QUEUE METHODS FOR VARIABILITY IN CONGESTED TRAFFIC

by

Nicholas B Taylor

A thesis submitted for the degree of
Doctor of Philosophy

at

University College London
Department of Civil, Environmental and Geomatic Engineering

DECLARATION

I, Nicholas B. Taylor confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

ABSTRACT

Time-dependent queue methods are extended to calculate variances of stochastic queues along with their means, and thereby provide a tool for evaluation and better understanding of travel time variability and reliability in congested traffic networks and other systems, including through probability distributions estimated from moments. Objectives include developing computationally efficient analytical methods, and achieving robustness by reflecting the underlying structure of queuing systems rather than relying on statistical fitting,

New deterministic and equilibrium formulae for queue variance are developed, acting also as constraints on estimating time-dependent queues generated by a range of processes, enabling improved accuracy and reliability estimates. New methods for approximating equilibrium and dynamic probability distributions use respectively doubly-nested geometric distributions and exponentially-weighted combinations of exponential and Normal functions, avoiding the need to rely on empirical functions, costly simulation, or equilibrium distributions inappropriate in dynamic cases.

For growing queues, corrections are made to the popular sheared approximation, that combines deterministic and Pollaczek-Khinchin equilibrium mean formulae in one time-dependent function. For decaying queues, a new exponential approximation is found to give better results, possibly through avoiding implicit quasi-static assumption in shearing. Predictions for M/M/1 (yield) and M/D/1 (signal) processes applied to 34 oversaturated peaks show good agreement when tested against Markov simulations based on recurrence relations.

Looking to widen the range of queues amenable to time-dependent methods, dependence of stochastic signal queues on green period capacity is confirmed by an extended M/D/1 process, for which new formulae for equilibrium moments are obtained and compared with earlier approximations. A simple formulation of queuing on multiple lanes with shared service is developed, two-lane examples with turning movements showing fair match to simulation.

The main new methods are implemented in a spreadsheet demonstrator program, incorporating a database of time-sliced peak cases together with a procedure for estimating dynamic probability distributions from moments.

ACKNOWLEDGEMENTS

I wish to thank the following:

My Supervisor, Professor Benjamin G Heydecker

Dr Taku Fujiyama, my Second Supervisor and Upgrade Examiner

Dr Eugeny Buldakov, Upgrade Examiner

Neil H Spencer, who contributed to the development of the Markov simulation program

Dr Alan Stevens, Transportation Chief Scientist at the Transport Research Laboratory, Professor Malachy Carey of Queens University Belfast, Professor Nick Hounsell of Southampton University, and Dr Abdulai Dumbuya for their support.

Many researchers and authors in the field, whose work has provided the grounding and inspiration of this research, including Professor Richard E Allsop of University College London; Dr Rod Kimber, Dr Dennis Robertson, Professor Mike Maher, Ian Summersgill, Erica Hollis, David Leonard, Peter Gower and P D Whiting while at Transport Research Laboratory or its precursors; A R Doherty and Ian Catling.

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 the concept is also very deceptive… Equilibrium has rarely been observed in real life [George Soros 1987]

Dedicated to

My Parents, Sidney Beresford Taylor DFC, and Elizabeth Dobinson
Professor Sir James Dunbar-Nasmith CBE, (PP)RIAS, architect and mentor
and John A Kenward MChem (Oxon), friend, biotechnologist and high-flyer to the end

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION AND REVIEW

The thesis of this dissertation is that general expressions for deterministic time-dependent and equilibrium queue variance can be derived, and when combined with corresponding expressions for the mean queue, including existing methods modified for accuracy, enable both mean and variance of queues to be estimated with improved accuracy. The results can be used to estimate general time-dependent profiles of queue mean and variance under transient oversaturation. Dynamic probability distributions can be estimated from queue moments, giving more detailed information about the reliability of facilities under congested conditions.

1.1. BACKGROUND AND CONTEXT

1.1.1 Link between variability and congestion in traffic networks and other facilities

Origins and destinations of journeys are distributed in space, so any traffic that does not have total freedom of movement tends to move through a network linking these points. All network links, generators and attractors are subject to capacity limitations and hence to delay and congestion when demand exceeds their capacity, and delay rises steeply when capacity is approached. Journeys tend to be constrained in time as well as route choice, so options to redirect demand may be limited. Where capacities are inflexible[1] and demand is not manageable at source[2], as is typical of roads, hospitals and other health services, call-centres, over-counter services etc, there is risk of congestion, leading to disruption and dissatisfaction. Variability of demand and service makes congestion to a degree unpredictable, while nonlinearity and sensitivity around saturation further complicate prediction. Where scope for new 'provision' is limited, reliability becomes a primary objective. In the quest for improved reliability, a first step is to understand how variability affects the network, and where queuing is the main effect the natural way to describe this is in terms of the variance of queue size. Direct simulation is costly. For some processes, deterministic[3] and analytical[4] methods are well developed for calculating mean queue sizes or waiting times in equilibrium, as well as some transient behaviour, and approximations are available for more general time-dependent behaviour. However, calculation of variance is limited to equilibrium values for a few processes, or empirical methods for particular scenarios. No general time-dependent method has been available for calculating variances or dynamic queue size probability distributions.

[1] Counter-examples of service *with* some flexibility include supermarket check-outs and airline check-in desks.

[2] Examples of demand controlled at source could include pre-booked airline or train seats.

[3] 'Deterministic' is used here broadly to mean an expression which does not require or depend directly on randomness, though it can include invariants of random processes.

[4] 'Analytical' is used here loosely to mean closed-form formulae embodying the structure (analysis) of a problem, not just an empirical or statistical fit to data.

1.1.2 Costs of road traffic congestion

While not the only area of interest, road traffic is a convenient, as well as intensively studied, paradigm. Road traffic congestion seems to increase relentlessly throughout the world. There is a view, summarised as ‘peak car’, that road traffic in the UK (and elsewhere) may have peaked and now be in decline (Goodwin 2010, Metz 2009,2010). However, this is not universally accepted, and spectacular extreme congestion and delay events occur particularly in the ‘developing’ world (e.g. Global Times 2010), while in ‘developed’ countries a high proportion of congestion is associated with recurrent demand (65% in the UK, 55% in the USA), as opposed to temporary loss of capacity through incidents, works or weather, and is therefore likely to recur at times and places where the network is most sensitive to variability. The annual cost of congestion in the UK has been estimated by Eddington (2006) as £8B, while a recent opinion survey (LTT 2012) produced estimates as high as £30B per annum. Apart from the question what constitutes ‘congestion’ as distinct from unavoidable delay (Lay 2011, Taylor 2012), the inability economically or politically, or even theoretically, to build a way out of congestion (‘predict and provide’) has led to a focus on reliability. Reliability is defined formally as the probability of a system performing its intended function during a given time period, but this is unhelpful for roads whose performance is not under the full control of operators. Therefore it tends to be measured in terms of journey time variability, which may be valued differently from average time (Gordon *et al* 2001, Fosgerau and Karlström 2010).

The UK Department for Transport has used two measures of journey time reliability in its published statistics: from July 2005 to March 2010 average delay on the 10% most delayed journeys on a set of routes; from April 2010 the average percentage of journeys arriving ‘on time’ set against reference speeds. These apply to strategic routes, where delays result mostly from transient overload or flow breakdown, or loss of capacity through incidents and road works. Speed/flow/density relationships are often used to describe local traffic conditions and to characterise level of service, though the latter is problematic (see e.g. Taylor 2012).

In urban road networks, delays may arise from random fluctuations in demand or capacity as well as overload. When average demand is below average capacity, some capacity remains unused, so upward fluctuations sufficient to cause a queue to form are not compensated by downward fluctuations, resulting in a net average queue that increases with variability. There are ongoing studies of macroscopic speed/flow/density relationships (MFD) for urban networks, analogous to those associated with motorways, that may lead to a concept of level of service for urban networks (B G Heydecker, personal communication). The present work is concerned only with delay arising from explicit queuing rather than from reductions in average speed. Some delay is an inevitable price of mobility, because the economic cost of relieving it completely could not be justified (Miller 1969, and other references in Taylor 2012). In the absence of control over demand and capacity, reliability implies keeping delay within tolerable limits by managing or allowing for variability. A first step towards this is to be able to predict variability of delay, which requires an understanding of the variance of queuing processes.

1.1.3 Development of time-dependent queuing theory

Queuing theory applied to all kinds of communication networks and service facilities has a long history going back to the work of Agner Krarup Erlang at the Copenhagen Telephone Company (Erlang 1909). In the 1930s interest developed in the performance of freeways in the USA, but this focused on continuous flow rather than queuing (Greenshields 1935). Application to road traffic began in earnest in relation to uncontrolled junctions and traffic signals in the late 1950s and 1960s (Webster 1958, Tanner 1962, Webster and Cobbe 1966), and later to the design of priority junctions and roundabouts. Allsop (1971) focused on the optimisation of signal timings to minimise total delay, as an alternative to Webster's principle of equal saturation on all arms.

In the 1970s and 1980s observations and track experiments led to empirical models of junction capacity and development of approximate formulae. Work funded over several years by the UK Department for Transport (DfT) used the multi-disciplinary staff and special track facilities at the Transport Research Laboratory (TRL)[5] to set up repeatable controlled experiments, as well as direct observations. In the late 1970s P D Whiting[6], Robertson and Gower (1977), Doherty (1977) and Catling (1977), working at the DfT or at TRL, developed coordinate transformed or 'sheared' methods, subsequently refined by Kimber and Hollis (1979). These methods combine deterministic queue development with random equilibrium queuing theory to give an approximation to time-dependent queues, including growth through saturation that neither component can handle separately. Akçelik (1980, 1998a,b) describes analogous methods developed in the 1980s for priority and signal junctions.

These results contributed to the development of several computer programs to assist priority and signal junction and network design including TRANSYT, CONTRAM, ARCADY, PICADY and OSCADY (Robertson and Gower 1977, Kimber *et al* 1985,1986, (Semmens 1985a,b), Burrow 1987, Leonard, Gower and Taylor 1989, Taylor 1990,2003). During this period, Kimber and Daly (1986) explored the variability of queues under consistent conditions. They found that predicted queue size matched the real average well, but individual observed queue sizes covered a range up to more than twice the average, as shown in Figure 1.1.1.

[5] 'TRL'may be taken to include all its previous manifestations, including RRL or TRRL as it was known until 1990 while still the research arm of the UK Department of Transport, and up to and after privatisation in 1996.
[6] Whiting left computer code but no documentation, but is also referenced with Hillier for optimum route search.

In the USA, Gordon F Newell (1960,1968a,b,c,1971,1982) was a pioneer of queuing theory applied to road traffic in the late 1960s, and may have been the first to suggest combining deterministic and quasi-static equilibrium functions, in his case using a diffusion approximation, but did not go so far as to develop a practical time-dependent analytical closed formula applicable to transient over-saturated queues. Figure 1.1.2 (a-c) reproduces some of the original figures of these authors, showing the common theme in their ideas and the idea of 'shearing' the equilibrium function into the time-dependent function. Figure 1.1.2 (d) shows that unlike the mean queue, variance does not 'shear' naturally.

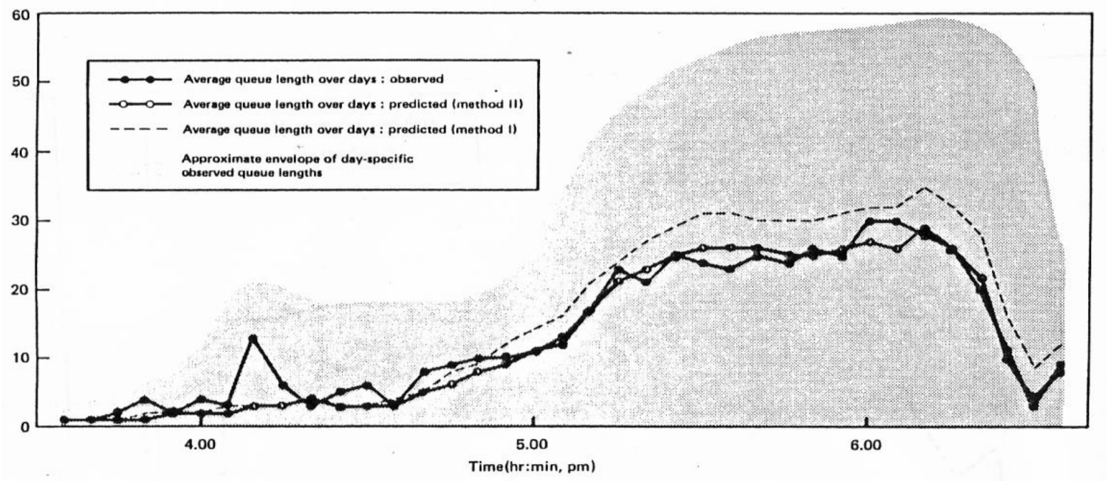


Figure 1.1.1 Original figure of Kimber and Daly (1986) showing observed queue variability

Figure 1.1.2 Original figures of four authors illustrating shearing and its precursors

1.1.4 Seeking improved accuracy

As will be discussed later, the sheared queue approximation contains a source of inaccuracy in its implicit assumption that a dynamically developing queue behaves as quasi-static or locally in equilibrium. Dissatisfied with its accuracy, Kimber and Hollis (1979) modified the sheared formula for growth by replacing the initial queue by a shift in the time origin. By so doing they hoped to avoid indeterminacy of the initial queue state (equivalent to an unspecified probability distribution) by making it the result of the growth process. For a decaying queue, they assumed that an initial queue greater than twice the equilibrium value (at the current level of demand) decays linearly to that value, after which it behaves like a mirror image of the growth function as it falls to its equilibrium value. Queue development profiles estimated by alternative methods are compared in Figure 1.1.3, where the subscript ‘0’ indicates the initial state at *t*=0, and subscript ‘e’ indicates the equilibrium state. This particular example does not support the idea that the origin-shifted/mirror-image method delivers better accuracy compared to benchmark Markov simulations (using step size 0.1 unit), and maintaining the initial rate of decay for an extended period tends to overestimate the rate at which the queue falls.

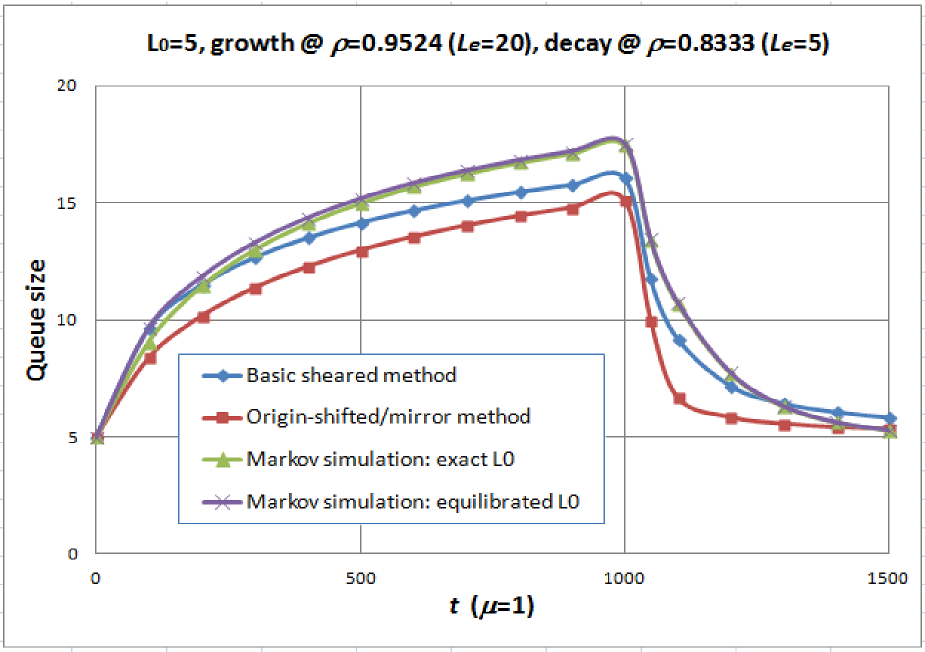


Figure 1.1.3 Growth/decay example using LR909 methods (time-advanced, not time-sliced)

While the origin-shifted/mirror-image method is ingenious, requires no additional empirical parameters or constants, and plays a part in this research, it appears to rely on graphical rather than physical arguments, and the accuracy of its approximation was not investigated further. This work aims to improve accuracy systematically by incorporating additional natural structural constraints including that imposed by variance.

1.1.5 Existing approaches to calculating queue variance

While the size of a queue is governed by conservation (what enters must either leave or stay), estimation of variance is inseparable from a probabilistic description. Queue variance can be estimated by repeated simulations, but this is not always a practical option. While queuing theory can provide formulae for the equilibrium variance along with the mean size of some queuing processes, there has been no efficient closed-form method available for calculating time-dependent variance, nor equilibrium variance for more general processes, nor specific probability distributions. This would for example enable estimation of the likelihood of occasional long tailbacks, or how likely it is that a queue will spill back across a junction or other facility upstream, a situation that can exacerbate congestion or lead to gridlock.

Exact descriptions of or close approximations to time-dependent queues exist for some simpler processes (e.g. Morse 1955,1958, Clarke 1956, Cantrell and Beall 1988). Morse (1955) gives the earliest known exact transient solution to queue development, which is limited to the simplest M/M/1 random queue, whose computation is demanding and subject to precision problems. The M/M/1 solution has been extended by Griffiths *et al* (2005) to M/E*k*/1 systems[7], representing staged service, effectively with reduced dispersion, and also multi-channel systems (Morse 1955). Markov simulation, evolving queue state probabilities in small steps, is more efficient than direct calculation and can be applied to any queue type describable in terms of recurrence relations or transition probabilities, but it too is computation-intensive. Microscopic event or 'Monte Carlo' simulation can represent any arbitrary process, but is the least efficient of all. These methods are useful as benchmarks, but are too burdensome for use in most practical predictions, unless compromised. For example, microscopic simulations of networks may require many randomised repeat runs to get an accurate assessment of the reliability of results, but these are seldom performed because of the cost (Wood 2012).

To the author's knowledge, estimation of queue variance has been addressed in three ways:

- Steady state equilibrium variance of waiting time or queue size, plus some analysis of transients (examples in most standard reference works). Addison and Heydecker (2006) extend this by expressing the rate of change of variance in terms of other moments;
- Regression models of symmetrical peaks calibrated from the results of simulations (Kimber *et al* 1986, and Figure 1.1.4), or statistical methods (Eliasson 2006);
- Dynamic relationships between observed or simulated variance and mean, giving loop-like graphs because variance lags mean (Arup, Bates *et al* 2004, Addison and Heydecker 2006, Fosgerau 2010).

[7] The author is grateful to the Examiners for drawing his attention to this result.

Kimber *et al* (1986), using 34 synthetic symmetrical peak cases representing idealised priority junctions with Gaussian demand profiles, fitted parameters of a 'low definition' model of queue variance to simulation results using empirical power functions of peak duration and average demand intensity. Variance was approximated by 'stitching together' two half-Gaussian functions fitted respectively to the growth and decay regimes of the variance, as shown in Figure 1.1.4. Used with the sheared or a low definition queue approximation, this enabled both queue mean and variance to be estimated, but only for symmetrical uni-modal peaks. Kimber *et al* (1986) proposed the regression formula (1.1.1), where T_m is the time of the maximum variance V_m and the B_i are the estimated scale parameters of the half-Gaussian fitting functions:

$$\begin{aligned} V(t) &= V_m \exp\left(-(t-T_m)^2/B_1^2\right) && 0 \le t \le T_m \\ &= V_m \exp\left(-(t-T_m)^2/B_2^2\right) && T_m \le t \le T_L \qquad (1.1.1) \end{aligned}$$

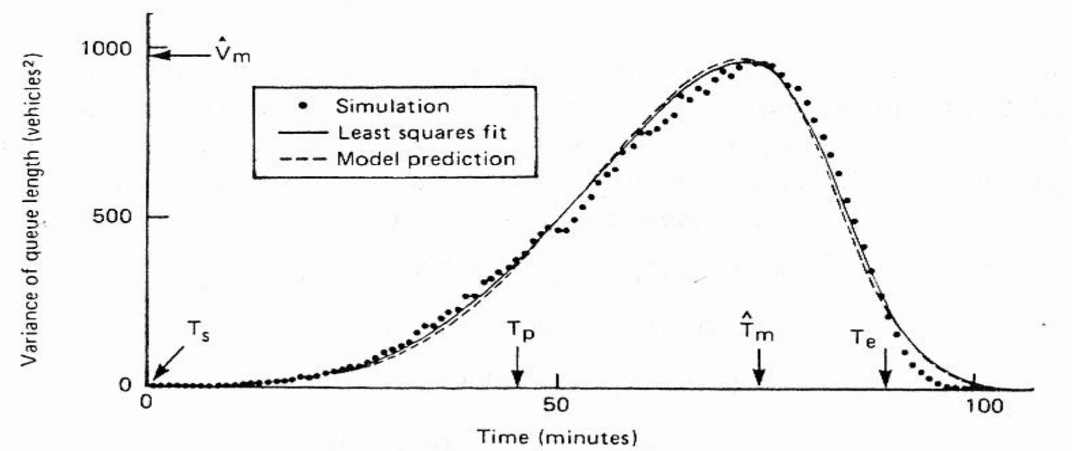


Figure 1.1.4 Original figure of Kimber *et al* (1986) showing bi-Gaussian fit to variance data

Low-definition methods cannot predict the development of a queue produced by arbitrary demand and capacity profiles, limiting their application compared with analytical time-dependent approximations like shearing. Conversely, analytical mean queue methods do not provide sufficient information to enable variances or probability distributions to be estimated.

