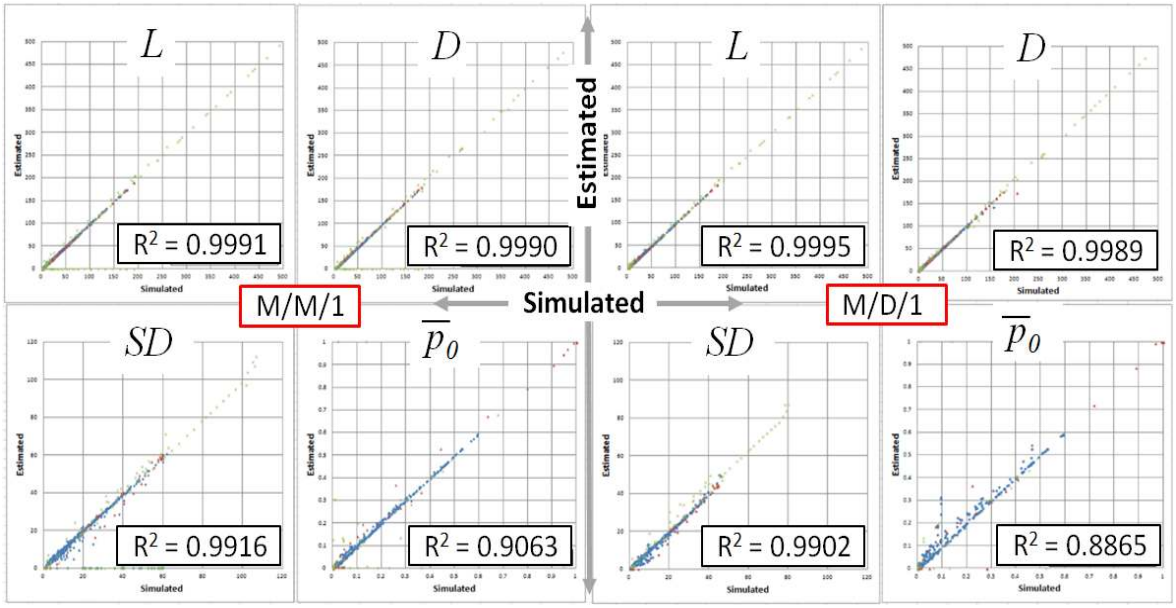

Figure 8. Estimated versus simulated results for all M/M/1 and M/D/1 test cases

9. Consequences for design, planning and policy

The results above show that estimation of variance can be added relatively simply to the established closed-form model of the mean queue, enabling estimation of day-to-day variability or uncertainty expected under similar average conditions. Actual underlying conditions may be difficult to pin down, so if the basic M/M/1 and M/D/1 models are not thought appropriate, observed traffic statistics could be used to estimate P-K parameters, although further work is needed to ensure that the methods of approximation and estimating probability distributions work in more general cases.

The form of and relationships between the time-dependent mean, variance and higher moments reflect the particular nature of queues, which have a simple physical generating process and a barrier at zero size as well as the possibility of a steady state, leading to relatively constrained behaviour and predictability of their probability distributions. Conversely, queue size can be volatile especially after oversaturated peaks. Figure 9 (left) illustrates a modelled highly skewed post-peak queue size distribution, with a 'heavy tail' creating a risk of impacts that may not have been anticipated or designed for. 'Heightened risks' reflect the consequential impacts of overloading a particular system. The shape of the probability distribution determines not only the magnitude of overloading risk at the current level of demand, but also how the system responds to exogenous growth in demand (e.g. Taylor, 2012). Even if mitigation by design or demand management is impractical, it will be advantageous to be able to analyse the risk of overloading, especially of critical services, and the likely effect of 'triage', metering or other selective measures.

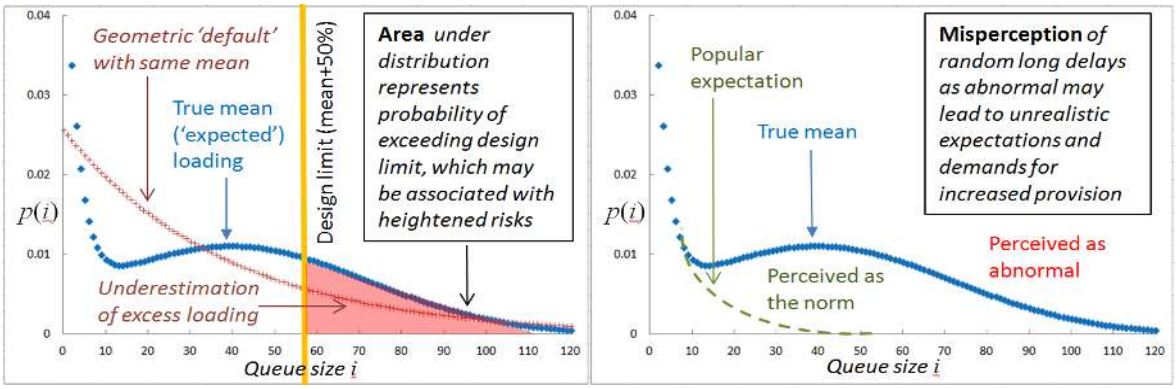

Figure 9. Illustrative queue size distribution showing consequences of 'heavy tail'

In Figure 9 (right), speculatively, random longer queues and delays are perceived as abnormal, which could encourage unrealistic expectations leading to demands for further provision, which in turn generate more demand (SACTRA, 1994), resulting in a self-reinforcing cycle. This may be exacerbated by behavioural 'baseline drift' such as the tendency to translate time savings into increased travel distances observed by Metz (2014).

10. Conclusion

Formulae have been derived for time-dependent and equilibrium queue variance, which are believed to be novel and general at least for processes covered by the Pollaczek-Khinchin mean formula. These results are used to develop an extension of macroscopic time-dependent queue approximation to include variance. Improved accuracy and correct asymptotic behaviour are assured by taking account of essential properties of queue development and embodying mutual consistency between queue size moments in the structure. Benchmarking against Markov simulations of transiently or randomly oversaturated test cases shows good results for M/M/1 (random service, similar to a

priority junction or roundabout) and M/D/1 (uniform service, similar to a traffic signal) queue processes, with aggregate $R^2>0.99$ for mean and standard deviation.

Although the structure of the approximation method is heuristic, it is anchored in theory and involves no free parameters or empirical calibration. While including variance is somewhat more complicated than calculating the mean alone, the use of compatible closed-form formulae means it could be incorporated as a module in existing junction design and assignment software with no changes to program structure or data. As it accommodates all the Pollazcek-Khinchin statistical parameters, with further work it should be extendible to processes other than M/M/1 and M/D/1. From a theoretical viewpoint, the time-dependent and equilibrium variance expressions may offer structural insights not available from considering only the mean, like the family of functions for moments, although it is doubtful that evaluating moments higher than variance is useful.

In the past, estimating queue size probability distributions, for example to evaluate risks of overloading services or exceeding critical limits while avoiding unjustified assumptions of equilibrium or normality, has required detailed simulation or *ad hoc* extensions limited to specific processes. By delivering *three* time-varying moments, namely the probability that the queue is zero ('0th moment'), as well as the mean and variance, the method described enables realistic probability distributions to be estimated. While this currently requires an iterative Newton method, there are standard codes for this, and distributions are likely to be needed only at particular places and times so may be calculated off-line.

Queue processes that can currently be modelled are those representing transient or periodic overloading at road junctions and transport services like car parks, airports and borders, but may be applied to hospitals and A&E/ER units, or anywhere subject to variability that can be represented by a single service process. High-capacity transport systems like motorways and railways tend to be dominated by deterministic variations rather than the random ones that affect urban road junctions, but they too are subject to unpredictable demand and incidents that can affect distributions of delay. The work may also have implications for design and policy where the shape of probability distributions of queue size or delay affects performance, consequential impacts, or perceptions of service.

Acknowledgments

The paper draws from the author's PhD research at University College London, under the supervision of Professor Benjamin D Heydecker of the Centre for Transport Studies and second supervisor Dr Taku Fujiyama. Some initial investigations took place at the Transport Research Laboratory including development of Markov simulation software with Neil H Spencer. Helpful comments by Professor Heydecker, Dr Alan Stevens at TRL and several reviewers are gratefully acknowledged. The work was presented at the UTSG 2017 Conference in Dublin, January 2017, hosted by Trinity College and Gresham's Hotel.