1.1.6 New expression for deterministic time-dependent variance

This work establishes a basis for efficient extended approximations by deriving a new deterministic formula for the variance of a queue and extending equilibrium formulae to embrace a wider range of queue processes. The deterministic variance formula, given by equation (1.1.3) below and first published in Taylor (2005a), has a structural resemblance to the well-known deterministic mean queue formula (1.1.2) reflecting conservation of queuing units or customers, where L_0 is the initial queue at time t=0, ρ is demand intensity (demand/capacity), and x is the average utilisation of capacity μ. L_e is the equilibrium queue size of the random process, and $D(t)$ is the average queue size or delay per unit time over the period $[0,t]$.

$$L(t) = L_0 + (\rho - x)\mu t \qquad \text{where } x = x(t) \to \min(\rho,1) \text{ as } t \to \infty \qquad (1.1.2)$$

$$V(t) = V_0 + L_0(L_0 + 1) + 2(1-\rho)(L_e - D)\mu t - L(L+1) \qquad \text{where} \qquad (1.1.3)$$

$$L_e = L_e(\rho, process) \qquad = \lim L(t)\big|_{t\to\infty} \text{ in the case } \rho < 1 \quad \text{and} \qquad (1.1.4)$$

$$D(t) \equiv \frac{1}{t}\int_0^t L(y)dy \qquad (1.1.5)$$

The variance formula is believed to be general, and conjectured to represent conservation of some higher quantity, though the physical interpretation is unclear. An equilibrium queue exists only for demand less than capacity, i.e. ρ<1. However, all equilibrium queue formulae include the divisor (1-ρ), so the product (1-ρ)L_e and hence equation (1.1.3) are defined for all ρ.

1.1.7 Approach to the problem of queue estimation

Wherever possible this research exploits precise relationships, some of which are new and many of which are well known, but it does not aim to repeat what has already been explored exhaustively, nor to achieve absolute mathematical rigour in its final results. It aims rather to enable approximate but robust and efficient computational methods to produce more complete and accurate predictions, and in particular to estimate reliability on a sound analytical basis rather than broad assumptions like equilibrium that are seldom tenable in reality. Problems in traffic as posed practically are approximations to situations that are not precisely known, either as a result of their complexity or uncertainties of measurement. A valuable property of an approximate solution is that it reflect the underlying structure and behaviour of a physical system, rather than just achieving a good empirical or statistical fit in particular cases. Such a solution is more likely to make useful predictions for cases not already encountered. This allows that there may be varying degrees of structural affinity between the model and original problems.

Any mathematical formula is inherently precise, but few practical problems in network and queuing analysis admit exact solutions like the M/M/1 formula of Clarke (1956) and Morse

(1955,1958), and they can be hard to use[8]. Therefore, for many practical applications robust accuracy may be preferable to fragile precision[9]. Precise but somewhat idealised formulae include the deterministic queue formula (1.1.2), reflecting conservation of units/customers, and the Pollaczek-Khinchin equilibrium queue formula relating to random processes, and may now be considered to include the deterministic variance formula (1.1.3). The sheared queue approximation has the advantage of elegantly fusing two mutually incompatible formulae, capturing gross properties of time-dependent queuing with a degree of accuracy, without the need for additional behavioural assumptions or calibration constants, and providing a convenient basis for further development. Some superficially similar methods, such as those reviewed in Appendix E, add empirical terms and calibration constants in pursuit of greater accuracy, but this is felt to be undesirable in the present research.

1.1.8 Information available from standard works on queuing

Much standard material on queuing theory is concerned with steady-state equilibrium and average waiting times, e.g. Kleinrock (1975), Newell (1982), Medhi (2003), Gross *et al* (2008). Since equilibrium is seldom achieved in reality, and exact time-dependent analysis tends to be intractable, transient response to change in inputs may be analysed. Although queuing is naturally a discrete process, fluid diffusion approximations using continuous functions have been exploited, e.g. by Newell (1982), Kleinrock (1976). Continuous models can be easier to work with and can provide useful insights. Variances are derived mostly for queues in equilibrium. Useful techniques for obtaining equilibrium probability distributions and moments are given by Kleinrock (1975) and Bunday (1996). Such results can be of limited relevance to dynamic traffic queues in oversaturated peak periods, where for example arrival and capacity profiles may be quite variable and inter-dependent. Sustained peaks are common in road traffic, so that predicting the course of queue growth and decay is more important than analysing transients, whose causes are likely to be events separate from recurrent queuing. Long road queues are often unavoidable because of limited control over demand and service facilities, while in some other contexts capacity and delay may be managed more readily. Networks of queues are addressed in the literature, though in an industrial or telecommunications context the networks may be simpler than in the road context, as flows may not interact and conditions be more controllable. In road traffic networks, despite the inter-dependence of their elements, the statistics of queuing processes are normally assumed to be those of isolated systems. Little research has been done to verify that such approximation is valid, with the exception of Kimber *et al* (1986), although the open-ended nature of such traffic problems argues against the likelihood of building manageable comprehensive solutions.

[8] Kleinrock (1975) describes the complexity of the M/M/1 solution given by Morse (1958) as “most disheartening” despite its describing the “simplest interesting queuing system”.

[9] This distinction is also associated with certain quotations of John Tukey (~1958).

In practical traffic modelling, simulation is increasingly preferred to analysis. The essential difference between competing traffic modelling approaches - microscopic, mesoscopic and macroscopic - lies in the *level* at which simulation takes place (e.g. Taylor 2003). In microscopic modelling, the statistical properties of queues arise from the behaviour of individual customers. In meso/macroscopic modelling, the statistics of queues at different points of a network are characterised in general terms, that can be based on general principles or calculated progressively by considering the competing effects of platooning, dispersion and mixing of streams, green waves etc, as addressed in the context of optimising coordinated signals by for example Robertson and Gower (1977). Stochastic User Equilibrium (SUE) assignments attempt to take account of the uncertainty or variance of demand and network variables, but tend to make simplifying assumptions about their distributions in order to be able to create global objective functions or closed-form solutions (e.g. Maher 1992,1998).

1.1.9 Gap-acceptance as an alternative capacity model

In many practical situations, the capacity of a traffic movement depends on the presence of other traffic, for example in the road context at priority junctions, roundabouts and also at opposed turns at signal junctions. Because of the complexity of possible movements, additional factors of flare lanes and visibility, and typically a large amount of scatter in data with a relative paucity of observations at extreme points compared to average conditions, linear regression models have been developed to calibrate empirical relationships between each movement capacity and opposing flows. Once the capacity of each movement is determined, standard queuing theory can be applied. A good example of this approach is PICADY (Semmens 1985b), using research and experiments at the Transport Research Laboratory by G F Maycock, R M Kimber and I Summersgill *inter alia*.

Gap acceptance is an alternative approach, which by-passes the capacity stage by directly estimating the waiting time in service based on the probabilistic distribution of gaps, and of accepting gaps, in opposing traffic or when overtaking, and the behavioural characteristics of individuals in service. An early use of the term is by Miller (1961) and similar problems are addressed by Tanner (1961,1962) and Hawkes (1968). The gap distribution is commonly assumed to be negative exponential, giving non-linear relationships between unblocked time and opposing flow, and between other pairs of variables (see for example review by Akçelik 2007). The gap acceptance approach can in principle deal explicitly with bunched arrivals, realistic minimum headways in opposing traffic, details of pedestrian behaviour at crossings (e.g. Griffiths 1981), etc. A dichotomy has arisen since capacity-based methods were developed as a response to a need for experimentally verified and calibrated computational tools. Zhang and Excell (2013) report a play-off between the two approaches illustrating the difficulties of matching methodology to reality in the practical case of a complex junction, as well as the emphasis on functionality as opposed to method. The concept of capacity allows a modular approach to traffic modelling which is relatively easy to generalise. The present research falls naturally into the capacity-based camp, with properties of arrivals and service being represented by simple statistics that can be accommodated within the methodology.

1.1.10 Chaotic processes, heavy tails and adaptive systems

As transportation systems become increasingly pressured, the effect of disruptions becomes more critical. This applies not only to road networks, where the nature of disruptions can be depressingly predictable even if their magnitude and timing is not, but to high-value services like airports. These are particularly prone to disruption because of knock-on delays, both within and between airports, heavily bunched movements of passengers, and 'banks' of flights by competing airlines at peak times (Peterson *et al* 1995). Arrivals in such cases can be bunched and highly variable, so that common random processes are considered inapplicable, and the arrival pattern may be better described as arising from a chaotic process.

Chow (2013) proposes that chaotic functions can be used to reproduce the statistical characteristics of queues resulting from transient events such as incidents, and in particular queue length distributions with 'heavy tails', where extreme events are much more probable than would be expected from exponential or Poisson processes. He points out that chaotic maps are deterministic, though they may mimic stochastic processes. Surges of scheduled arrivals at an airport could be considered deterministic, although some dispersion and randomness is likely to creep in between aircraft and exit. Mirchandani and Zou (2007) focus on analysis of adaptive signal control, which also leads to more complex statistics.

The focus of this research is on analytical methods that can be applied to time-dependent scenarios in a piecewise continuous manner, regardless of whether parameters change regularly and smoothly, or randomly and abruptly. This is for partly for reasons of computational efficiency and partly in order to produce repeatable average results. Microscopic and Markov simulation are used here only for creating benchmarks for verifying the analytical methods. However, beyond some level of system complexity, microscopic or event-based simulation may be the only option for getting usefully realistic results[10], although even such methods have limitations (Wood 2012). Establishing how far practically efficient analytical methods can be taken in traffic modelling is left for future research.

[10] Examples could include weather forecasting, nuclear and molecular physics, galaxy formation and collision.

1.2. CONGESTION AND RELIABILITY

1.2.1 The economic costs of congestion and unreliability

The theoretical cost of congestion and delay in road traffic is almost incalculable. It has been estimated by the Automobile Association as £20 billion per annum in the UK (AA 2009), while Eddington (2006) quotes the Department for Transport's more conservative estimate that eliminating congestion would be worth £7-8 billion of GDP. Recent less formal estimates are as high as £30B (LTT 2012). The cost of congestion in the USA has been estimated as $63.1B in 2005, "larger than the GNP of most countries in the world" (Prashker 2008, Schrank *et al* 2010), and by the US Department of Transportation as $1 trillion per annum in total on all roads, equal to the total cash amount spent on road transportation, plus $200M cost of accidents and fatalities (NCTIM 2002). Much of this cost is associated with queuing delay.

However, it is arguable that some delay is an unavoidable price for a cost-effective transport system. Miller (1969) said "The only objective criterion for deciding what is a tolerable level of congestion is an economic one in which the cost of increasing the capacity is matched against the benefits so achieved". The cost of providing more capacity is high and increasing, but it is difficult to put a numerical value on this macro-economic criterion because it seeks to internalise so many factors. Cost-Benefit Analysis (CBA) as used in transport scheme appraisal attempts to do this, although its focus on saving journey time at the expense of pollution, social severance etc has attracted criticism (Kelly 2012). Fundamental questions are now being asked about the sustainability of a lifestyle dependent on motorised transport.

Congestion is often measured pragmatically in terms of the 'Congestion Index' (CI), the ratio of average to minimum travel time. This measure seems to be increasing in popularity, but has certain drawbacks. A more consistent measure would be marginal delay caused to a body of traffic by the addition of one traveller (Taylor 2012). Figure 1.2.1 shows that this is relatively insensitive to the standard and capacity of road considered, and independent of section length, unlike the CI which is 'diluted' on longer sections.

The UK Department for Transport (DfT) has found that a majority of road users do not consider recurrent congestion a serious problem for them personally, and have various tactics for coping with it, but do become frustrated by *unanticipated* congestion (DfT 2005). Figure 1.2.1 shows that sensitivity of delay is greatest around saturation. Variability or unreliability is potentially more amenable to control than absolute congestion, because in principle this can be achieved by improved management of an existing system without changing demand or capacity.

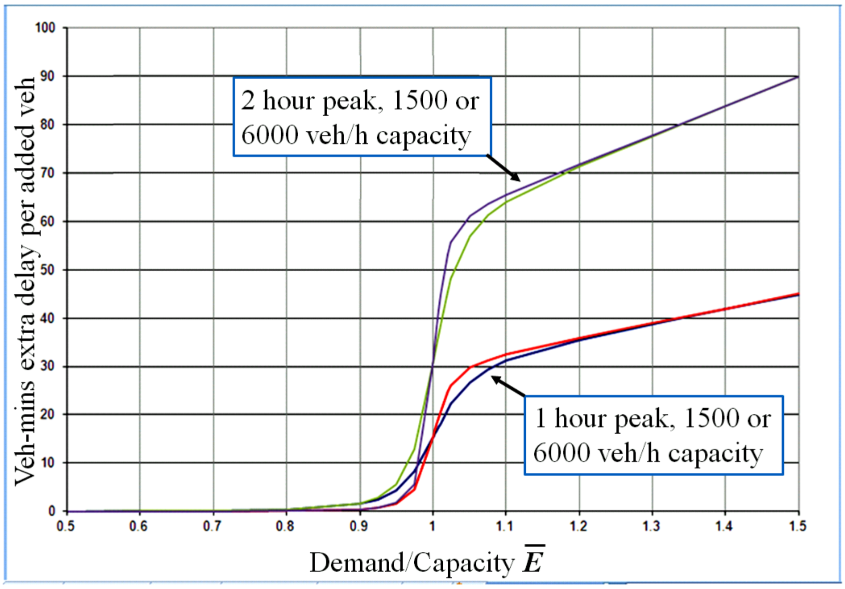


Figure 1.2.1 A possible definition of congestion: vehicle-minutes marginal total delay per additional demand in 1 and 2 hour peaks, as a function of demand/capacity (Taylor 2012)

Value can be ascribed to the reliability of travel time, or the variability of waiting or queuing time. This is more controversial than placing a value on delay, though the latter is arguably *more* subjective since it depends on the reference 'undelayed' travel time. It was once held that the value of travel time uncertainty should be absorbed into the average value of time, all time savings being valued equally. However, the perceived value of uncertainty has been found to vary greatly, 0.55-3.22 times that of average travel time (Gordon *et al* 2001). The 'value' of travel time really reflects opportunity cost, that can be very non-linear for critical journeys such as to an airport or hospital, and this could have a bearing on the valuation of reliability.

The DfT regards journey time reliability as a policy objective. Under the Public Services Agreement (PSA) in the period 2004-2008, it defined the PSA1 measure and target for 'Average Vehicle Delay' (AVD) as the average delay incurred on the most delayed 10% of synthetic journeys on a spanning set of around 100 routes constructed from the around 2,500 links and 4,200 km of the English primary road network managed by the Highways Agency (the HATRIS network). In April 2011, AVD was replaced by an 'On Time Reliability' measure that calculates the percentage of journeys 'on-time' on a link-by-link basis, by comparing average speed with a reference speed. The reference speeds are essentially the same as those used to calculate delay under PSA1, but the new measure is considered easier to relate to the percentage on-time measures used by the rail and air industries. Various other definitions of reliability have been used or proposed, including one based on the assumption that journey time distributions are approximately LogNormal, implying that there is a significant chance of journey times being very much longer than average (Kaparias *et al* 2008).

1.2.2 Recurrent and non-recurrent congestion

Much attention has been and is being focused on preventing or managing incidents, because of their avoidability. While incidents are the largest *single* cause of journey time unreliability, they are responsible for only 10-25% of congestion across Europe (CEDR 2009,2011), while another 10% is estimated to result from road works (construction) and planned events, implying that at least 65% is ascribable to *recurrent* congestion[11]. Non-recurrent congestion tends to involve a sharp drop in capacity, so can be treated by deterministic methods, where the queue size depends essentially on the difference between demand and capacity. Sources of incidental variability tend to be independent of the traffic demand, for example depending on the severity of an incident or the number of lanes closed by road works.

Recurrent congestion tends to recur at the same places. On motorways/freeways these are usually, though not invariably, near junctions, merges or diverges, collectively known as ‘seed points’. Capacity drop may be associated with spontaneous flow breakdown (Kerner and Klenov 2003), although spontaneity has been disputed (Daganzo *et al* 1999). The drop may be only in the range 5-15% (Bertini *et al* 2005), but high demand flows mean large queues can develop. Deterministic methods are therefore applicable, although random queuing theory might help to explain why flow breaks down in the first place (Kühne and Lüdtke 2013). Figure 1.2.2 shows a trajectory of flow and speed measurements at a site upstream of a motorway bottleneck, with typical bent-back appearance (e.g. Carey and Bowers 2011, Heydecker and Verlander 1998, Taylor *et al* 2008 and references therein). A less common cause of non-recurrent congestion is a *moving* bottleneck, such as a large or slow abnormal load, see Figure 1.2.3 overleaf. Analysis is more complex than for fixed bottlenecks, requiring ‘horizontal’ modelling of queues that are both in motion and extended in space, and involving the use of at least basic speed/flow/density relationships (Taylor 2005b,2009 and references therein). Indeed, where traffic is moving at considerable speed, a ‘queue’ in the common sense may not be discernible by an observer.

In contrast to motorways/freeways, urban networks are characterised by: (a) the relatively predictable nature of average delays and the filtering effect of junctions upstream, tending to hold traffic volumes around the saturation level where random variability has the greatest impact; (b) low speeds, high densities and mostly stationary queues, so physical queue lengths are closely related to queue size and ‘horizontal’ methods are not essential; (c) traffic often building up and decaying according to recurrent peak profiles. Since it considers arbitrary profiles, the present work is not limited to such cases, and should be applicable to more complex cases such as traffic flows to or from special events or airports.

[11] These figures originate from 1999 and may have changed, but no more recent figures appear to be available.

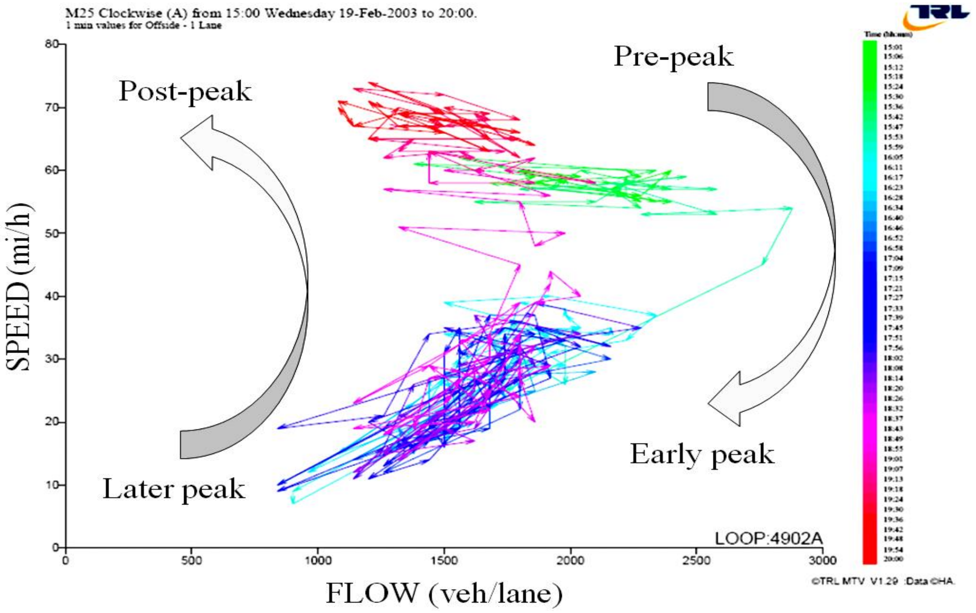


Figure 1.2.2 State and phase transitions 3.8km upstream of a flow breakdown, showing clustering into ‘free flowing’ and ‘congested’ regimes with large short-term variability and uncertain meaning of ‘capacity’. Note also the recovery of speed at constant reduced flow.

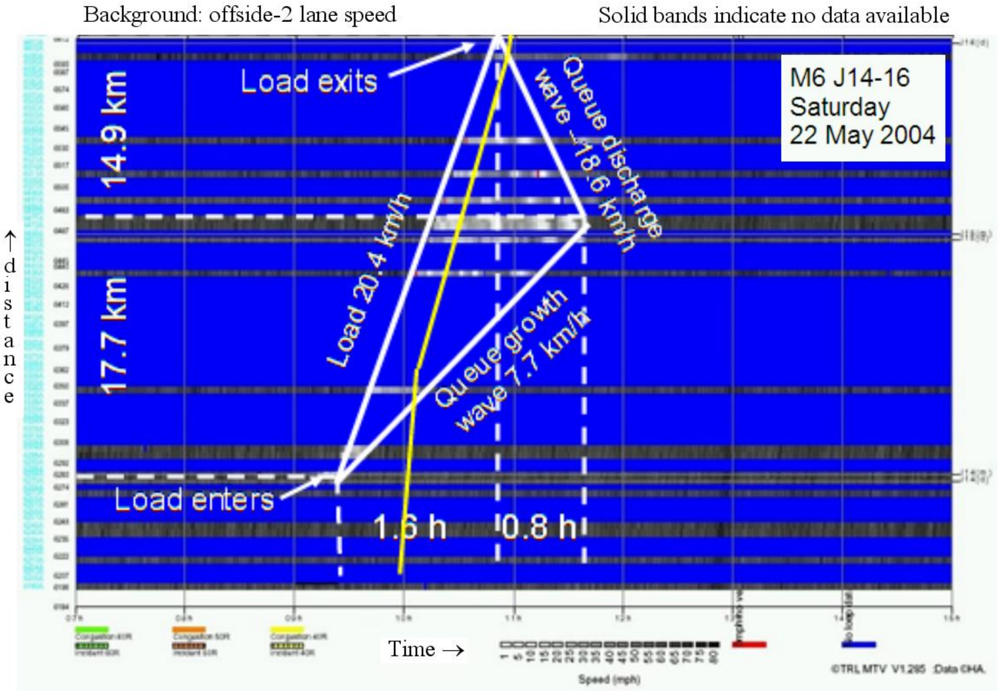


Figure 1.2.3 MTV[12] space/time plot with horizontal queue model superimposed on queue caused by large slow moving vehicle. Low speed traffic is indicated by lighter areas in thin dark bands. Wide blank areas are where loops were disconnected from data network because of limited capacity (no better original is available). See also Taylor (2005a,2009).