References

Addison, J. D. (2006). Behaviour of variance of delay. *Proc. UTSG Conference January 2006*, Dublin.

Addison, J. D. and Heydecker, B. G. (2006). Journey time variability on a congested link. *Proc. UTSG Conference January 2006*, Dublin. http://iris.ucl.ac.uk/iris/publication/44326/1 [accessed 7/3/17]

Akçelik, R. (1980). Time-dependent expressions for delay, stop rate and queue length at traffic signals. *ARRB Internal Report AIR 367–1*. Australian Road Research Board.

Arup, Bates, J., Fearon, J. and Black, I. (2004). *Frameworks for Modelling the Variability of Journey Times on the Highway Network.* UK Department of Transport.

Binning, J. (2004). ARCADY 6 User Guide. *TRL Application Guide AG49*. Transport Research Laboratory, Crowthorne House.

Bunday, B. D. (1996). *An introduction to queueing theory*. Arnold (Hodder Headline).

Burrow, I. J. (1987). OSCADY: a computer program to model capacities, queue and delays at isolated traffic signal junctions. *TRL Report RR 105.* Crowthorne House.

Catling, I. (1977). A time-dependent approach to junction delays. *Traffic Engineering and Control*, 18, 520-526. Hemming.

Clarke, A. B. (1956). A Waiting Line Process of Markov Type. *Annals of Mathematical Statistics*, 27, 452-459.

Doherty, A. R. (1977). A comprehensive junction delay formula. *LTR1 working paper*. UK Department for Transport, London.

Erlang, A. K. (1909). The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B20*, Købnhavn.

Fosgerau, M. (2008). On the relation between the mean and variance of delay in dynamic queues with random capacity and demand. *MPRA Paper No 11994/Technical University of Denmark.* http://mpra.ub.uni-muenchen.de/11994/

Fu, L. and Hellinger, B. (2000). Delay variability at signalized intersections. *Transportation Research Record 1810*, 215-221. Transportation Research Board, Washington DC.

Griffiths, J. D., Leonenko, G. M. and Williams, J. E. (2005). The transient solution to M/E*k*/1 queue. *Operations Research Letters*, 34(2006), 349-354.

HCM (2010/2016). *Highway Capacity Manual*. Washington DC: Transportation Research Board. http://hcm.trb.org/?qr=1 (2016 update of 2010 edition available for purchase)

Heidemann, D. (1994). Queue length and delay distributions at traffic signals. *Transportation Research Part B,* 24B(5), 377-389. Amsterdam: Elsevier.

Heydecker B. G. and Verlander N. Q. (1998). Transient delay in oversaturated queues. *Proc. 3rd IMA International Conference on Mathematics in Transport Planning and Control*, Cardiff, 1-3 April 1998.

Kendall, D. G. (1951). Some problems in the theory of queues. *J. Royal Statistical Society B (Methodological)*, 13(2), 151-183.

Kimber, R. M. and Hollis, E. M. (1979). Traffic queues and delays at road junctions. *TRL Report LR 909.* Transport Research Laboratory, Crowthorne House.

Kimber, R. M. and Daly, P. (1986). Time-dependent queuing at road junctions: observation and prediction. *Transportation Research Part B*, 20B(3), 187-203.

Kimber, R. M., Daly, P., Barton, J. and Giokas, C. (1986). Predicting time-dependent distributions of queues and delays for road traffic at roundabouts and priority junctions. *J. Operational Research Society*, 37(1), 87-97. Palgrave Macmillan.

Kleinrock, L. (1975). *Queueing systems: Volume 1 Theory*. Hoboken NJ: Wiley

Lindley, D. V. (1952). The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(2), 277–289.

Little, J. D.C. (1961). A simple proof of $L=\lambda W$. *Operations Research* 9, 383–387.

Leonard, D. R., Tough, J. B. and Baguley, P. C. (1978). CONTRAM: a traffic assignment model for predicting flows and queues during peak periods. *TRL Report LR 841*. Transport Research Laboratory, Crowthorne House.