[12] MTV and the plots above were developed by Dr Brian Williams and Peter Still at TRL.

1.3. QUEUING IN NETWORK TRAFFIC MODELLING

1.3.1 Modelling journeys

Where traffic has a choice of routes available, prediction usually requires a traffic modelling computer program, ideally with the ability to assign routes according to a principle such as minimum perceived cost, although this surely oversimplifies real route choice if it ignores knowledge, preferences and inertia. Where delay is a monotonic function of volume, the User Equilibrium (UE) principle, as described by Wardrop (1952) and formulated as an optimisation problem by Beckmann (1952), is frequently applied, according to which journeys between the same origin and destination distribute themselves so all used routes have equal cost.

A tree-building computer algorithm for finding minimum cost routes was developed independently by Dijkstra (1959) and Whiting and Hillier (1960). This has since been improved through focusing effort on more productive directions (e.g. Hart, Nilsson and Raphael 1968, van Vliet 1977, Taylor 1989), and recently by origin-based assignment (Bar-Gera 2002, Bar-Gera and Boyce 2003), which greatly improves convergence by calculating routes in terms of whole trees based on origins. The User Equilibrium principle is an idealisation, but alternatives that seek more realism by accepting uncertainty (Stochastic User Equilibrium) or limited information or unwillingness to change routes unless there is a clear advantage (bounded rationality) can be seen and formulated as extensions to the equilibrium concept (Sheffi and Powell 1982, Mahmassani and Chang 1987, Maher 1998).

1.3.2 Modelling queues

Assignment methods need to estimate delays in order to find minimum time or cost routes. Queues occur where the demand for a service approaches or exceeds the effective capacity of the service. Over-saturation is necessarily transient, otherwise queues and delays would increase indefinitely. The diurnal cycle of peaks and rests tends to ensure this, and the ability of traffic to adjust over time justifies the concept of network equilibrium. On short time scales, random fluctuations in demand and service result in a net queue whose mean value in theory tends to a steady-state equilibrium value provided *average* demand is below *average* capacity. On medium time scales queuing occurs because of demand peaks, which while varying greatly in duration, magnitude and steepness are generally limited to a few hours during a day. Catastrophic events, such as serious incidents, severe weather or major disruptions, can lead to queuing on a time scale of days or even weeks (e.g. Global Times 2010) but could be mitigated by adaptation of underlying behaviour and expectations. On the time scale of years we move onto the socio-economic level of land use, investment, policy and ultimately sustainability.

1.3.3 Modelling time-dependent traffic

Understanding time-dependent queuing is essential for accurate journey time prediction in conditions of near-saturation and transient over-saturation. Instantaneous time-dependent functions describing queuing processes accurately can be intractable to integrate or impractical to evaluate numerically. Gradual changes in demand or capacity can be approximated, and step changes accommodated directly, by formulating the problem in terms of appropriate time slices, giving a piecewise continuous profile as in Figure 1.3.1. Any resulting queue, being cumulative, is necessarily continuous. Heydecker and Verlander (1998) point out theoretical inaccuracies associated with this type of approach, in particular non-transitivity, though it is unclear how severely this could affect traffic predictions or economic appraisal.

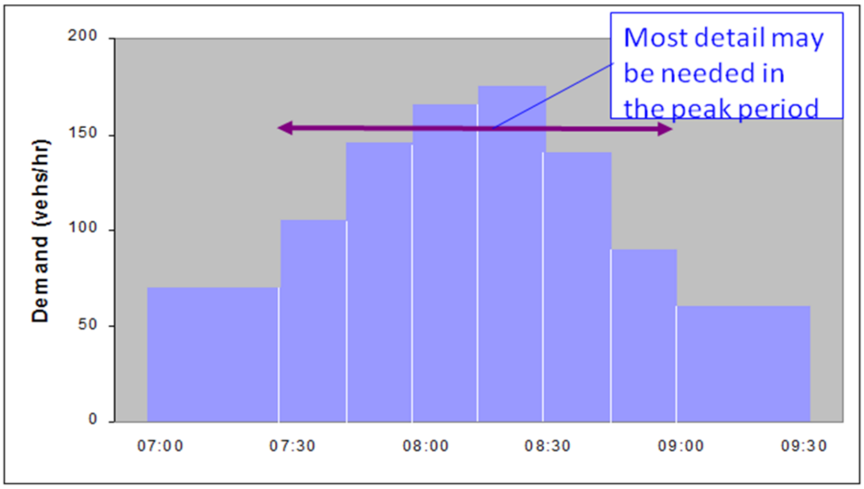


Figure 1.3.1 Illustration of time-slicing applied to a traffic demand peak

1.3.4 Probabilistic queues and probability distributions

Arrivals and service at a queue tend to have an element of randomness, so if the same *average* profiles of mean arrival and service are repeated each day, the queue size at a given time on a given day will be described by a probability distribution (see Figure 1.1.1 earlier). The simplest case is where either or both of the inter-arrival and inter-service intervals is exponentially distributed (variance ≈ mean^2). The number of events in a given period is then Poisson distributed (variance ≈ mean). Variations include uniform service, an idealisation of the effect of a traffic signal, and Erlang-*k* interval distributions where the ratio of variance to mean is modified by a parameter representing arrivals or service in bunches or over several stages.

When calculating queue development from one time slice to the next, ideally a whole probability distribution should be carried over, but in typical modelling only the mean value is available to be passed. The probability distribution of a queue in equilibrium can be quite simple (e.g. geometric), but in a dynamic process it need not be either simple to describe or easy to calculate. In fact, this work will show that most queue size probability distributions during a demand peak are nowhere near an equilibrium form. Having a value for the variance as well as the mean makes it possible in principle to estimate the true shape of the probability distribution, which has not previously been generally possible.

1.3.5 Estimating delay in real networks

Several factors tend to confound the purity of queuing processes in traffic networks:

- Time-varying arrival and service patterns seldom persist long enough or recur consistently enough for their probability distributions to be precisely defined; and granularity of traffic also limits the precision with which distributions can be defined;

- Capacities cannot be calibrated exactly because of the variability of traffic, indeed the definition of capacity is somewhat circular, since it is a theoretical or empirical estimate of whatever maximum thoughput can actually be achieved;

- In mixed traffic, different vehicles use capacity differently, an effect that can be approximated by using PCU factors, that are only indirectly related to vehicle size; e.g. Kimber, McDonald and Hounsell (1985), with references to work by Heydecker (1982) and earlier studies going back to 1964;

- Turning and opposing movements result in complicated local interactions;

- Signals incur ‘lost time’ caused by finite reaction times and acceleration rates;

- ‘Non-separability’ occurs where conditions affecting one traffic movement depend on other movements. Road network traffic is highly interactive in both space and time, although Bar-Gera (2002) claims this is not an issue with origin-based assignment.

As a result, there is a degree of reliance on regression or calibration of parameters of capacity, delay or gap-acceptance models, using experimental data either from observation of actual traffic or from track experiments, e.g. Kimber, McDonald and Hounsell (1986). This work does not aim to deal with these practical issues, except in a highly idealised form in the case of multi-lane queues, but takes as given the queue process statistics that have been established either theoretically or empirically and are widely accepted.

1.3.6 Traffic simulation techniques and computational efficiency

Since around 1990, microscopic simulation of individual events, agents, persons, vehicles etc, has grown in capability and popularity. In principle it is capable of simulating in full detail the behaviour and interactions of individual agents in an environment of arbitrary complexity and realism, and is therefore used extensively in simulations of complex systems in climate, weather, physics and structures. It is particularly valuable where complexity or chaotic behaviour mean that outcomes are emergent rather than analysable. However, in relation to road traffic, microsimulation has been criticised on several grounds. While most of these criticisms can be answered (Wood 2012), persistent ones are that methodology is obscure or proprietary, and that inputs must be varied randomly to generate a representative spread of outputs, while cost constraints may limit the number of repeat runs that can be performed.

Macroscopic, including stochastic, equilibrium methods subsume any randomness into closed formulae and hence produce a single 'typical' result. New equilibrium assignment methods, such as the origin-based method of Bar-Gera (2002), have achieved exceptional levels of convergence as measured by the gap between the solution and the theoretical ideal where alternative used routes have equal cost (Slavin 2012). However, as Wood (2012) points out, and as expressed in a wider economic context by Soros (1987), equilibrium seldom occurs in reality, raising the question whether consistency is a sufficient guarantee of realism. Markovian simulation, which develops a probabilistic description by evaluating all the possible state probabilities over time in small steps, can be applied effectively to a one-dimensional problem like queuing, but could lead to an exponential explosion of calculations in a network unless results are pruned or aggregated to remove contributions of low probability.

At the present time, neither microscopic nor macroscopic method has been declared the 'winner', but microscopic simulation will surely continue to gain ground, thanks to ever increasing computing power and ability to accommodate incremental advances in behavioural modelling. However, an ability to simulate detail need not lead to a proportionately better understanding of whole systems. Even if the present research is not exploited by macroscopic traffic models, it can provide more complete information about how queuing systems can be expected to behave on aggregate, which could be lost in the detail of microscopic simulations. This may particularly apply where the probability distribution of results has an extended or complex shape.

1.4. OBJECTIVES AND METHODS

1.4.1 Objectives

The main objective of this work is to develop an efficient method of predicting the most characteristic queue moments, mean and variance, and the probability of zero queue, for arbitrary time-profiles of demand and capacity, and a range of queuing processes of practical importance. Thereby to provide a tool for evaluation and better understanding of travel time variability and reliability in congested traffic networks and other systems, where equilibrium is inappropriate, with the wider motivation of making time-dependent queuing methods more complete and internally consistent. Benefits will include better understanding of network performance and sensitivities to demand and capacity, and hence better design and management of networks and policies for sustainability. Specific objectives are to:

- Acquire a sufficient understanding of existing methods relevant to the main objective, through reviewing and interpreting relevant past work;

- Implement benchmark programs to validate new methods against established theory and new analytical approximations against simulations;

- Develop a theoretical and computational approach to calculating the time-dependent variance of queues for arbitrary traffic profiles;

- Apply the results of the above to enhance time-dependent closed-form queue estimation methods by improving the accuracy of estimation of the mean, estimating variance, and providing a way to estimate time-dependent probability distributions;

- Produce a computational implementation for the purpose of demonstration;

- In the light of potential applications, pursue simplicity, efficiency, robustness and repeatability, once the basis of the approximations is considered sound.

1.4.2 Scope and application

Equilibrium queue properties tend to depend on the ratio of demand to capacity, not on absolute traffic levels. So in channels of high capacity, queuing tends to be dominated by deterministic effects, making this work most relevant to systems like urban road networks where elements have moderate capacity, there is significant randomness, and long queues can develop on a few channels. However, apart from assuming the standard results of random queuing theory the methods do not presuppose any particular system and so may be more widely applicable.

A typical application could be similar to that addressed by existing traffic modelling methods, an oversaturated peak acting on a junction in a network, made tractable by approximating by a sequence of time slices each characterised by basic parameters of arrival and service rates and statistical properties, that are assumed to remain constant within each time slice. If better resolution is required, more and shorter time slices can be used. Heydecker and Verlander (1998) point out the errors that can occur through time-slicing, yet it not only simplifies computations but arguably reflects the fact that data are often available only as average values in finite time periods. Since queues are cumulative in nature, conservation of quantities and realistic continuity of queue profiles are maintained. The methods developed promise greater accuracy and information about variability and reliability within this context, but are not restricted to it any more than are existing methods of queue and traffic modelling.

1.4.3 Technical approach, methods and validation

The main lines of the technical approach can be summarised as:

- Development or verification of certain basic results, in particular new results for deterministic and equilibrium queue variance, and feasible approximations to probability distributions;

- Integration of mean and variance into a single internally consistent description;

- Heuristic rules or corrections to further reduce estimation error, avoiding the use of empirical parameters as far as possible; and

- Markov chain simulations based on recurrence relations, used to provide time-dependent and equilibrium benchmarks to verify methods and validate them against theory. Experimental validation is outside the scope of this ‘desktop’ research, but reference is made to past experimental verification of some existing methods.

Three methods that could be applied independently are also developed:

- Fitting equilibrium probability distributions to three moments using a doubly-nested geometric template, providing a simple approximation for a range of queue processes;

- Approximate formulae for the mean, variance and other properties of stochastic queues at signals taking account of green period capacity, a factor ignored in some signal calculations, extending the scope of earlier work by various authors;

- Estimating queues on multiple lanes with shared service, or lanes shared by different movements. The case of two lanes with turning movements where ahead streams can use either lane is considered in some detail. Analysis is limited to priority-controlled movements but could be extended to other situations in future research.

Several queue processes are subjected to critical analysis to establish some common principles that help to extend the range of processes that can be approximated. Where new types of function are introduced, they are constrained as far as possible to depend only on natural properties of the system. With two moments, plus the probability of the queue being zero, it is possible to estimate the queue size probability distribution, providing a tool to address detailed questions of reliability and overspill. Mathematical tractability is a criterion, for example for probability functions, tempered by compatibility with realistic constraints and tests of performance compared to alternatives. A demonstrator program is described in Chapter 7.

1.4.4 Practical and computational considerations

Network-wide time-dependent traffic modelling involves simulating typically hundreds of thousands or even millions of trips, and even more transits of links and junctions making up assigned routes, so function evaluations need to be computationally efficient and kept to the minimum necessary for realism. Methods should be resistant to approximation error to the extent that they in some sense conform structurally to the system being modelled.

Road traffic queues have their own particular complexities. They tend to form on a few lanes or channels where there may be some initial choice of lane and possibly opportunity for swapping between lanes ('jockeying') while queuing. Queues can become long and probability distributions consequently extended. Other types of service facility may be designed to avoid this precisely because of the reliability and predictability issues it creates, for example by opening extras service channels according to demand at airport check-ins and supermarket check-outs (airports somehow manage to process most long queues in time for flights!). This kind of adaptability is seldom possible in geographically or service-constrained facilities where demand is difficult to manage, such as road networks and hospitals.

The complicating effect of turning movements has already been referred to. Signal control tends to cause platooning, mitigated somewhat by dispersion, so the statistics of arrival processes can be complex. Variability of individual behaviour and demand, and unpredictability of circumstances affecting supply, favour the use of relatively simple and flexible methods with parameters that can be related readily to data.

1.4.5 Wider applications

Queuing theory has applications in any field in which a service facility has capacity with limited flexibility and may be subject to unpredictable or excess demand. This includes most forms of transportation, health and other services, hospital management, airport operations, telecommunications, call-centres, retail facilities, scheduling and industrial production lines. In addition to time profiles that do not conform to a classic AM/PM peak pattern, these can involve special configurations such as multiple service channels, the possibility of 'defection' or systematic 'censoring' of customers, addition or removal of servers according to demand, and bunching of customers in arrival or service.

Road networks are further complicated by interactions between different movements and routes. The methods described may be applied to these cases to the extent that they can be described in terms of the available parameters.

Some aspects may merit further investigation by researchers, including the deeper meaning of the variance formula, and further development of estimation methods for multi-lane systems and probability distributions. A feature of all the systems studied is that they are 'well-behaved' in the sense that even if unpredictable at the microscopic level, their degree of unpredictability is predictable.

In the real world, many processes appear to be highly sensitive to initial conditions, or conform to higher principles such as the 'power law' in which the probability of an event is roughly inversely proportional to its magnitude, or involve extreme and unpredictable ('black swan') events which overwhelm rational analysis, though it could be argued that these arise from correlations that have not been taken into account, so it may be the problem that is mis-specified. In the last decade or so the study of 'heavy tails' and rare events has become a discipline in itself. While the present research will not address these cases, knowing the boundaries of 'good behaviour' may help to identify cases where it does not apply.

1.5. STRUCTURE AND CONTENT OF THE DISSERTATION

This dissertation is arranged in several main Chapters. Chapter 1 is this introduction. Chapters 2-4 describe the main work and results in queuing theory and approximation. Chapter 5 addresses methods of estimating probability distributions based on the moments calculated by the modelling methods described earlier. Chapter 6 explores certain aspects of queuing on multiple lanes, Chapter 7 covers computational issues and Chapter 8 concludes. More fully:

Chapter 1 Has addressed the motivation of the research and its historical background, scope and applications, and reviewed current traffic and queue modelling methods in general terms. Further references to previous work and existing methods are made at appropriate points in the following Chapters.

Chapter 2 Reviews the core queue models commonly associated with transportation, in particular random-and-oversaturation queues at give-way/yield or signalised junctions, to give coverage. From this basis a new formula for the deterministic variance of queue size is derived. Finally, benchmark methods for modelling arbitrary traffic profiles in time-dependent queuing are described, against which the approximate methods developed later are tested.

Chapter 3 Extends steady-state mean and variance formulae to more general arrival and service processes, with the primary objective of establishing the appropriate parameterisations of the Pollaczek-Khinchin mean queue formula to allow exploitation by efficient time-dependent approximations. A method of fitting doubly-nested geometric probability distributions to known queue moments is described. Dependence of the stochastic signal queue on the green period capacity is discussed and approximations to equilibrium moments derived.

Chapter 4 Develops time-dependent approximations incorporating extensions of the sheared approximation and a new exponential model of queue decay, constrained so as to reproduce the correct equilibrium variance where defined. The method is tested by comparing with Markov simulations of a set of peak profiles with periods of oversaturation against benchmark simulations, for both priority (M/M/1) and signal-type (M/D/1) queuing processes.

Chapter 5 Starting with the example of diffusion approximations, this Chapter develops simplified methods that can be used to estimate queue size probability distributions from known time-dependent queue moments (while a distribution can be estimated at any time in a queue's development the approximating functions are not themselves time-dependent).

Chapter 6 Explores an approach to modelling queuing on multiple lanes, where there may be a choice of lane and sharing of the service processes between lanes, and different turning movements may share a lane. While somewhat detached from the main argument, this differs from multi-channel queuing, and may give additional insight into relevant queuing processes.

Chapter 7 Summarises computational tools and issues associated with exploiting results, specifically to predict the evolution of queue mean, variance and probability distributions together, for arbitrary traffic profiles. A demonstration software tool is described.

Chapter 8 Summarises the results and discusses potential impacts and future work.

Appendices A-F contain derivations, discussions or examples linked to the main text.

Chapters are divided into Sections, the first of which introduces the Chapter and the last of which summarises the main conclusions. Within each Section there can be a number of sub-sections, the first of which can be a form of introduction where this is thought useful, titled ‘Motivation and approach’ to avoid confusion with the main Introduction Section.

Figure 1.5.1 illustrates the logical links between the Chapters.

Figure 1.5.2 visualises the work plan in terms of the flow of information from methodological sources to results and applications.

Figure 1.5.3 pictures the framework of the research in terms of its technical components, arranged in three parallel streams, the main thrust being the time-dependent approximation, with queue processes feeding into it, and producing output that can be post-analysed to estimate probability distributions. Some special cases are brought within the scope to fill perceived gaps in existing methods. To avoid unnecessary fragmentation of topics, some Review is embedded in the sections in which it is relevant, or in Appendices.