Leonard D. R., Gower, P. and Taylor, N. B. (1989). CONTRAM: Structure of the model. *TRL Report RR178*. Transport Research Laboratory, Crowthorne House.

Machta, B. B., Chachra, R., Transtrum, M. K. and Sethna, J. P. (2013). "Parameter space compression underlies emergent theories and predictive models." *Science*, 342, 604-607.

Medhi, J. (2003). *Stochastic models in queueing theory*. Elsevier Academic Press.

Metz, D. (2014). *Peak car - the future of travel*. Landor.
Morse, P. M. (1955). Stochastic properties of waiting lines. *J. Operations research Society of America*, 3(3), 255-261, August 1955.
Morse, P. M. (1958). *Queues inventories and maintenance*. Wiley.
Newell, G. F. (1971). *Applications of queuing theory*. Chapman and Hall. (Revised 1982)
Olszewski, P. S. (1990). Modelling of queue probability distribution at traffic signals. *Proc. International Symposium of Traffic and Transportation Theory, 1990*.
Robertson, D. I. (1969). TRANSYT: a traffic network study tool. *TRL report LR253*. Transport Research Laboratory, Crowthorne House.
SACTRA (1994). *Trunk roads and the generation of traffic*. Standing Advisory Committee on Trunk Road Assessment. UK Department for Transport.
Semmens, M. C. (1985). PICADY2: An enhanced program to model capacities, queues and delays at major/minor priority junctions. *TRL Report RR36*. Crowthorne House.
Sharma, O. P. (1990). *Markovian queues*. Ellis Horwood.
Shore, J. E. (1978). Derivation of equilibrium and time-dependent solutions to M/M/∞//N and M/M/∞ queueing systems using entropy maximization. *Proc. National Computer Conference*, 1978.
Soros, G. (1987). *The alchemy of finance*. John Wiley & Sons, Hoboken NJ.
Spiess, H. (1989). *Conical volume-delay functions*. Available at: http://www.inrosoftware.com/en/pres_pap/papers/conic.pdf [accessed 24/3/14]
Taylor, N. B. (2003). The CONTRAM dynamic traffic assignment model. *J. Networks and Spatial Economics - special issue on Dynamic Traffic Assignment, 3(2003)*, 297-322, Kluwer.
Taylor, N. B. (2005). Variance and accuracy of the sheared queue model. *Proc. IMA Mathematics in Transport Conference,* University College London, September 2005.
Taylor, N. B. (2012). A recipe for jam - Can congestion be define consistently? *Proc. UTSG 2012 Conference*, Aberdeen University, January 2012.
Taylor, N. B. (2014). *Queue methods for variability in congested traffic*. PhD dissertation. Department of Civil, Environmental and Geomatic Engineering, University College London. Awarded 28 January 2014.
Taylor, N. B. and Heydecker, B. G. (2014). The effect of green time on stochastic queues at traffic signals. *Transportation Planning and Technology*, 37(1), February 2014. Available at: http://dx.doi.org/10.1080/03081060.2013.844907
Taylor, N. B. and Heydecker, B. G. (2015). Estimating probability distributions of dynamic queues. *Transportation Planning and Technology*, 38(1), February 2015. Available at: http://dx.doi.org/10.1080/03081060.2014.976987
TRL (2015). *TRANSYT: Traffic Network and Isolated Intersection Study Tool.* https://trlsoftware.co.uk/products/junction_signal_design/transyt [accessed 7/3/17]
van Vliet, D. (1982). SATURN- a modern assignment model. *Traffic Engineering and Control*, 23(12), 578-581. Hemming.
Viti, F. and van Zuylen, H. J. (2004). Modeling queues at signalized intersections. *Transportation Research Record 1883*, 120-135.
Viti, F. and van Zuylen, H. J. (2010). Probabilistic models for queues at fixed control signals. *Transportation Research Part B*, 44, 120-135.
Webster, F. V. and Cobbe, B. M. (1966). Traffic signals. *Road Research Technical Paper 56*. HMSO.
Whitt, W. (1983). Untold horrors of the waiting room: what the equilibrium distribution will never tell about the queue-length process. *Management Science*, 29(4), 395-408.
Willmot, G. E. (1986). Mixed compound Poisson distributions. *Astin Bulletin*, 16S, 59-79.