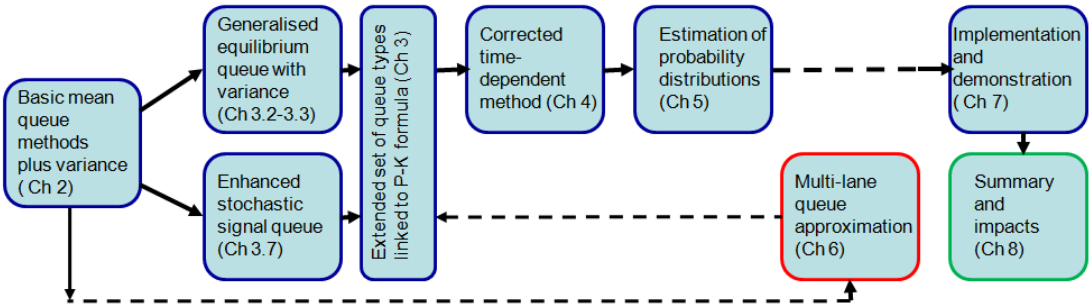

Basic mean queue methods plus variance (Ch 2)
Generalised equilibrium queue with variance (Ch 3.2-3.3)
Enhanced stochastic signal queue (Ch 3.7)
Extended set of queue types linked to P-K formula (Ch 3)
Corrected time-dependent method (Ch 4)
Estimation of probability distributions (Ch 5)
Implementation and demonstration (Ch 7)
Multi-lane queue approximation (Ch 6)
Summary and impacts (Ch 8)

Figure 1.5.1 Linkage of Chapters developing the technical argument

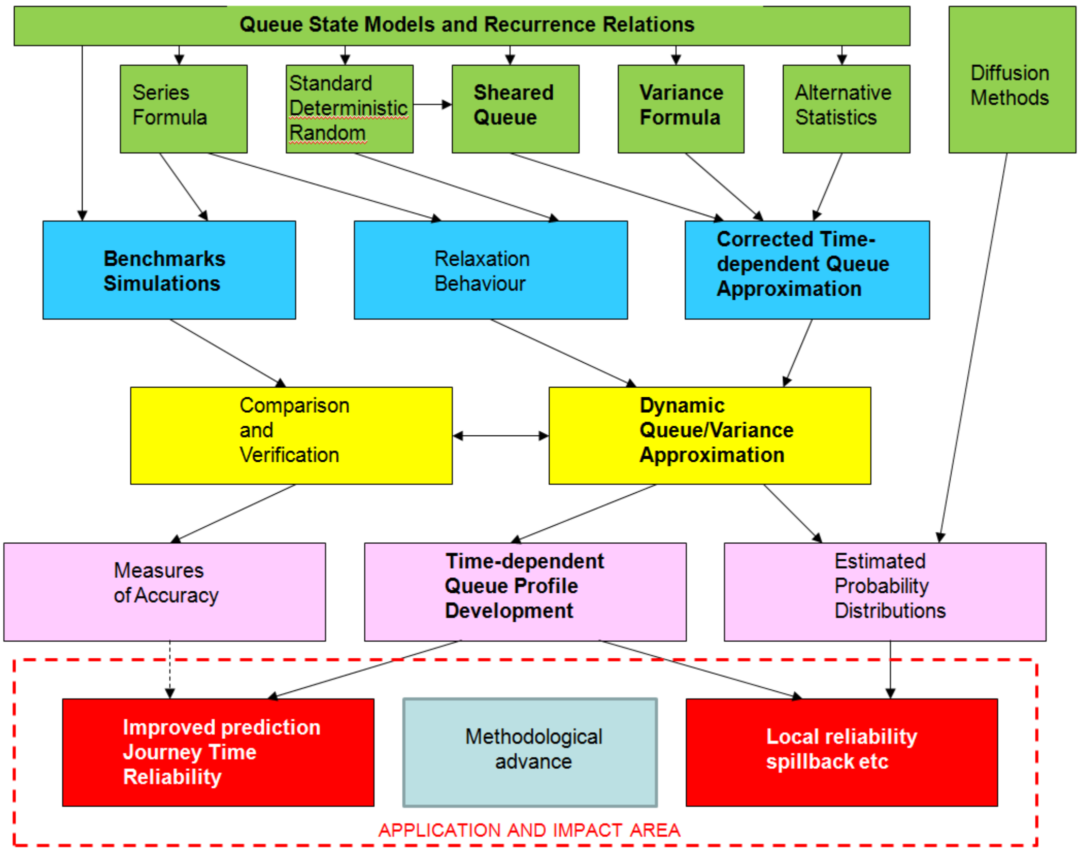

Queue State Models and Recurrence Relations
Series Formula
Standard Deterministic Random
Sheared Queue
Variance Formula
Alternative Statistics
Diffusion Methods
Benchmarks Simulations
Relaxation Behaviour
Corrected Time-dependent Queue Approximation
Comparison and Verification
Dynamic Queue/Variance Approximation
Measures of Accuracy
Time-dependent Queue Profile Development
Estimated Probability Distributions
Improved prediction Journey Time Reliability
Methodological advance
Local reliability spillback etc
APPLICATION AND IMPACT AREA

Figure 1.5.2 Work plan in terms of information flows from sources to objectives and impacts

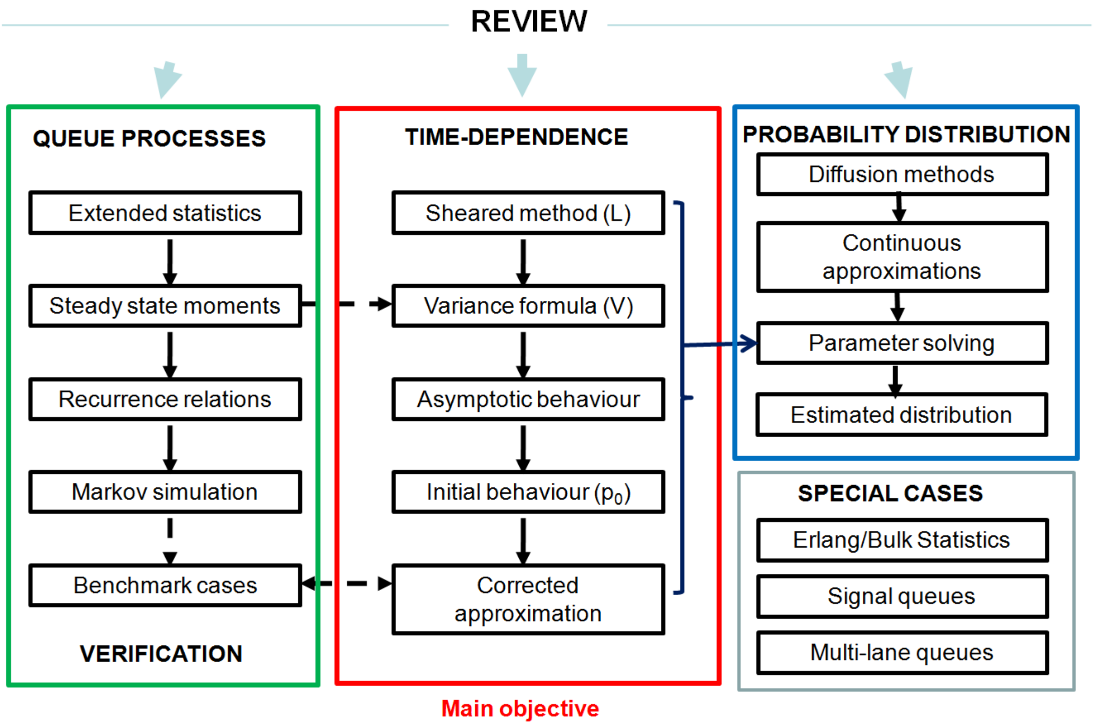


Figure 1.5.3 Framework of the research in terms of technical components

CHAPTER 2: QUEUING PROCESSES

2.1. INTRODUCTION

The core queuing processes relevant to transportation are derived or reviewed, in particular the M/M/1 model relevant to give-way/yield processes and the M/D/1 model relevant to signalised junctions, the deterministic queue formula representing conservation, and the Pollaczek-Khinchin (P-K) formula describing the mean equilibrium queue that takes account of randomness. Recurrence relations between queue states are derived from first principles. In the process, a formula for the deterministic variance of queue size is derived, which is believed to be new. Variations in the statistics of arrivals and service found in many standard works on queuing theory are discussed and linked to statistical parameters in the P-K formula. Benchmark methods are described for modelling arbitrary traffic profiles in time-dependent queuing, using Morse's series formula or Markov simulation, that reveal the details of queue size probability distributions and against which the methods developed later will be tested.

2.2. CONVENTIONS AND DEFINITIONS

Usage related to queuing varies between authors and particular subject areas. A conscious effort has been made to keep to a consistent use of symbols throughout the document. Conventions and variables used are given below, with referenced formulae being 'translated' where necessary. The notation for variables is similar to that used by Kleinrock (1975), Kimber and Hollis (1979) and Bunday (1996). Use of subscripts may vary locally.

2.2.1 Definitions

(2.2.1)

a is a generic symbol for the arrival process, $a(t)$ = arrival time distribution

b is a generic symbol for the service process, $b(t)$ = service time distribution, where

t = time

a and b subscripts will identify quantities associated with arrivals or service

$E(X)$ = the expectation value of any variate X, var(X) = variance of X

λ = the mean arrival[13] rate or demand (constant or function of t)

μ = the mean service rate or capacity (constant or function of t)

s = saturation flow, or conventionally a variable associated with Laplace transformation

[13] In general queuing or 'renewal' theory, 'arrivals' are often referred to as 'renewals'. This term is particularly appropriate to cyclic processes, but less so to traffic processes where each arrival is a new individual.

$\rho = \lambda/\mu$ the demand intensity[14]

τ_{re} = stochastic relaxation time towards equilibrium $(1/\mu)(1-\surd\rho)^{-2} =^{\mathrm{def}}(1/\mu)\tau_{re1}$

I_a = index of dispersion of arrivals (after Kendall, Heydecker) = $1/r$ ($r\geq1$)

I_b = index of dispersion of service (in quoted formulae only)

c_b = coefficient of variation of service = $1/\surd m$ ($m\geq1$)

where r and m are used to represent Erlang parameters of arrivals and service distributions respectively, rather than the usual k

$C \equiv \frac{1}{2}\left(1+c_b^2\right)$ is the 'randomness coefficient'

c_a = coefficient of variation of arrivals (seldom used)

I = unit-in-service parameter, normally 0 or 1

$I^* =_{def} I + \frac{1}{2}\left(I_a - 1\right)$

u = instantaneous utilisation (time dependent)

x = average utilisation or degree of saturation over a time period (time dependent)

L = mean queue size (time dependent)

D = average queue or delay per unit time over a time period

V = queue variance (time dependent)

W = variable related to second moment: $V+L(L+\delta)$ where $\delta\leq1$

L_0, V_0, W_0 = initial values of the above

L_e, V_e, W_e = equilibrium (steady-state) values of the above

L_x, D_x, V_x, W_x = specific forms or estimates of the above

g, c, s, $G=gs$, $\Lambda=g/c$: signal green, cycle time, saturation flow, green capacity, ratio

p_i = probability that a queue is in discrete state i, so $\sum_{i=0}^{N\sim\infty} p_i = 1$

$p(x)$ = probability density of continuous variate x, where $\int_0^\infty p(y)dy = 1$

π_i may be used to represent the ith term of the Poisson distribution

[14] Except in a steady state, the term'traffic intensity' is ambiguous, as 'traffic' could be interpreted as throughput.

2.2.2 Conventions in formulae

The symbol μ is sometimes used to represent *service time interval,* but here it is reserved for *capacity* (service rate) which turns up more often in the transportation context. Likewise λ is reserved for arrival rate, because this often appears with μ in the above context, so Λ is used to represent the ratio of signal green to cycle time. Time can be expressed in any convenient units, such as mean service interval, since results never depend on time *t per se* but on throughput or throughput capacity μt. However, capacity μ is left explicit for two reasons.

First, it is not common practice to express time in units of the service interval. Doing so could cause confusion when comparing results with other sources. Second, the instantaneous capacity *rate* is a meaningful property that can depend on instantaneous and material (e.g. geometric) factors rather than merely reflecting throughput over a particular time period. This does not normally affect results, since the rate of change of capacity is not normally involved in formulae, but it could cause confusion where a time-sliced problem is being evaluated. In some calculations, however, it may be convenient to set μ=1 so that t is effectively expressed in units of service interval and is numerically equivalent to throughput capacity. It may also sometimes be convenient work in multiples of the stochastic relaxation time τ_{re}.

D G Kendall in 1953 invented a notation for labelling queue processes, employed here in a simplified and slightly extended form:

$$\mathrm{A}r/\mathrm{B}m/n/N[G] \qquad \text{where A and B can be} \qquad (2.2.2)$$

M = Markovian, random memoryless process
D = Deterministic, or uniform process
E = Erlang, a random process with a modified shape parameter

r, m are respectively Erlang parameters of the arrival and service processes, omitted if equal to 1 (see definitions);

n is the number of parallel independent service channels;

N is the maximum size of the queue, omitted if effectively infinite (see Section 2.5 later);

G is the author's addition, omitted if equal to 1, representing green phase capacity in a signal queue (see definitions) and used to identify an extended form of M/D/1 (see Section 2.4 later).

‘E*k*’ is interpreted as Erlang-*k*, where the distribution of intervals between arrival or service events is given by equation (2.2.3), λ representing the event *rate* and *k* is the shape parameter (particularlised to *r* or *m* as above). If *k* is allowed to be non-integral this generalises to the Gamma distribution (2.2.4):

$$Erlang(k,\lambda,t)=\frac{k\lambda(k\lambda t)^{k-1}e^{-k\lambda t}}{(k-1)!} \qquad (2.2.3)$$

$$Gamma(k,\lambda,t)=\frac{k\lambda(k\lambda t)^{k-1}e^{-k\lambda t}}{\Gamma(k)} \qquad (2.2.4)$$

Both distributions have mean *interval*=1/λ, variance=1/(*k*λ^2) (2.2.5)

The Erlang distribution can be interpreted as arrival or service passing through *k* stages each with rate parameter *k*λ. Staged arrivals/service can alternatively be represented as ‘bulk’ service/arrivals, where the ‘customers’ arrive or are served simultaneously, represented by 'M*k*' in Kleinrock (1975). Newell (1982), Medhi (2003) and Gross *et al* (2008) also discuss these processes. The alternative scale parameter θ=(*k*λ)$^{-1}$ may be used, in which case:

$$Gamma(k,\theta,t)=\frac{\left(\frac{t}{\theta}\right)^{k-1}e^{-\frac{t}{\theta}}}{\theta\Gamma(k)} \qquad (2.2.6)$$

For which:

$$Mean=k\theta, Variance=k\theta^2, Mode=(k-1)\theta, Maximum=\frac{[(k-1)/e]^{k-1}}{\theta\Gamma(k)} \qquad (2.2.7)$$

When *k*=1, the Gamma distribution reduces to its simplest form, the exponential distribution, that results in a Poisson distribution of the number of events *i* in a given time interval *t*:

$$Exponential(\lambda,t)=\lambda e^{-\lambda t} \qquad (2.2.8)$$

$$Poisson(\lambda,t,i)=\frac{(\lambda t)^i}{i!}e^{-\lambda t} \qquad (2.2.9)$$

2.2.3 Conventional use of terms

Following advice, ‘model’ is as far as posssible used more strictly than is common, to imply an exact, or at least accepted, mathematical description of a process. For implementations the terms ‘approximation’ and ‘method’ are preferred, even where these are thought to reflect the structure of a process as well as predicting its results to some degree of accuracy.

Following the usage of others, ‘deterministic’ is used here to mean a formula that depends on average quantities independent of their statistics, rather than a formula that does not involve randomness in any way. Hence the ‘deterministic queue formula’ does not exclude the possibility that it depends implicitly on random behaviour, but does represent conservation of a quantity, which is necessarily exact.

Generally ‘equilibrium’ is preferred to ‘steady state’ since it is not restricted to the absence of change, but includes dynamic conditions resulting from balance of forces. In practical queuing each tends to imply the other, so it may be read as shorthand for ‘steady state equilibrium’.

2.3. BASIC QUEUE PROPERTIES INCLUDING NEW VARIANCE RESULT

2.3.1 Motivation and approach

In a general queuing process, arrival and service rates can vary with time. This is helpful conceptually but can be difficult to evaluate unless simplified. The simplest queue process is one whose arrivals and service headways are random around known mean rates, i.e. exponentially distributed inter-arrival and service times. Provided average demand is below average capacity the queue tends to an equilibrium condition with a constant probability distribution and moments (at any given moment the queue is not in a steady state). Time-dependent (deterministic) and equilibrium properties of the queuing process, including a new formula for time-dependent variance, can be derived from recurrence relations between queue states, obtained using standard methods. Detailed derivations are given in Appendix A.

2.3.2 Deterministic queuing

Newell (1982) analyses the development of a *deterministic* queue through saturation and through a mild (undersaturated) peak, using a fluid approximation. With modified notation, the mean and variance are calculated by:

$$L(t) \approx L_0 + \int_0^t [\lambda(y) - \mu(y)] dy \qquad (2.3.1)$$

$$V(t) \approx V_0 + \int_0^t [I_a \lambda(y) + I_b \mu(y)] dy \qquad (2.3.2)$$

where λ and μ are arbitrary mean arrival rate and capacity functions and I_a and I_b are stated to be the indices of dispersion of the arrival and service interval distributions. 'Deterministic' in this context means the calculation involves only the means of the random variates, *not* that the system does not involve randomness, and (2.3.1) at least represents conservation of number.

In principle these simple integral formulae allow moments of a queue to be calculated at any time. However, the initial mean values raise an issue, hence the approximate equality signs. For the queue, the single value L_0 disguises the possible role of an initial probability distribution, that logically ought to affect the early development of the queue, though not the final value if the distribution converges ergodically (forgetfully) to a steady-state equilibrium value. This is possible because the negative sign within the integral allows it to both increase and decrease.

In reality, particularly in under-saturated conditions, the throughput rate depends on the queue size, because random variation means there can be periods when the queue is empty, during which capacity is unutilised. Unless this is the case it is not possible for the integral in (2.3.1) to converge naturally to a finite steady-state value. Therefore, (2.3.1) is replaced by the modified equation (2.3.3) where *u*(*t*) is the utilisation of capacity or the degree of saturation, which lies in the range [0,1]. Equality now holds because *u* can absorb the effect of the initial probability distribution as well as modulating convergence to equilibrium:

$$L(t) = L_0 + \int_0^t [\lambda(y) - u(y)\mu(y)]dy \qquad (2.3.3)$$

However, there is a problem with the variance (2.3.2), since this can only increase, regardless of the form of *u*(*t*) and μ(*t*). So while (2.3.1) can form the basis of a realistic time-dependent model of the mean queue, (2.3.2) cannot do the same for the variance.

If λ and μ are constant, then the only variable function on the RHS is *u*. Defining the function *x*(*t*) to represent the average value of *u* over [0,*t*], (2.3.3) can be rewritten as:

$$L(t) = L_0 + (\rho - x(t))\mu t \qquad \text{where} \qquad (2.3.4)$$

$$x(t) = \frac{\int_0^t u(y)dy}{t} \qquad (2.3.5)$$

2.3.3 Steady-state equilibrium

This possibility of convergence is essential to equilibrium theory. Ergodicity is normally assumed, meaning loosely that the queue converges to an equilibrium value independent of its early states. For computational simplicity it is now assumed that the arrival and service capacity rates λ and μ are constant over finite time slices, and that these can be cascaded to approximate time-variation, although Heydecker and Verlander (1998) point out the errors inherent in this.

M/M/1 is the simplest queuing process, in which units/customers arrive and are served in a single channel randomly according to exponential distributions of inter-arrival and inter-service times at constant mean rates, and Poisson distributions of the number of arrivals or service opportunities in a given period. It is ergodic and ‘forgets’ its initial state but only as O(1/*t*), a fact that can be exploited. It is ‘Markovian’ in that the probability of a state change depends only on the current state, not on what has gone before. It is also a ‘Birth-Death process’, in which only movements from the current state to its immediate neighbours need be considered.

Another property that applies to all queues of practical interest in traffic is 'Chapman-Kolmogorov homogeneity' where the probability of any current state can be calculated by multiplying the probabilities of initial states by the transition probabilities along all paths between them. This ensures that distributions can be developed progressively.

Standard works usually draw a distinction between the number in the system and the number in the queue, or total time in the system and waiting time for service. Mean waiting time is often calculated rather than queue size. Most sources use the convention:

Total_time_in_system = Waiting_time + Time_in_service (2.3.6)

This is reflected in decomposition of the queue into two components:

Total_queue = Waiting_queue + Unit_in_service_component (2.3.7)

The two are related by Little's formula (Little 1961), which says:

Total_queue = Total_time_in_system * Departure_rate (2.3.8)

Note that in the steady state: Departure_rate = Arrival_rate = Service_rate * Traffic_Intensity

In traffic modelling the 'in service' distinction is mostly academic, because waiting time and time in service are equally unproductive, and the concerns are presence of queues and total delay. Service time is likely to increase in importance where it is highly variable, as at toll plazas and other borders, bank/post-office counters, airport check-ins and loading facilities etc. At traffic signals time in service tends to be ignored or subsumed in a parameter adjustment, because saturation flow is much greater than the average capacity over the cycle. Generally, therefore, the total number and total time in the system are of primary interest.

Symmetry with a change of state viewpoint implies that far enough away from the empty state the probability distribution of a queue process in equilibrium should be independent of its state apart from a scaling factor, so that its gradient should be proportional to its value, implying an exponential (geometric if discrete) form. Ergodic and Markovian properties imply that to approach equilibrium, the difference between the mean system state and the steady state must eventually decrease with time. Independence of scaling with a shift in time viewpoint implies that its rate of change should also be proportional to its value[15]. Therefore, the long-term behaviour of the system is likely to resemble exponential relaxation to equilibrium.

[15] This will not hold for non-equilibrium states, as seen later with drift/diffusion models where the rate of change depends on gradients of the probability distribution.

2.3.4 The simplest random queue, M/M/1/1, and the significance of utilisation

The simplest random equilibrating queue problem is the 'parking space'[16] where there are only two possible states – unoccupied (0) and occupied (1). Its value is in illustrating the relaxation process, and also revealing the nature of utilisation. Assuming constant arrival and service rates the process can be visualised through the state transition diagram, Figure 2.3.1.

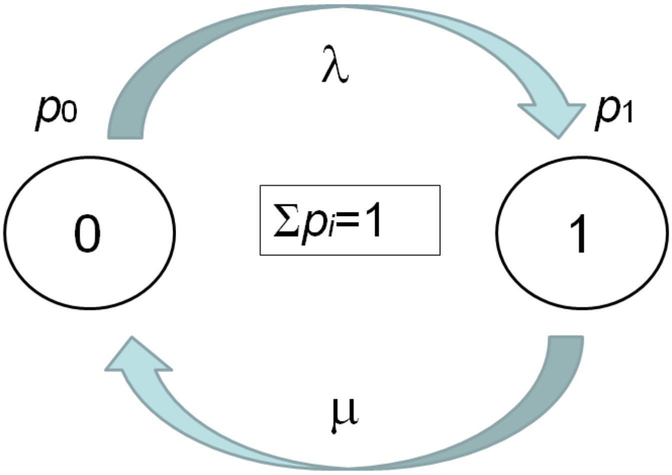


Figure 2.3.1 Basic state transition diagram for M/M/1/1 'parking' process

The rate of change of p_0 is the difference between the rate of transitions out of state 0, and the rate of transitions back into state 0 that equals the rate of transitions out of state 1, so:

$$\frac{dp_0}{dt} = \mu p_1 - \lambda p_0 \tag{2.3.9}$$

Since the $\{p_i\}$ sum to 1, p_1 can be eliminated giving:

$$\frac{1}{\mu}\frac{dp_0}{dt} = 1 - (1+\rho)p_0 \tag{2.3.10}$$

If some function f satisfies:

$$p_0 \equiv f(t) + \frac{1}{1+\rho} \tag{2.3.11}$$

[16] I am grateful to Prof. B G Heydecker for suggesting that this example be included.

then equation (2.3.10) becomes (2.3.12) whose solution is (2.3.13):

$$\frac{1}{\mu}\frac{df(t)}{dt} = -(1+\rho)f(t) \qquad \text{where} \qquad (2.3.12)$$

$$f(t) = K\left(\frac{e^{-(1+\rho)\mu t}}{1+\rho}\right) \qquad \text{(for some constant } K\text{)} \qquad (2.3.13)$$

The mean occupancy of the system is just p_1, so after some manipulation:

$$p_1(t) = L(t) = L(0)e^{-(1+\rho)\mu t} + \left(\frac{\rho}{1+\rho}\right)\left(1 - e^{-(1+\rho)\mu t}\right) \qquad (2.3.14)$$

where $L(0)\in[0,1]$, so the equilibrium value when $t\to\infty$ is:

$$L_e = p_{1e} = \frac{\rho}{1+\rho} \qquad \text{with} \qquad p_{0e} = \frac{1}{1+\rho} \qquad (2.3.15)$$

Equation (2.3.14) shows that the system relaxes exponentially from its initial value to its equilibrium value with a time constant equal to $\mu^{-1}(1+\rho)^{-1}$, and provided that arrivals do not exceed capacity it achieves a maximum mean occupancy of ½ when $\rho\approx1$.

However, there is something wrong with this model since it is inconsistent with (2.3.4), which in order to remain finite as $t\to\infty$ requires that $u_e=x_e=\rho$ at equilibrium. Since the utilisation represents the proportion of time capacity μ is fully used, capacity being instantaneously unutilised if the queue is zero and fully utilised if the queue is occupied[17], the average probability of the queue being empty can be identified with the complement of the utilisation:

$$p_0 = 1-u \qquad p_{0e} = 1-u_e = 1-x_e = 1-\rho \qquad (2.3.16)$$

Equation (2.3.15) is incompatible with this. The explanation is that Figure 2.3.1 has left out those customers who arrive to find the parking space full, and depart without contributing to the 'queue', which in this case represents occupancy of a single parking space. Figure 2.3.2 allows for this.

[17] 'Occupied' means a customer at the front of the queue. It is a frequent cause of frustration that the stop line serves only the first customer in line so any behind it have no influence.

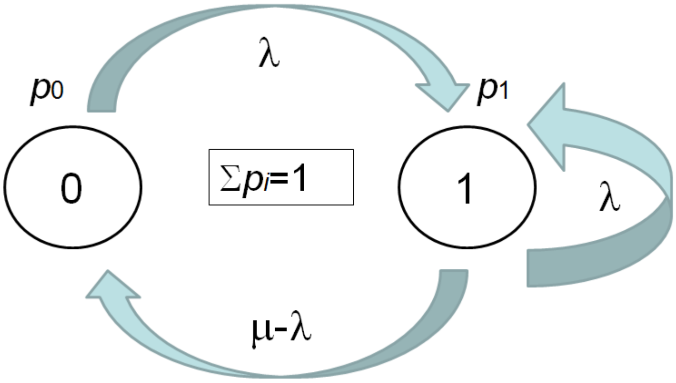


Figure 2.3.2 Completed state transition diagram for M/M/1/1 'parking' process

Customers who arrive when the parking space is inconveniently occupied are said to 'defect'. They could also be systematically 'censored' by some external influence, such as by excluding every other arrival, but it assumed here that they simply react to availability of parking. In this completed system (2.3.9) is replaced by (2.3.17) or in terms of demand intensity (2.3.18):

$$\frac{dp_0}{dt} = (\mu - \lambda)p_1 - \lambda p_0 \qquad (2.3.17)$$

$$\frac{1}{\mu}\frac{dp_0}{dt} = (1-\rho) - p_0 \qquad (2.3.18)$$

The solution is got by substitution and integration in a similar way to (2.3.11-13):

$$p_1(t) = L(t) = L(0)e^{-\mu t} + \rho\left(1 - e^{-\mu t}\right) \qquad \text{so letting } t \to \infty \qquad (2.3.19)$$

$$L_e = p_{1e} = \rho \qquad\qquad p_{0e} = 1 - \rho \qquad (2.3.20)$$

Hence (2.3.4) is satisfied.The wrong solution (2.3.15) was the result of mis-specification, but in other systems the value of p_{0e} can disagree with (2.3.16) through difference of interpretation (see M/D/1 later). Be that as it may, no description of a queuing process that obeys (2.3.4) can be considered realistic unless it also satisfies (2.3.16), so this acts as a 'reality check' on any practical formulation, and any different result calls for explanation and resolution.

2.3.5 The simplest realistic traffic queue: M/M/1/∞

Given that mixing and dispersion tend to randomise traffic at least at the microscopic level, M/M/1 is one of the most useful processes in traffic modelling, where arrivals and service are random as at an idealised priority junction. Although roads often have more than one lane, as long as it is possible to switch lanes freely a single server model can be assumed (a multi-lane model is considered in Chapter 6). Real road sections have finite storage capacity, but as long as a queue is unlikely to exceed this, infinite storage can be assumed, so the '∞' is usually dropped. As this is a 'birth-death' process with a theoretically infinite number of states a general recurrence relation between adjacent states is set up, and as there is no particular time scale (such as a finite green period) transitions can be evaluated for an infinitesimal interval:

$$p_i(t+dt) = p_{i+1}(t)(1-\lambda dt)\mu dt + p_i(t)\left(\lambda\mu dt^2 + (1-\lambda dt)(1-\mu dt)\right) + p_{i-1}(t)\lambda dt(1-\mu dt) \tag{2.3.21}$$

Terms in infinitesimal dt^2 vanish, and introducing $\rho=\lambda/\mu$ as before, the relations reduce to the following, noting that there is no state below the absorbing zero state:

$$\frac{1}{\mu}\frac{dp_0}{dt} = p_1 - \rho p_0 \tag{2.3.22}$$

$$\frac{1}{\mu}\frac{dp_i}{dt} = p_{i+1} - (1+\rho)p_i + \rho p_{i-1} \tag{2.3.23}$$

Where i is limited to some finite N, the formula for dp_N/dt is truncated, but this is not relevant to the present argument. When summed, terms cancel between successive expressions, and since the probabilities sum to 1, equations (2.3.22-23) add up to zero on both sides.

The rate of change of the mean queue is calculated by taking the first moment of equations (2.3.22-23), with some cancellation of terms and utilisation defined as before:

$$\frac{1}{\mu}\frac{dL}{dt} = \frac{1}{\mu}\sum_{i=1}^{N} i\frac{dp_i}{dt} = \rho - 1 + p_0(t) = \rho - u(t) \tag{2.3.24}$$

When this is integrated, the deterministic mean queue formula (2.3.4-5) is obtained:

$$L(t) = L_0 + \mu\left(\rho t - \int_0^t u(y)dy\right) \equiv L_0 + (\rho - x(t))\mu t \tag{2.3.25}$$

Equation (2.3.25) expresses conservation of units/customers/vehicles and so will be inherent in any correctly formulated queuing process. In the parking place model in section 2.3.2, conservation was violated as long as ρ was interpreted as the true demand intensity because customers who defected were not recorded.

2.3.6 Derivation of the deterministic formula for time-dependent variance

In a similar way to the foregoing, evaluating the rate of change of the *second* moment of equations (2.3.22-23) gives initially (see Appendix A for derivations):

$$\frac{1}{\mu}\frac{d\left(S^2\right)}{dt}=\frac{1}{\mu}\sum_{i=1}^{N} i^2\frac{dp_i}{dt}=2(\rho-1)L(t)+\rho+1-p_0(t)-\rho(2N+1)p_N \qquad (2.3.26)$$

Letting $N\to\infty$, using (2.3.24) to eliminate p_0 in favour of dL/dt, integrating, and using the definition of variance $V=S^2-L^2$, the following expression for variance results:

$$V=V_0+L_0(L_0+1)+2(1-\rho)(L_e-D)\mu t-L(L+1) \qquad \text{where} \qquad (2.3.27)$$

$$D(t)=\frac{1}{t}\int_0^t L(y)dy \quad \text{is the } \textit{average} \text{ mean queue or delay over } [0,t] \qquad (2.3.28)$$

L, D and V all being functions of time, and L_0, V_0 being initial values at t=0. On making the natural definition $W\equiv V+L(L+1)$ (sum of first two moments), a tidier formula results:

$$W=W_0+2(1-\rho)(L_e-D)\mu t \qquad (2.3.29)$$

Equations (2.3.27-29) constitute the key new result that will be exploited here. As the structural similarity to (2.3.4-5) is immediate, one is tempted to conjecture that:

(1) The result applies to *all* queues, not just M/M/1

(2) Some property of traffic is conserved, though not simple units of traffic

(3) Analogous formulae for higher moments may exist.

Conjecture (1) will be reinforced later when it is shown that the variance formula is also satisfied by the M/D/1 process and can be derived from diffusion models (see Chapter 5). Conjecture (2) does not affect the present discussion but is touched on informally in Appendix A. The purpose here is to explore the implications for calculating queue variance and hence reliability. The exact shape and tail of a probability distribution depends also on higher moments (3), in particular skewness and kurtosis, but the natural asymmetry of a queue distribution means that large skewness is likely to be associated with large variance. Kurtosis might have a stronger association with 'heavy tails', but these and 'rare events' may more usefully be approached directly. However, it is conceivable that a formula for skewness might be used to constrain queue development. This could be a topic for further research.

2.3.7 Steady state invariants of M/M/1

Steady-state moments can be got from the recurrence relations by setting the rate of change of the *next highest* moment to zero. Equation (2.3.24) yields the utilisation immediately:

$$u_e = 1 - p_{0e} = \rho \tag{2.3.30}$$

Using (2.3.22-23) inductively with (2.3.30) gives the steady-state probability distribution:

$$p_{ie} = (1-\rho)\rho^i \tag{2.3.31}$$

Setting the rate of change of the *second* moment to zero in (2.3.26) and substituting for p_{0e} from (2.3.30) yields the equilibrium mean queue size:

$$L_e = \frac{\rho}{1-\rho} \tag{2.3.32}$$

To get the equilibrium variance it is necessary to evaluate the rate of change of the *third* moment, whence using substitutions from (2.3.30-32):

$$V_e = \frac{\rho}{(1-\rho)^2} \tag{2.3.33}$$

These are all well-known results, and there are other ways of extracting them for more general queue processes (e.g. Kleinrock 1975, Bunday 1996), but they are *not* obtainable from the deterministic formulae of the same order. This is not surprising since the deterministic formulae have to apply equally to all queuing processes.

2.3.8 Exact Series formulation of the time-dependent M/M/1/N queue

To describe a queue process completely, the transient queue size probability distribution must be specified for all time from any starting condition. The M/M/1 queue was the first queue process for which a closed-form solution for the transient probability distribution was derived (Morse 1955, Clarke 1956, Morse 1958), in various but equivalent forms involving exponential, trigonometric or Bessel functions, and later Generalised Q-functions (e.g. Cantrell 1986, Cantrell and Ojha 1987, Cantrell and Beall 1988). Sharma (1990) gives an alternative formula involving binomial coefficients. More recently, Griffiths *et* al (2005) have described the transient solution to the M/E*k*/1 process, where service is generalised to the Erlang-*k* distribution. Kleinrock (1975) finds the complexity of the transient solution “most disheartening”, despite its representing the “simplest interesting queuing system”. For present purposes, the following discrete form quoted by Kimber and Hollis (1979), using only standard exponential and trigonometric functions, is found computationally convenient. If the maximum queue size is *N*, the probability ${}_m p_n$ that a queue initially of size exactly *m* has size *n* at time *t* is:

$$ {}_m p_n(t) = \frac{(1-\rho)\rho^n}{1-\rho^{N+1}} + \frac{2}{N+1}\rho^{n-m}\sum_{i=1}^{N}\frac{S_{i,m}S_{i,n}}{X_i}\exp(-X_i\mu t) \qquad (2.3.34)$$

where

$$X_i = 1+\rho-2\sqrt{\rho}\cos\left(\frac{i\pi}{N+1}\right) \qquad (i,j\in 1..N) \qquad (2.3.35)$$

$$S_{i,j} = \sin\left(\frac{ij\pi}{N+1}\right) - \sqrt{\rho}\sin\left(\frac{i(j+1)\pi}{N+1}\right) \qquad (2.3.36)$$

The queue size probability distribution at time *t*, given that at t_0, is obtained by convolution of the transition probabilities with the initial distribution:

$$p_n(t) = \sum_{m=0}^{N} {}_m p_n(t-t_0)p_m(t_0) \qquad (2.3.37)$$

The first term in (2.3.34) represents the equilibrium distribution. Although equilibrium is undefined for $\rho\geq 1$, the formula as a whole remains valid, although the maximum *N* needed for accuracy is no longer bounded. As in simple cases (2.3.14, 2.3.19) the description involves exponential relaxation, but (2.3.34) behaves like a linear superimposition of processes with different time scales, as if each queue state evolves independently of all the others, like waves propagating in a linear medium. As in (2.3.14), there is effective censoring/defection of arrivals if the queue is already ‘full’, so (2.3.4) is *not* exactly satisfied for finite *N*.

Efficiency and accuracy of calculation depends on the choice of *N* in equations (2.3.34-37), as well as restraint in the value of *t* when ρ>1. For ρ>1 the highest possible resolution is advisable, subject to powers of ρ not exceeding the available computational range. For ρ<1 an adequate minimum value for *N* is found empirically to be:

$$N+1 \geq L_0 + 7/(1-\sqrt{\rho}) \qquad \text{when } \rho< 1 \qquad (2.3.38)$$

In (2.3.34-36) all terms apart from the exponentials are independent of *t*, so can be precalculated for ρ, while the *sin* and *cos* functions depend only on *N*. If (*N*+1) is chosen to be a power of 2, with *N* sufficiently large that p_N is expected to be negligible, the *sin* and *cos* terms can be precalculated and stored for all *i* and *j*. Values for smaller *N* can then be interpolated by selecting from the array. Calculation of p_n can be suppressed once the probability falls below some lower limit, and the remaining probabilities normalised to sum to 1. For example, if the lower limit is set at 0.0000001, and this occurs before *i* reaches a maximum (*N*+1) of 1024, then the maximum error introduced by truncation is 0.01%.

2.3.9 Characteristic relaxation times

Relaxation times are useful for making approximations that have the right dynamic or transient behaviour. When *N*→∞, or the *cos* term approaches 1, (2.3.35) simplifies to the inverse of the commonly quoted relaxation time of a random queue towards the steady-state, which also happens to be the longest time scale:

$$\tau_{re} = \frac{1}{\mu}\left(1-\sqrt{\rho}\right)^{-2} \qquad (2.3.39)$$

Morse (1955) quotes a value twice this, but the practical difference is small where τ_{re} is used as a time scale indicator on a logarithmic scale. In the early stages of development the shortest time scale dominates, got by setting the *cos* term to zero, being already familiar from the ‘censored’ parking place model (2.3.14):

$$\tau_{r0} = \frac{1}{\mu}\left(1+\rho\right)^{-1} \qquad (2.3.40)$$

Otherwise, (2.3.35) ranges in value between these two, which diverge as ρ approaches 1. For Erlang-*m* service, equation (2.2.3), where the variance of the service time is divided by a factor *m* as a result for example of bunching, (2.3.40) generalises to (2.3.41), although this breaks down in the case *m*=∞ representing uniform service:

$$\tau_{r0}(m) = \frac{1}{\mu}\left(m+\frac{\rho}{m}\right)^{-1} \qquad (2.3.41)$$

2.3.10 Lags and loops

It is observed that the variance of travel time or delay tends to lag the mean. Addison (2006) demonstrates this behaviour through a diffusion approximation (see later in Chapter 5). This effect leads to relationships between the variabilities and means of travel times that resemble anti-clockwise loops, as illustrated by Arup, Bates *et al* (2004), and Fosgerau (2010). It is also observed that the peak of a queue lags that of the demand profile causing it. This is intuitively obvious because a deterministic queue caused by transient oversaturation will go on rising as long as the demand exceeds the capacity, including a period while the demand is falling from its peak. Equation (2.3.24) expresses this formally. The rate of change of the queue is positive as long as the demand intensity ρ exceeds the utilisation. If ρ>1, since $u \leq 1$ always, oversaturation will cause the queue to grow even while the demand is falling.

The variance equation (2.3.27) or (2.3.29) has a similar general structure to (2.3.25). The derivative of (2.3.27), making use of (2.3.24) and (2.3.28), is:

$$\frac{dV}{dt} = 2\mu(1-\rho)(L_e - L) - \mu(2L+1)(\rho - u) \qquad (2.3.42)$$

This can also be written as (2.3.43), which for the M/M/1 queue simplifies to (2.3.44):

$$\frac{dV}{dt} = 2\mu(1-\rho)(L_e + .5)\left[1 - \left(\frac{L+.5}{L_e+.5}\right)\left(\frac{1-u}{1-\rho}\right)\right] \qquad (2.3.43)$$

$$\frac{dV}{dt} = \mu(1+\rho)\left[1 - \left(\frac{L+.5}{L_e+.5}\right)\left(\frac{1-u}{1-\rho}\right)\right] \quad \text{(M/M/1)} \qquad (2.3.44)$$

Pre-peak, when the queue is rising, $u<\rho$ and L must be less than L_e, but this need not be a problem if the queue has not had time to approach equilibrium. Within the peak, where ρ>1, $dV/dt>0$ certainly. Post-peak, where ρ<1 and the queue is falling then necessarily $u>\rho$, so the second inner bracket <1, and $dV/dt>0$ as long as in the first inner bracket L is not so large compared to L_e that it overwhelms the second inner bracket making the whole expression to go negative. Considering the state produced during the peak, L_e is unbounded as ρ is in lower neigbourhood of 1, whereas L is bounded by some value less than the cumulative arrivals, so there must be a period *after* demand falls below capacity when L is less than L_e. Hence the peak variance lags the peak mean. The same conclusion is reached for other queue types since their equilibrium formulae invariably include the divisor (1-ρ) ensuring the multiplier in (2.3.43) is >0. Figure 2.3.3 shows hysteresis loops of L against ρ, and V against L, taken from one of a number of over-saturated symmetrical peak cases (J2P4, see Table 2.5.1 later), that lasts 108 minutes including 36 minutes over-capacity and is divided into 9-minute time slices. Addison and Heydecker (2006) prove this effect in theory, and Arup, Bates *et al* (2004) show that it extends to whole networks.

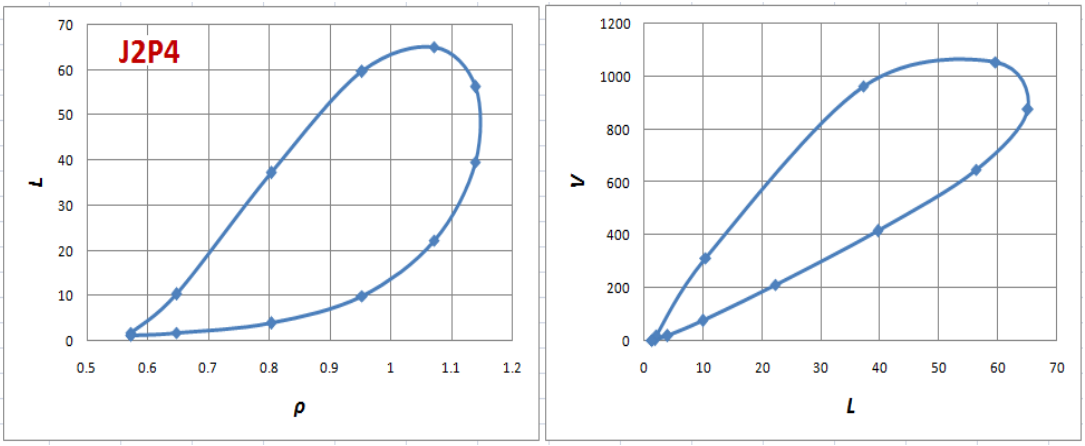


Figure 2.3.3 Lag-loops generated by an oversaturated peak, with points every 9 minutes. Left: mean queue against demand intensity; Right: variance against mean queue.

2.3.11 The Pollaczek-Khinchin mean queue formula

The Pollaczek-Khinchin[18,19,20](P-K) mean queue formula (2.3.45) is the simplest form of a generalisation of the steady-state mean queue formulae given earlier.

$$L_e = I\rho + \frac{C\rho^2}{1-\rho} \qquad \text{where} \quad C \equiv \tfrac{1}{2}\left(1+c_b^2\right) \tag{2.3.45}$$

This is a standard result in text books and extensively used although necessarily an idealisation of real processes. It is derivable either by considering a queuing process with general (or at least Erlang) arrival and service distributions (see later in Chapter 3 where an extended form is derived), or via a Laplace transform approach developed by the same authors and found in most standard works. The parameters I and C are as defined in (2.2.1). C is nominally 1 for M/M/1 where c_b=1 and 0.5 for M/D/1 where c_b=0, but in practical signal models an empirical value in the range 0.5-0.6, such as 0.55, is used (Catling 1977, Branston 1978, Kimber and Hollis 1979, Burrow 1987), reflecting that the service is not perfectly uniform. The form of C in equation

[18] We use the modern English transliteration of the Russian name rather than the French Khintchine.

[19] Félix Pollaczek and Aleksandr Khinchin developed their formula for G/G/1 (‘General’) queues independently in 1930-32, though this has also been credited to Harald Cramér in 1930.

[20] The convention adopted here is consistently to use subscript a for arrivals, and b for departures or service.

(2.3.45) takes account of the service process through its coefficient of variation c_b, but does *not* involve the arrival process (see later in Chapter 3). However, Kimber, Summersgill and Burrow (1986) proposed the alternative (2.3.46) that is trivially valid when ρ and the index of dispersion of arrivals are close to 1:

$$C \cong C^* = \tfrac{1}{2}\left(c_a^2 + c_b^2\right) \qquad (2.3.46)$$

The unit-in-service coefficient *I* does not appear in standard texts and is believed to have been introduced by Kimber, Hollis and/or Summersgill to reflect the contribution of the average time in service at a priority junction. This is neglected at a signal. *I* is conventionally 1 for the M/M/1 queue and 0 for the M/D/1 queue.

2.3.12 The deterministic limit

Equation (2.3.24) contains the time-variable degree of saturation *x*(*t*). In equilibrium this converges to ρ so the shrinking of the bracket compensates for the unlimited growth of the time multiplier. If service is assumed, somewhat simplistically, to be permanently saturated (*x*=1), the queue formula reduces to:

$$L = L_0 + (\lambda - \mu)t \qquad \text{where} \quad \lambda = \rho\mu \qquad (2.3.47)$$

Since this is linear in *t* the delay-per-unit-time (average queue) function is simply:

$$D = L_0 + \tfrac{1}{2}(\lambda - \mu)t \qquad (2.3.48)$$

In the variance formula (2.3.27), since L_e always has (1-ρ) in the denominator, the first term in the central bracket 2(1-ρ)L_e is always defined. In the case of M/M/1 it is just 2ρ, so using (2.3.47-48), noting that all the terms in L_0 cancel:

$$\begin{aligned} V &= V_0 + L_0^2 + L_0 + \left(2\lambda - (\mu - \lambda)(2L_0 + (\lambda - \mu)t)\right)t \\ &- L_0^2 - 2L_0(\lambda - \mu)t - (\lambda - \mu)^2 t^2 - L_0 - (\lambda - \mu)t \\ &= V_0 + (\lambda + \mu)t \end{aligned} \qquad (2.3.49)$$

The final expression in equation (2.3.49) is well known and demonstrates its consistency and that of the more general equation (2.3.2) with the variance formula (2.3.27).

The constant multipliers of time in (2.3.47) and (2.3.49) feature in simple diffusion approximations considered later in Chapter 5 as respectively measures of drift (bodily movement) and diffusion (spreading) of the probability distribution. When more general process statistics are introduced as in the P-K formula (2.3.45), the form of (2.3.47-48) is not affected, but that of L_e is. Rewriting (2.3.45) as:

$$L_e = \frac{I^*\rho + (C-I)\rho^2}{1-\rho} \quad \text{where} \quad I^* = I + \tfrac{1}{2}(I_a - 1) \qquad (2.3.50)$$

where I_a is the index of dispersion of arrivals. Using I^* has advantages that will be apparent later (see Chapter 3), and dropping terms in L_0 that cancel anyway, (2.3.49) is modified to:

$$\begin{aligned} V &= V_0 + \big(2\big(I^* + (C-I)\rho\big)\lambda - (\mu-\lambda)(\lambda-\mu)t\big)t - (\lambda-\mu)^2 t^2 - (\lambda-\mu)t \\ &= V_0 + \big(\big(2I^* - 1 + 2(C-I)\rho\big)\lambda + \mu\big)t \end{aligned} \qquad (2.3.51)$$

So the net effect is to apply a factor to the arrival rate λ, but this depends on the service process as well as the arrival process. For an M/D/1 signal-type queue (I^*=0, C=0.5) the factor on time reduces to (ρ-1). This is positive since the assumption of deterministic growth implies ρ>1. However, the significance of (2.3.51) is unclear since neither the factor on time nor the equation as a whole seems to correspond to anything familiar.

2.3.13 Accommodating hyper-exponential processes in the P-K statistics

Real processes can involve hyper-exponential headway distributions. For example, arrivals might be drawn from several different streams with different headways, and service might be affected by intermittent blocking of several exit routes with different capacities. This research is aimed specifically at exploiting and enhancing the P-K formula as it is used in many traffic modelling tools, so hyper-exponential technically falls outside the scope. However, it is possible to estimate how parameters of the P-K formula could represent the *average* effect of a hyper-exponential headway distribution. Figure 2.3.4 shows how the average coefficient of variation (c.v.) of headways, i.e. c_a of arrivals, or c_b in the P-K randomness coefficient, is affected by varying the spread of three rate parameters $\{\lambda_m\}$ expressed in terms of their c.v. from 0 to 0.75, with mean headway normalised to 1, and various permutations of mixture probabilities $\{h_m\}$, again with c.v. in the range 0 to 0.75. In addition to the average value, the distribution of results assuming each permutation to be equally probable is indicated. All process headway c.vs are greater than 1 and less than 3, so effective randomness is always increased but might potentially be accommodated approximately in the P-K model.

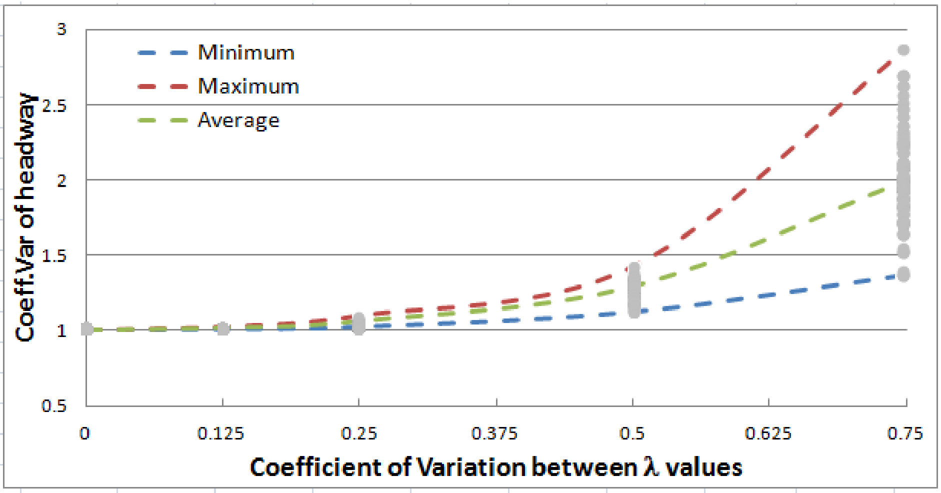


Figure 2.3.4 Headway coefficient of variation for three-way hyper-exponential cases

2.3.14 Limited applicability of equilibrium

Figure 2.3.5 shows that the practical range of traffic intensities where equilibrium conditions can be considered to apply is quite limited. This graphs the M/M/1 equilibrium queue (2.3.31) and the relaxation time (2.3.37), assuming throughput capacity of 3600 units/hour, equivalent to two average lanes. A lower throughput would increase the relaxation time in proportion, making the region of interest even smaller, although in some networks reassignment might tend to maintain traffic levels in this region. ‘Heavy traffic’ generally refers to 0.95≤ρ< 1.

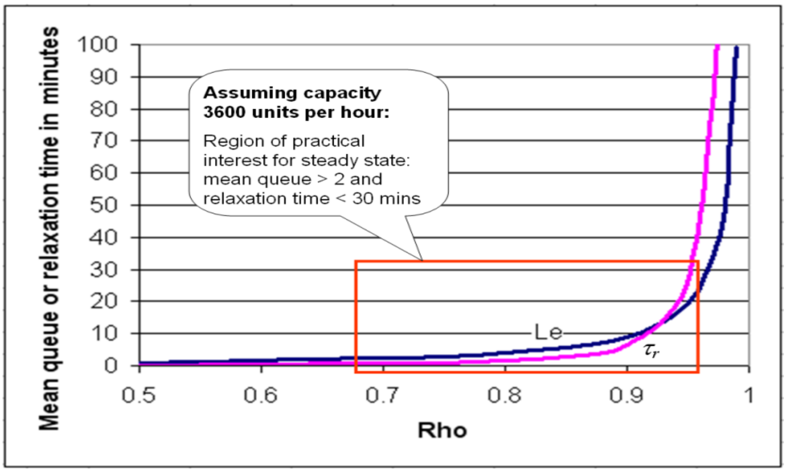


Figure 2.3.5 M/M/1 steady state queue L_e and relaxation time τ_{re} v. demand intensity

In practical problems such as modelling traffic peaks with transient oversaturation, the traffic levels are often either too low to produce interesting results, or too high for the system to get anywhere near equilibrium in the time available. The latter time-dependent case is of the greatest interest because predictions for any particular time can depend sensitively on past development, and calculations become correspondingly more complicated.

2.4. M/D/1 PROCESSES DERIVED USING RECURRENCE RELATIONS

2.4.1 Motivation and approach

M/D/1 with ‘Markovian’ (random) arrivals and ‘Deterministic’ (uniform) service idealises the queue remaining stochastically at the end of an intermittent short period of throughput like green at a signal, where traffic can be either stopped (red) or flowing freely (green), sometimes referred to as the ‘overflow’ queue. The service interval is notional in that it is based on average capacity and does not explicitly reflect a signal cycle. A separate phase queue formula accounts for the linear build up of queue in the red phase and its discharge in the green phase (e.g. Webster and Cobbe 1966). The derivation of M/D/1 differs from M/M/1 in that instead of a single arrival in an infinitesimal time interval, multiple arrivals can occur during the finite service interval. Derivations of the basic M/D/1 moments including variance formulae are given in Appendix A. However, it allows for only one customer to be serviced in its notional service period. Furthermore, like M/M/1, the M/D/1 equilibrium queue depends only on the traffic intensity, which in turn depends only on the *ratio* of green to cycle time.

An extended formulation is therefore required to represent the actual capacity of signal green periods. This has been recognised since signal queues began to be analysed, which can be traced to A J H Clayton in 1940 (see Allsop / Hutchinson 1972). Various formulae have been developed that take account of the green period capacity (Appendices D, E), but the importance of M/D/1 is that it is a special case of the Pollaczek-Khinchin equilibrium queue formula, and so can be incorporated in time-dependent methods like shearing (see later in Chapter 4). It is therefore advantageous to have an extension of M/D/1 that can account for green period capacity. In practice, unlike the basic M/D/1, exact closed formulae for the extended moments do not appear to be available, so empirical approximations to them are required, which will be discussed later in Chapter 3, along with other derivations of equilibrium formulae.

2.4.2 The basic M/D/1 process

The basic M/D/1 process can be interpreted as one in which customers are served singly in fixed service time intervals, during which more than one random arrival can take place. The probability of *j* arrivals in a unit time period is Poisson distributed:

$$\pi_j = \frac{\rho^j e^{-\rho}}{j!} \qquad (2.4.1)$$

The probability of i units in the queue at the point after μt+1 service periods (cumulative throughput capacity) is the sum of the probabilities of i+1 in the queue at μt and no arrivals in [μt, μt +1), i at μt and one arrival, i-1 at μt with 2 arrivals, and so on, hence:

$$p_i(\mu t+1)=\sum_{j=0}^{i+1}\pi_j\, p_{i+1-j}(\mu t)=\sum_{j=0}^{i+1}\frac{\rho^j e^{-\rho}}{j!}\, p_{i+1-j}(\mu t) \quad \text{and} \qquad (2.4.2)$$

$$\Delta p_i(t)=p_i(\mu t+1)-p_i(\mu t) \qquad \text{in finite differential form} \qquad (2.4.3)$$

In the steady state $\Delta p_i(t)=0$ for all i, so rearranging (2.4.2), and in calculating p_0 allowing that the initial queue may be zero and there are no arrivals, i.e. the service is not utilised:

$$p_1=\left(e^{\rho}-\rho-1\right)p_0 \qquad (2.4.4)$$

$$p_i=\left(e^{\rho}-\rho\right)p_{i-1}-\sum_{j=2}^{i}\frac{\rho^j}{j!}\, p_{i-j} \qquad (i> 1) \qquad (2.4.5)$$

The derivation of (2.4.4) for the final state i=0 in (2.4.2) can be visualised as follows:

Table 2.4.1 Construction of recurrence formula (2.4.4) for p_1

Initial state	**Arrivals**	**Departures**	**Final state**	***j* in (2.4.2)**	**Factor**	p_{i+1-j}
1	0	1	0	0	$e^{-\rho}$	p_1
0	1	1	0	1	$\rho e^{-\rho}$	p_0
0	0	0	0	0	$e^{-\rho}$	p_0

The first two terms contribute to a *notional* $p_{(0)}$ as calculated from (2.4.2), which is incomplete as the final zero state, and the third term is a notional $p_{(-1)}$ term, again as calculated from (2.4.2). These two notional probabilities must be added to give the probability of the *real* zero state, so using brackets around the indices to indicate notional states:

$$p_0=p_{(0)}+p_{(-1)}=e^{-\rho}\left(p_1+(\rho+1)p_0\right) \qquad (2.4.6)$$

Multiplying this through by e^{ρ} and rearranging yields (2.4.4). The significance of this procedure is twofold: first, it formalises the working of the absorbing boundary at zero; second, it is a prototype for more general use of notional non-positive states later. Equation (2.4.5) is derived in a similar way but without needing a notional term: i.e. setting i=1 yields p_2 etc.

As in M/M/1, moments of the steady-state distribution can now be got from (2.4.4-5) (see Appendix A for derivation):

by evaluating $\sum_i i p_i$:
$$p_{0e} = e^{\rho}(1-\rho) \qquad (2.4.7)$$

by evaluating $\sum_i i^2 p_i$:
$$L_e = \frac{\rho^2}{2(1-\rho)} \quad \text{(M/D/1 without u.i.s.)} \qquad (2.4.8)$$

by evaluating $\sum_i i^3 p_i$:
$$V_e = \frac{\rho^2(6-2\rho-\rho^2)}{12(1-\rho)^2} \qquad (2.4.9)$$

Equation (2.4.8) is the idealised form of random queue at a signalised stop line (I^*=0, C=0.5 in equation (2.3.50)), although Kimber and Hollis (1979) and references therein, and Burrow (1987), find that setting C=0.55 gives a better match to observation. It represents a queue without 'unit in service' because departures from the queue can proceed at saturation flow without having to wait for a gap. While queue size probabilities sum to unity as expected, equation (2.4.7) is inconsistent with equilibrium utilisation, which as long as L remains finite must equal ρ to satisfy (2.3.4), as explained in the previous Section. So *average* p_0 should equal (1-ρ). In fact, p_0 in (2.4.7) applies at the *end* of the service period, whereas average p_0 takes into account queues that exist transiently during the green period but discharge before the end of it. The final value is greater than average p_0 by a factor $e^{\rho} \geq 1$, consistent with this interpretation. This important distinction will be justified later.

2.4.3 M/D/1[G] queue dependent on green period capacity

Olszewski (1990) reports an observation made by Miller (1969) and previously by Newell, subsequently picked up by Cronjé (1983a) who quotes the relevant formulae, that the mean size of the stochastic ('overflow') queue at a signal tends to fall with increasing throughput capacity in the green period G. He develops transition probabilities for a signal-like process with general arrival distribution and variable service period, and using Markov simulation shows that the mean queue size does indeed decrease in a manner very similar to Newell's model[21]. Miller's and Newell's formulae include dispersion of arrivals I_a as a multiplying factor, the logic presumably being that removing all randomness should eliminate the stochastic queue.

According to equation (2.3.50), where I_a enters differently, the queue is zero when I_a=(1-ρ), so apparently it is not necessary to eliminate arrivals dispersion to make the queue zero, but the queue can still be reduced by any required factor. At a real signal the randomness of stop-line arrivals during green could be less than that of a distant demand source because some short-term randomness in arrivals is absorbed during red.

[21] See Figure 3 in Olszewski (1990). G is used here in place of Olszewski's B.

However, if throughput capacity in the green should then increase, the uniform discharge flow from any red queue should represent a *smaller* proportion of the throughput, so *increasing* the dispersion of stop-line arrivals. This would tend to counteract any direct effect of I_a on the mean queue size, so the mean queue reduction cannot be linked so directly to a change in the dispersion of arrivals. Since the M/D/1 model assumes c_b=0, changing the service statistics also cannot reduce the mean queue size. Therefore the model needs radical modification beyond that achievable within the P-K formula.

2.4.4 Allowing for different green period capacities

The basic M/D/1 model describes a system like a ramp meter where only one customer can be served in each green period. A more realistic model allows for G customers to be serviced in each green period, and the recurrence relations become more complicated. Queues of a specified size at the end of the green period can be achieved by the following conditions:

Table 2.4.2 Conditions for getting given final queue in green period of capacity *G*

Initial state	**Number of arrivals in green period**
Queue i = 0 at end of green period	
0	Up to G
1	Up to G-1
j	Up to G-j
>G	*Not possible*
Queue i> 0 at end of green period	
0	Exactly G+i
1	Exactly G+i-1
j	Exactly G+i-j
>G+i	*Not possible*

Table 2.4.2 shows that development of p_i with i>0 is similar to equation (2.4.2), while that for p_0 must be more complicated, and further that notional states down to -G must be considered:

$$p_i(\mu t + G) = \sum_{j=0}^{i+G} \frac{(G\rho)^j e^{-(G\rho)}}{j!} p_{i+G-j}(\mu t) \qquad (i \geq \text{-G}) \qquad (2.4.10)$$

As a device for calculating an overflow queue probability distribution, Cronjé (1983a) (in his Table 2) introduces notional negative overflow queue sizes, summing these terms to give the real p_0 at the end of the cycle, although he does not appear to pursue the concept further. Tables 2.4.3 visualise this approach for G = 1, 2, and 5 respectively, where the notional states are represented by the rows with bracketed indices in the left column.

Table 2.4.3 Charts relating final to initial state for different green capacities

G=1	Initial state												
↓←	0	1	2	3	4	5	6	7	8	9	10	11	12
(-1)	0												
(0)	1	0		Numbers in interior cells are numbers of arrivals in period									
1	2	1	0		Mean arrival rate = ρ								
2	3	2	1	0		Departures in period = 1 exactly							
3	4	3	2	1	0								
4	5	4	3	2	1	0							
5	6	5	4	3	2	1	0						

G=2	Initial state												
↓←	0	1	2	3	4	5	6	7	8	9	10	11	12
(-2)	0												
(-1)	1	0		Numbers in interior cells are numbers of arrivals in period									
(0)	2	1	0		Mean arrival rate = *G*ρ = 2ρ								
1	3	2	1	0		Departures in period = 1 exactly							
2	4	3	2	1	0								
3	5	4	3	2	1	0							
4	6	5	4	3	2	1	0						
5	7	6	5	4	3	2	1	0					

G=5	Initial state												
↓←	0	1	2	3	4	5	6	7	8	9	10	11	12
(-5)	0												
(-4)	1	0		Numbers in interior cells = numbers of arrivals in period									
(-3)	2	1	0		Mean arrival rate = *G*ρ= 5ρ								
(-2)	3	2	1	0		Departures in period = *G* exactly							
(-1)	4	3	2	1	0								
(0)	5	4	3	2	1	0							
1	6	5	4	3	2	1	0						
2	7	6	5	4	3	2	1	0					
3	8	7	6	5	4	3	2	1	0				
4	9	8	7	6	5	4	3	2	1	0			
5	10	9	8	7	6	5	4	3	2	1	0		

Numbers in interior cells are arrivals, which translate into Poisson coefficients, in terms of $G\rho$ rather than ρ, multiplying the initial probabilities (columns) as in equation (2.4.10). To convert from these sequential relationships to steady-state recurrence relations, the final and initial state probabilities are equated and terms rearranged as follows:

(1) move the final state probability from the left-hand column to the right and combine it with the corresponding initial state term (black-bordered cells)

(2) move the highest state (always initial) to the equation LHS (red bordered cells)

(3) divide through by coefficient of the new LHS (always $e^{G\rho}$)

(4) adjust indices as necessary.

This procedure is straightforward for G=1, yielding (2.4.4-5) from (2.4.2), but gives a more complex result for G>1. However, the coefficient of the final-state probability is always the G^{th} Poisson coefficient (content of black bordered cells), and the coefficient of the highest state is always the 0^{th} Poisson coefficient (content of red bordered cells). At equilibrium steps (1-2) produce equation (2.4.11) and steps (3-4) rearrange this to (2.4.12):

$$e^{-(G\rho)} p_{i+G}^{(e)} = p_i^{(e)} - \sum_{j=1}^{i+G} \frac{(G\rho)^j e^{-(G\rho)}}{j!} p_{i+G-j}^{(e)} \qquad (i>0) \qquad (2.4.11)$$

$$p_i^{(e)} = e^{(G\rho)} p_{i-G}^{(e)} - \sum_{j=1}^{i} \frac{(G\rho)^j}{j!} p_{i-j}^{(e)} \qquad (i>G) \qquad (2.4.12)$$

To get the recurrence relations for $i \leq G$ consider the upper triangular parts of Tables 2.4.3. In the case of G=1, the components of the recurrence relation can be constructed as in Table 2.4.4 (where π_j represents the j^{th} Poisson coefficient):

Table 2.4.4 Construction of recurrence relation for p_1 in the case G=1

Target equilibrium state	**Components of recurrence relation**	
$\pi_0 p_1 =$	$p_{(0)}$	$-\pi_1 p_0$
$\pi_0 p_0 =$	$p_{(-1)}$	

The *real* zeroth state at the end of green is given by (2.4.13), as earlier in (2.4.6):

$$p_0 = p_{(-1)} + p_{(0)} \qquad (2.4.13)$$

Adding the terms, moving the (unknown) p_0 target term to the RHS, applying (2.4.11) to leave three terms, and dividing through by π_0, results in equation (2.4.4). In this case the notional terms can be found explicitly:

$$
\begin{aligned}
p_{(-1)}^{(e)} &= (1-\rho) \\
p_{(0)}^{(e)} &= \left(e^{\rho}-1\right)(1-\rho)
\end{aligned}
\qquad (2.4.14)
$$

In the case G=5, the structure is more complicated, as shown by Table 2.4.5. From this, an expression for p_5 can be got in terms of lower states. However there seems no straightforward way to determine the probabilities of states below 5. This might be possible if the $p_{(-i)}$ could be evaluated independently, but there seems to be no way to do this.

Table 2.4.5 Construction of recurrence relations for states in the case G=5

Target	**Components of recurrence relation**					
$\pi_0 p_5$ =	$p_{(0)}$	$-\pi_1 p_4$	$-\pi_2 p_3$	$-\pi_3 p_2$	$-\pi_4 p_1$	$-\pi_5 p_0$
$\pi_0 p_4$ =	$p_{(-1)}$	$-\pi_1 p_3$	$-\pi_2 p_2$	$-\pi_3 p_1$	$-\pi_4 p_0$	
$\pi_0 p_3$ =	$p_{(-2)}$	$-\pi_1 p_2$	$-\pi_2 p_1$	$-\pi_3 p_0$		
$\pi_0 p_2$ =	$p_{(-3)}$	$-\pi_1 p_1$	$-\pi_2 p_0$			
$\pi_0 p_1$ =	$p_{(-4)}$	$-\pi_1 p_0$				
$\pi_0 p_0$ =	$p_{(-5)}$					

Seeking a numerical method, a Markovian sequence can be set up that calculates iteratively the raw probabilities, including notional final state probabilities (left hand column of Tables 2.4.3) from the previous ‘real’ state probabilities, then calculates the final real probabilities from the raw probabilities. A Markov simulation program developed for benchmarking, Qsim[22], has been enhanced to calculate recurrence relations (2.4.10) for in principle any value of G, in practice up to G=100 and ρ up to 0.95. It is difficult to guarantee accuracy at higher values of G because the recurrence relations become increasingly complex, and convergence becomes increasingly hard to achieve as ρ approaches 1. Some compensation for this inaccuracy is possible by ensuring that the sum of probabilities is always normalised to 1.

In equations (2.4.13-14), $p_{(-1)}$ represents the probability that the queue remains zero throughout the entire green service period, while $p_{(0)}$ represents the probability that one arrival occurs and is served in the period. This logic can be extended to any value of G.

[22] The implementation of Qsim is discussed in the Section on benchmarking.

Since all the notional probabilities contribute zero queue at the *end* of green, in general:

$$p_0 = \sum_{i=0}^{G} p_{(-i)} \tag{2.4.15}$$

Because it represents a queue present for a negligible part of the green period, $p_{(0)}$ does not contribute to the *average* probability of the queue being zero *during* the service period, that determines the utilisation in practice. The notional probabilities contribute varying amounts because some arrivals during the green period can have 'disappeared' by the end.

Imagining Table 2.4.3 extended to any values of *G*, evaluating the steady-state mean value of the 'triangles' by initially treating the $\{p_{(-i)}\}$ as real probabilities of negative queue states, and since all blank cells are identically zero, the k^{th} column mean is given by:

$$K_k = \left[\sum_{i=0}^{\infty}(i+k-G)\frac{(G\rho)^i e^{-G\rho}}{i!}\right]p_k = [G\rho + k - G]p_k \tag{2.4.16}$$

The left-hand column total must equal the total of the right-hand columns, namely:

$$\sum_{0}^{\infty} K_k = L_e - G(1-\rho) \qquad \text{hence} \tag{2.4.17}$$

$$\sum_{1}^{\infty} ip_i - \sum_{0}^{G} ip_{(-i)} = L_e - \sum_{0}^{G} ip_{(-i)} = L_e - G(1-\rho) \qquad \text{so finally}$$

$$L^*_{[G]} = \sum_{0}^{G} ip_{(-i)} = G(1-\rho) \qquad \text{or} \qquad \frac{1}{G}\sum_{0}^{G} ip_{(-i)} = (1-\rho) \tag{2.4.18}$$

Also: $$p_{(-G)} = e^{-G\rho} p_0 \tag{2.4.19}$$

Thus the *average* probability of the queue being zero during the service period is indeed consistent with the definition of utilisation, that requires it to equal ρ at equilibrium as discussed earlier in Section 2.3. It is seen now that the possibility to get result (2.4.7) in the case *G*=1 depends on the fact that there are two equations, (2.4.13) and (2.4.18), that can be solved for two unknowns. This apparently does not extend to *G*>1.

From Table 2.4.3, the notional probabilities for G=2 can be expressed in terms of p_0:

$$\begin{aligned} p_{(-2)} &= e^{-2\rho} p_0 \\ p_{(-1)} &= 2(1-\rho) - 2e^{-2\rho} p_0 \\ p_{(0)} &= \left(1+e^{-2\rho}\right) p_0 - 2(1-\rho) \end{aligned} \qquad (2.4.20)$$

Subsequently (2.4.12) leads to relations analogous to (2.4.4-5):

$$p_1 = e^{2\rho} p_{(-1)} - (2\rho) p_0 \qquad (2.4.21)$$

$$p_2 = e^{2\rho} p_{(0)} - (2\rho) p_1 - \frac{(2\rho)^2}{2!} p_0 \qquad (2.4.22)$$

$$p_i = e^{2\rho} p_{i-2} - (2\rho) p_{i-1} - \sum_{j=2}^{i} \frac{(2\rho)^j}{j!} p_{i-j} \quad (i > 2) \qquad (2.4.23)$$

Summing either $\{p_i\}$ or $\{ip_i\}$ using these equations yields an identity, and summing $\{i^2 p_i\}$ gives a formula for the steady-state mean queue confirmed for G=2 by Markov simulation (up to i=14), that depends on p_0:

$$L_e = \frac{2\rho^2 - 2\rho + 1 - e^{-2\rho} p_0}{2(1-\rho)} \qquad (G\text{=2 only}) \qquad (2.4.24)$$

The formulae so far do not give an explicit value for p_0. If a formula corresponding to (2.4.18) could be found for the variance of the notional probabilities, then this might yield p_0. The results show at least that the variance of the notional terms satisfies:

$$\operatorname{var}\left\{p_{(-i)}\right\}_{[G]} \to G\rho \qquad \text{in the limit when } p_0 \to 1 \qquad (2.4.25)$$

This is certainly true when ρ→0, and also when G is large. However, this gives no clue as to the dependence on p_0 in other cases.

From (2.4.15) and (2.4.20) in teh cases G=1 and G=2 respectively:

$$\operatorname{var}\left\{p_{(-i)}\right\}_{[G=1]} = \rho(1-\rho) \qquad (2.4.26)$$

$$\operatorname{var}\left\{p_{(-i)}\right\}_{[G=2]} = 2(2\rho-1)(1-\rho) + 2e^{-2\rho} p_0 \qquad (2.4.27)$$

The presence of p_0 in (2.4.20) and (2.4.27) when G=2, and its absence in (2.4.14) and (2.4.26) when G=1, suggests there may be an inherent difficulty in extrapolating these equations, for example those for G=3 may involve p_1 and so on. This would not invalidate (2.4.25) since in the equilibrium distributions p_i<p_0 for all i>0.

Although some useful results have been found for M/D/1[G], this Section has failed to produce any definitive results for its important moments. Historically, researchers have resorted to empirical approximations. These and some new empirical formulae are described later in Chapter 3.

2.5. BENCHMARKING METHODS

2.5.1 Peak cases used for benchmarking

A set of benchmark peak cases is available from Kimber *et al* (1986)[23] who approximated the M/M/1 variance profiles for 34 symmetrical Gaussian peaks by gluing together half-Gaussian functions. The cases represent queuing at several hypothetical give-way junctions 'J*m*' with peak profiles 'P*n*' producing transient overloading up to various amplitudes, and durations ranging from 45 to 120 minutes. For the purpose of simulation, the profiles have been divided into at least 12 equal time slices, in each of which demand and capacity are assumed to be constant. A similar time-sliced approach is used in respectively roundabout, T- and signal junction models ARCADY (Semmens 1985a, Binning 1996), PICADY (Semmens 1985b), OSCADY (Burrow 1987), and dynamic traffic assignment model CONTRAM (Taylor 1990,2003).

One of the test cases, J2P4, provides a useful test of 'middle-weight' symmetrical Gaussian peak with substantial overload, ρ_{max}=1.1384, over a substantial but not excessive peak period lasting 90 minutes. Because this case was originally formulated to represent a give-way junction with opposing flow, capacity falls somewhat as demand increases, as in Table 2.5.1.

Table 2.5.1 Definition and properties of symmetrical peak case J2P4 as used here

Time Slice	**End time (min)**	**ρ**	**μ (veh/min)**	***p*₀**	***L***	***V***
0	0[24]	0.5717	15.906	0.4283	1.3348	3.1165
1	9	0.6472	15.492	0.3528	1.8343	5.1975
2	18	0.8032	14.694	0.1986	3.9755	19.0354
3	27	0.9520	14.004	0.0736	9.9567	76.6503
4	36	1.0711	13.494	0.0170	22.2552	213.1961
5	45	1.1384	13.224	0.0034	39.5938	416.7314
6	54	1.1384	13.224	0.0012	56.3156	647.6774
7	63	1.0711	13.494	0.0015	65.1355	876.5028
8	72	0.9520	14.004	0.0065	59.6147	1056.5750
9	81	0.8032	14.694	0.0498	37.2192	963.7484
10	90	0.6472	15.492	0.2465	10.4176	314.5935
11	99	0.5719	15.906	0.4148	1.9363	22.0238
12	108	0.5719	15.906	0.4276	1.3519	3.5187

[23] This was joint work between the Transport Research Laboratory and Halcrow Fox and Associates, who also investigated (unpublished) the use of Gamma Distributions to model queue size probability distributions.

[24] This represents an equilibrium initial state.

2.5.2 Accuracy of M/M/1 Series calculations

Issues relating to the choice of maximum queue size N and the accuracy of the Series calculation have been discussed earlier. The author first implemented the Series method in a Fortran program. A Visual Basic pseudocode listing is given in Appendix B. Results of the Series calculations are consistent for maximum resolutions (N+1) of 256 and above, but begin to show errors at maximum resolution 128, that become significant at maximum resolution 64. The errors become evident at ρ>1, where it is not possible to analyse equilibrium distributions. In an attempt to determine the required resolution, which amounts to estimating where probabilities become ‘insignificant’, probability distributions for each of the 12 time slices of J2P4 have been compressed in Figures 2.5.1 using natural logarithms.

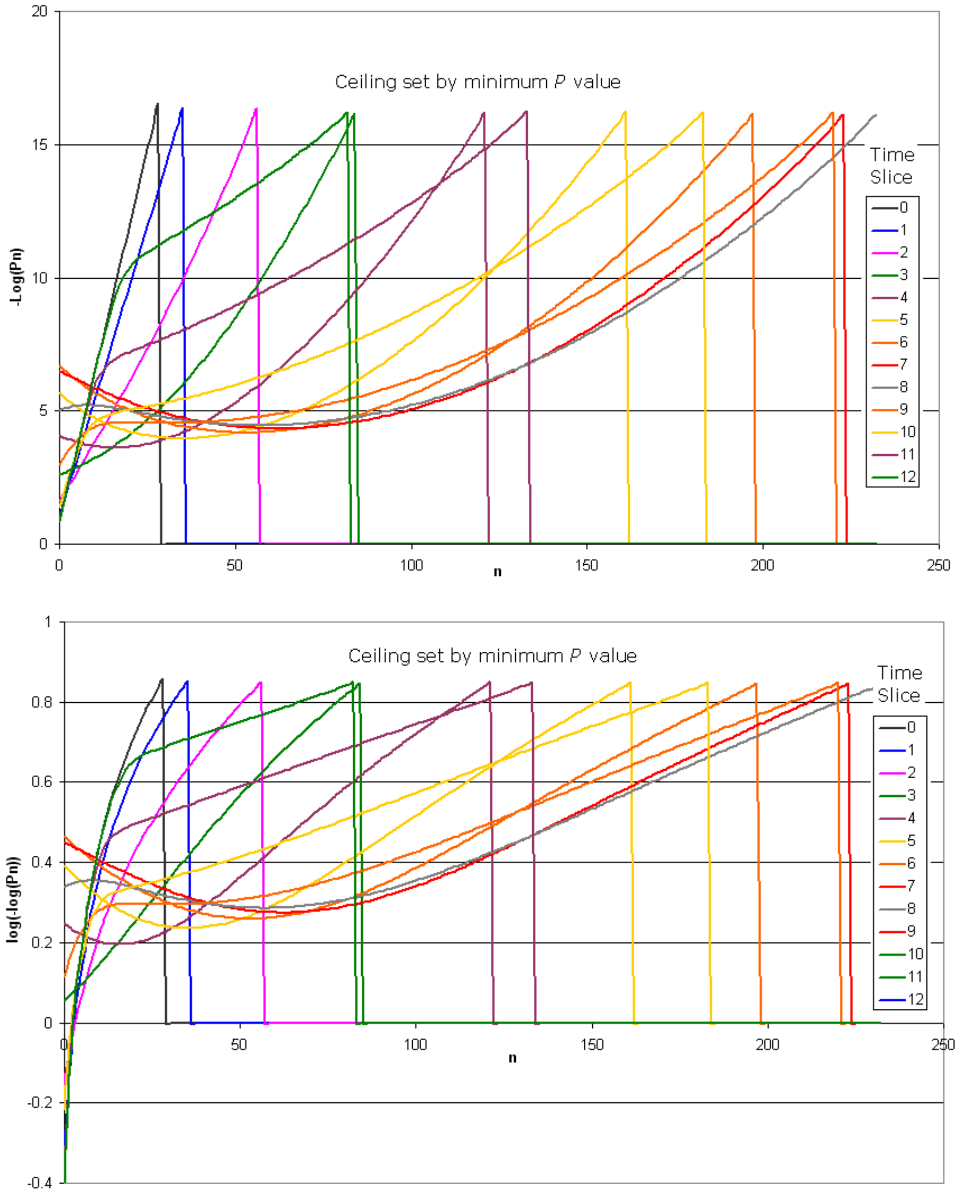


Figure 2.5.1 -ln(p_i)and –ln(ln(p_i))(queue size probability) distributions for peak case J2P4

The maximum values plotted are determined by the arbitrary lower limit on p_i. If the tail of the distribution should approach a geometric equilibrium form, ln(p_i) should become linear with i. However, in Figure 2.5.1 (left), where the actual -ln(p_i) are plotted against i for the initial equilibrium distribution and, this appears not to be the case except in the first two time slices. In practice, in most time slices, ln(-ln(p_i)) appears to become almost linear against i as in Figure 2.5.1 (right). Why this should be the case is not clear, but it implies that most of the probability distributions, including those post-peak, are nowhere near equilibrium, or nearer Gumbel than exponential. This is not pursued as it has been possible to generate distributions at sufficient resolution, and the results compare well with more efficient Markov simulations next.

2.5.3 Markov simulation

Markov simulation calculates the probability of each queue state at each point in time by calculating all possible transitions from the possible states at the previous time point according to transition probabilities. It can therefore be applied where only differential recurrence relations are available, without explicit state formulae. The initial distribution at each cycle can most readily be an exact state, e.g. zero, or an initial equilibrium state, or a distribution carried over from the previous time point. As time t is advanced in small steps δt, the $\{p_n(t)\}$ are calculated at successive points in time, using finite differential equations of the general linear form:

$$\frac{\delta p_n}{\delta t} = a_{nm} p_m \qquad (2.5.1)$$

Specific recurrence relations exist for M/M/1 (2.3.22-23) and M/D/1 (2.4.2), (2.4.10). In addition to its greater computational efficiency, Markov simulation's ability to calculate M/D/1, M/D/1[G] and other distributions lacking explicit formulae for state probabilities (see also Chapter 3), gives it a significant advantage over Series calculation which is limited to M/M/1. With **Neil H Spencer**, then a sandwich student at TRL, the author developed a Markov algorithm implemented in a program 'Qsim'. This uses the above method to develop a queue length distribution stepping through consecutive time-slices in which the arrival and service rates are assumed constant. The program has since been extended to include M/D/1[G], and queues with Erlang-k arrival or service distributions (see Chapter 3 later). The effect of continuously varying demand and capacity can in principle be simulated by using very short time-slices (e.g. 1 second). Alternatively, mean arrivals and capacity can be held constant over periods of several minutes to allow comparison with time-sliced traffic models such as ARCADY (Semmens 1985a, Binning 1996), PICADY (Semmens 1985b), OSCADY (Burrow 1987) and CONTRAM (Taylor 1990, 2003). A Basic pseudocode listing and example of output are given in Appendix C.

2.5.4 Sensitivity of Markov method to maximum queue and step size

The step size δt needs to be chosen with care, since if it is too large the calculation risks becoming unstable, whereas if it is too small computation time is increased without any useful gain in accuracy. As with the Series method, practical calculations require a maximum queue length to be set. The Series method is assumed to be exact provided that the maximum queue size (array dimension) in the calculation is sufficient. For all the J*m*P*n* test cases, 1024 is considered more than adequate, being over five times the largest mean queue expected. For the J2P4 peak case, maximum queue size of 256 is sufficient, and reducing it to 128 has little effect. However, the calculation fails when maximum queue size is reduced to 64, comparable with the maximum *mean* queue size. Computation time can be reduced, without significant loss of accuracy, by not suppressing calculating once $p_i(t)$ falls below 10^{-7}, but this needs to be smaller than the cut-off value for the Series probabilities to avoid accumulating error through many calculation steps. If the calculation does reach the maximum queue size N then the results *could* be inaccurate. The last calculated p_m could be a guide to the degree of inaccuracy if the question of asymptotic behaviour raised in section 2.5.3 were resolved, allowing the missing ‘tail’ to be estimated.

Step size needs to be smaller for the M/M/1 model, which is based on differential recurrence relations, than for the M/D/1 model, which is based on finite recurrence relations. In both cases, the calculation fails sharply if the step size is too large. Time steps of 1 second give consistent results, though 0.1 second can be used for added security or as a check.

Results of using different maximum queue and step sizes are shown in the four panels of Figure 2.5.2. Run time appears to be little affected by step size but appears to be roughly proportional to the logarithm of maximum queue size: e.g. if it is x1 for N=64, it is x2 for 128 and x3 for 256 etc.

2.5.5 Microscopic simulation

By simulating directly the random arrival and service at units, at given mean rates, probability distributions can be built up empirically over a large number of trials. This leads one to consider the practical interpretation of the probability distribution, usually seen somewhat artificially as representing the state at a particular time when the same conditions are repeated day after day. Being conceptually the simplest method and so the least prone to coding error, random simulation is useful for checking the other methods and for verifying general features of the probability distribution, but it is also the most time consuming, and can produce only a ragged approximation to the underlying distribution.

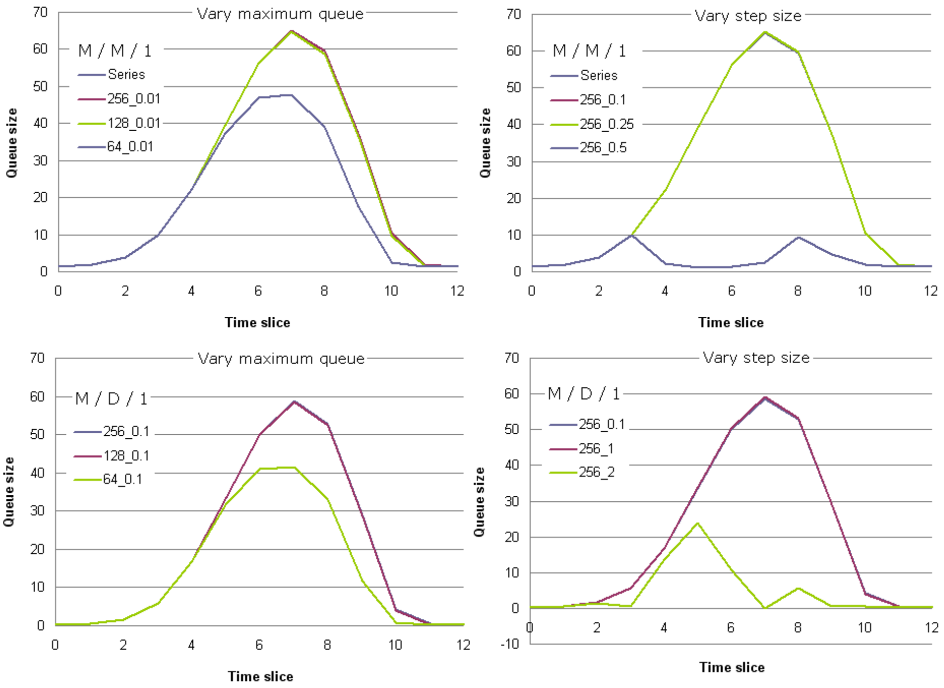


Figure 2.5.2 Showing abrupt failure cases of otherwise consistent Markov models. The lower graph in each case relates to smallest *N* or largest step size. Other graphs cannot be separated

2.5.6 Comparison of actual results of different calculation methods

Figure 2.5.3 shows the probability distribution in the J2P4 case at the end of the time slice (Ts) giving the greatest variance: Ts 8 at 72 minutes. Interesting features are the accuracy of the Markov simulation even when using very large time slices, and the bimodal shape with a dip at *i*=8, that arises in all cases. The lower left figure was generated using 1 second time *slices*, i.e. each of the original 9 minute time slices was broken down into 540 consecutive time slices with the same values of ρ and μ.

Figures 2.5.4-6 compare probability distributions and profiles for the case J2P4 calculated by the M/M/1 Series, M/M/1 Markov and M/D/1 Markov methods. In Figure 2.5.6 the peak profile ρ is shown together with the mean *L* and variance *V* of the queue at the end of each time slice, and the delay *D* (time-average of the mean queue) over each time slice is also plotted. In each case both the variance obtained directly from the probability distribution, *V* and that calculated using *D* and the variance formula, V_{cal}, are shown, but they are so close that broken lines have been used to prove that both are shown. This confirms not only that the variance formula (2.3.27) is correct but that it applies to M/D/1. Since its form does not appear specific to a distribution, it is inferred that like (2.3.4) it should apply to any queue process, and while this is not absolutely proved further evidence will be provided later. Note the lag of variance relative to the mean, as well as the lag of mean relative to the demand peak, as remarked earlier.

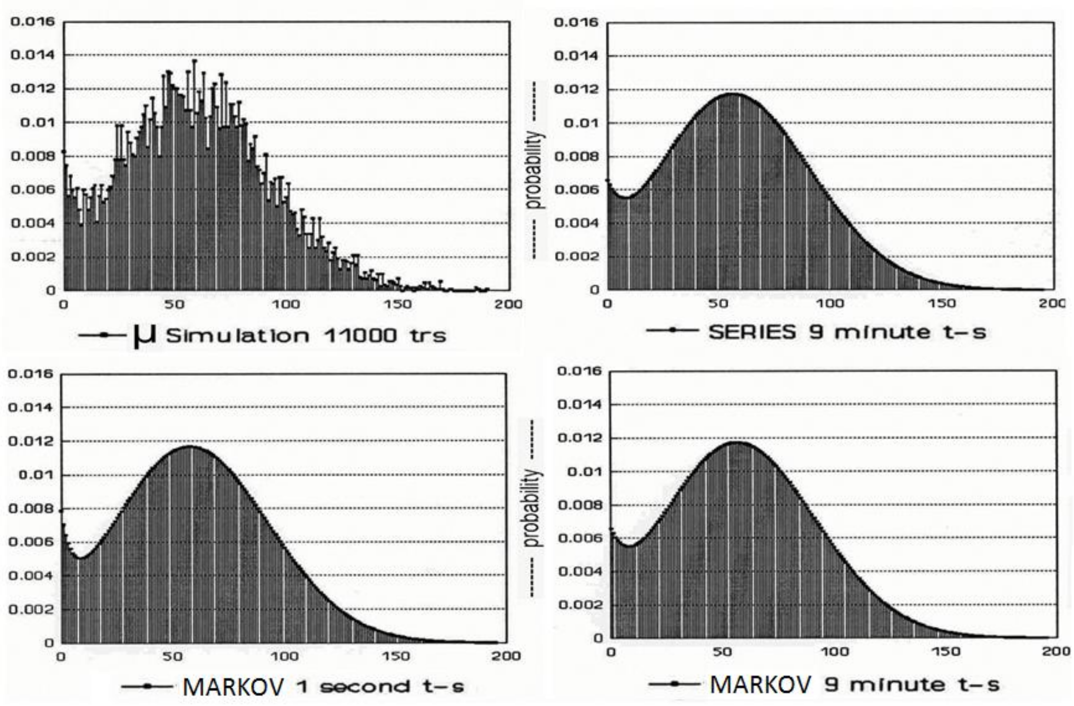


Figure 2.5.3 Alternative probability distributions for J2P4 M/M/1 at maximum variance

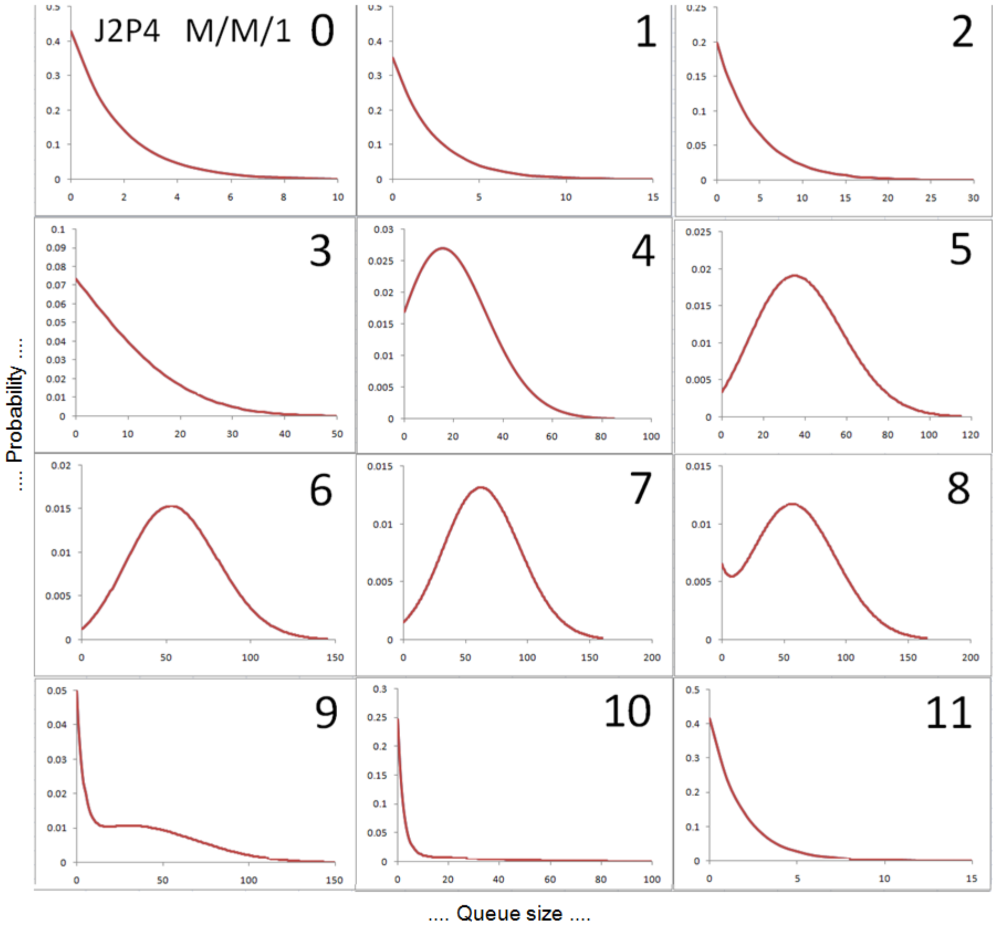


Figure 2.5.4 Series and Markov M/M/1 probability distributions compared for J2P4 for initial state and in 9-minute t/s through the peak (scales vary). The graphs cannot be separated.

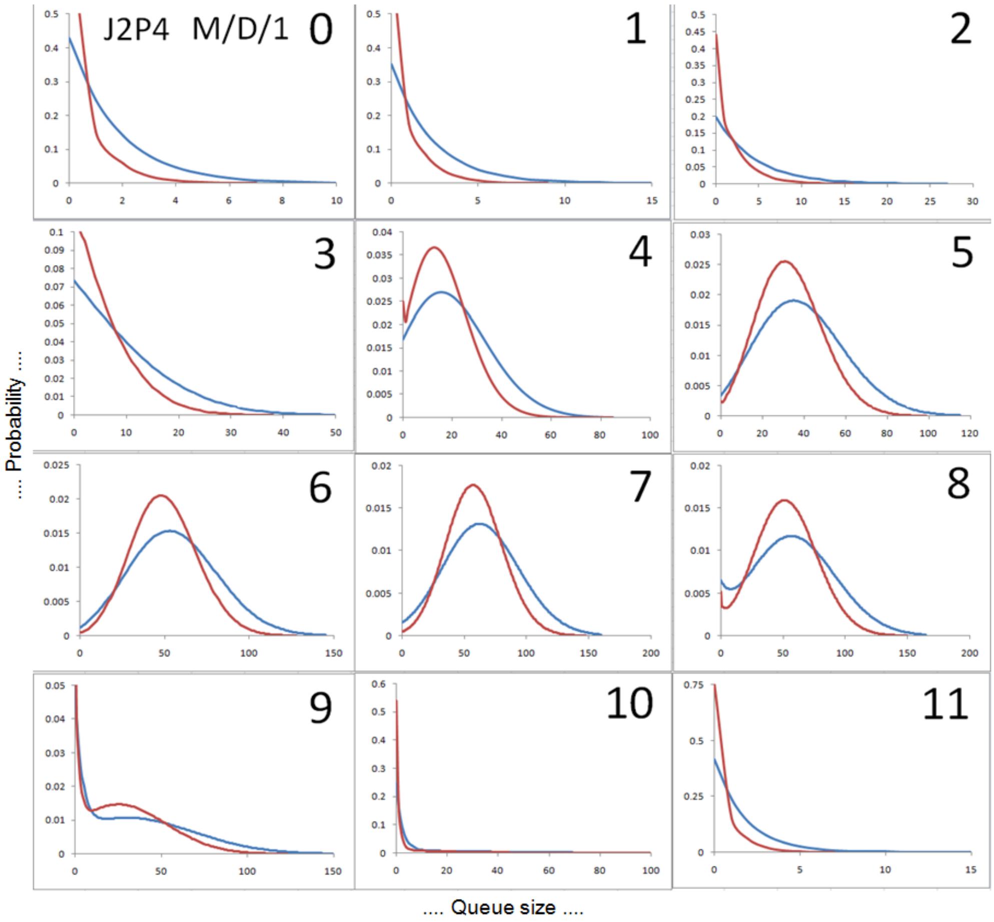


Figure 2.5.5 Markov M/D/1 probability distributions for J2P4 for initial state and in 9-minute t/s through the peak – compared with flatter M/M/1 Series graphs (scales vary)

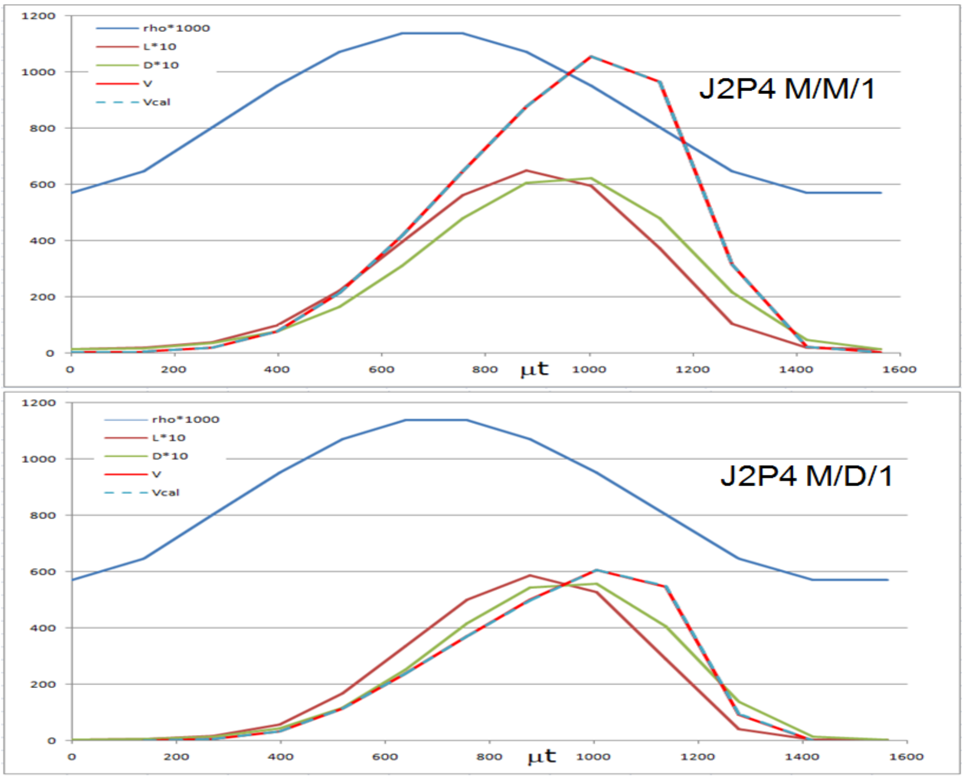


Figure 2.5.6 Simulated profiles for J2P4 peak case, showing ρ, *L*, *D*, *V* (not to same scale)

(Upper graphs are profile of demand ρ. *V* is shown by skewed graphs with alternating colours)

2.5.7 Deviation from equilibrium in queue development over time

The Markov simulation program can calculate development of a queue size probability distribution over an extended time period leading to equilibrium. Figure 2.5.7 shows a sequence of distributions for ρ=0.9, μ=1, that get increasingly spread out but tend towards a similar 'geometric' shape. However, at most points in time the shape cannot be precisely geometric.

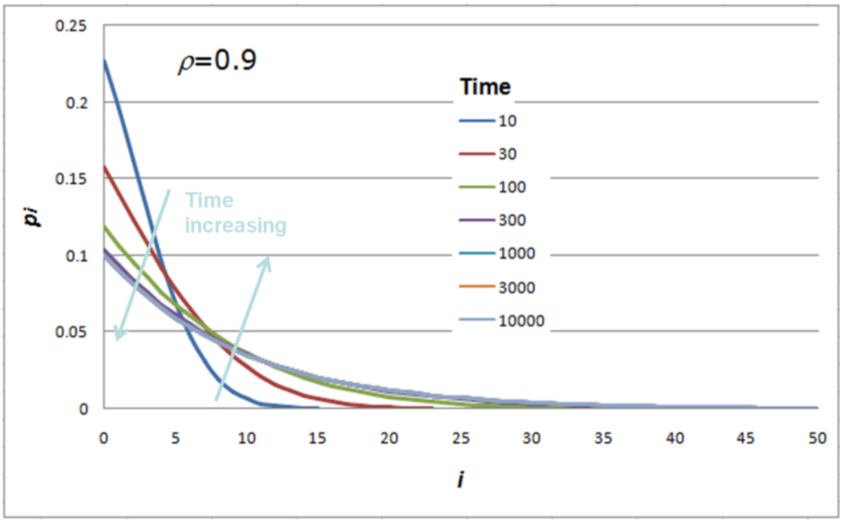


Figure 2.5.7 M/M/1 distribution sequence generated by Markov simulation

According to equations (2.3.32-33) there is a fixed relationship between equilibrium mean and variance, L^2/V=ρ, so if 'deviation from equilibrium' is defined broadly by equation (2.5.2). This deviation rises before it falls to zero, as shown in Figure 2.5.8. Equilibrium is established after approximately 3 times the relaxation time, 379.7 in this case, with μ=1[25].

$$Deviation = \frac{L^2}{V} - \rho \qquad (2.5.2)$$

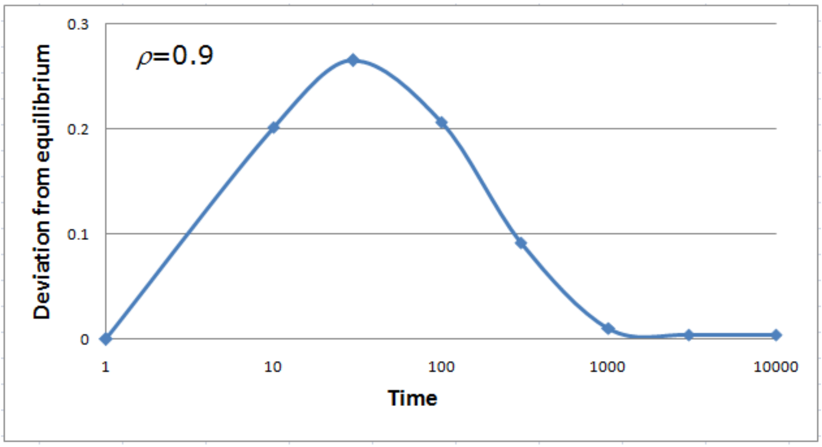


Figure 2.5.8 Deviation from equilibrium in M/M/1 queue development

[25] Units of time are arbitrary, since all queue processes as modelled here depend only on throughput.

2.6. CONCLUSIONS ON QUEUING PROCESSES

Chapter 2 has defined basic queuing methods, the M/M/1 and M/D/1 processes representing queues resulting from random arrivals and respectively random and uniform service. The elemental processes have been described using recurrence relations, that have been used to specify Benchmark programs against which analytical approximations can be compared. In road traffic terms these represent queues at (idealised) give-way/yield and signal junctions. The latter case has been extended to allow explicitly for green period capacity.

Based on the same recurrence relations, a new formula for the deterministic time-dependent variance of a queue has been derived.This arises from both the M/M/1 and M/D/1 cases, and the form of the result suggests that it is general. In what follows it will be used extensively, with the aim of enabling the variance of a time-dependent queue to be estimated alongside the mean, and realistic probability distributions to be estimated from these moments.

CHAPTER 3: MORE GENERAL EQUILIBRIUM QUEUES

3.1. INTRODUCTION

In modelling road traffic, processes are usually restricted to random arrivals with random or uniform service, representing yield or give-way junctions and signals respectively, but there can be variations, associated for example with bunching. The Pollaczek-Khinchin mean queue formula allows for some variation in queue process statistics through a randomness parameter that technically refers to service only. This Chapter derives an extension to include dispersion of arrivals, then applies a similar method to get an equivalent formula for equilibrium variance. It goes on to investigate a number of other processes that feature in standard works, with their probability distributions, and to relate these to the extended equilibrium formulae. It then obtains a relationship between queue moments and certain probability distributions, that in principle allows the latter to be estimated for processes that do not fit the P-K model. Finally, in the absence of an exact formulation, it develops new empirical approximations to the moments of the M/D/1[G] stochastic signal queue with different green times.

3.2. QUEUE PROCESSES WITH GENERAL STATISTICS

3.2.1 G/M/1 with modified variance of arrival headways

Bunday (1996) shows that, for a general arrival headway distribution $a(t)$ and random service, queue moments are got by replacing ρ by an 'effective rho' η, being the solution of:

$$\eta = \int_0^\infty a(t) e^{-(1-\eta)\mu t} dt \qquad (3.2.1)$$

The Erlang-*r* arrival headway distribution, with arrivals at mean rate *q*, is usually interpreted as bunched so that their variance is reduced by the factor *r*:

$$a(t) = \frac{rq(rqt)^{r-1} e^{-rqt}}{(r-1)!} \quad \text{so that} \quad \text{mean}[a]=1/q,\ \text{var}[a]=1/(rq^2) \qquad (3.2.2)$$

In this case, η is given by the solution of (3.2.3), as shown in Figure 3.2.1:

$$\eta = \left(\frac{r\rho}{r\rho + 1 - \eta} \right)^r \qquad (3.2.3)$$

Transformation between ρ and η in principle enables a variety of processes to be accommodated, parameterised by the relationship between the mean and variance of headways, though a physical interpretation may no longer be clear. When *r*=1, this reduces to η=ρ which is just M/M/1.

3.2.2 G/M/1 with scheduled arrivals

As *r* becomes larger, representing arrival headways with *smaller* variance, the relationship approaches the lowest curve in Figure 3.2.1:

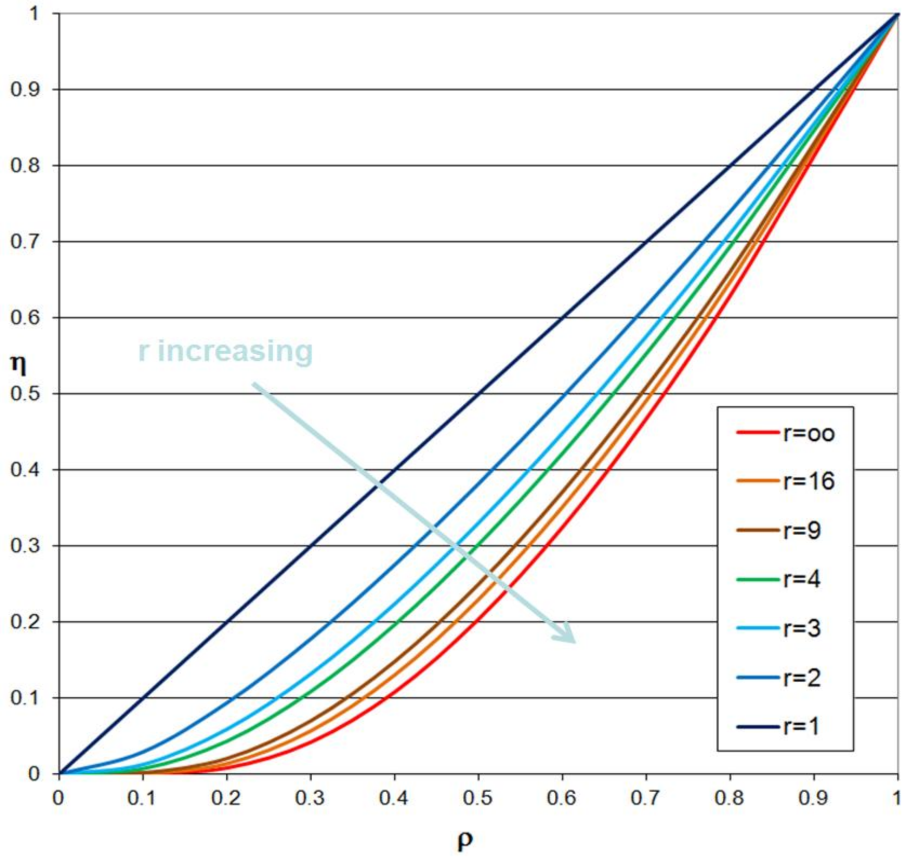


Figure 3.2.1 Relationship between η and ρ for various E*r*/M/1

Bunday shows η for *r*=∞ is the solution of (3.2.4), pointing out that this is equivalent to 'scheduled' arrivals, that occur at regular intervals but still with random service, in which case *a*(*t*) in (3.2.1) is replaced by the Dirac delta function, the relationship becoming similar to the defining form of Lambert's W function:

$$\eta = e^{-\frac{(1-\eta)}{\rho}} \qquad (3.2.4)$$

It is not surprising that this is independent of the schedule interval, and depends only on the demand intensity, given that no long-term average result can depend on the time scale. Scheduled arrivals are unusual in free-moving traffic but may occur in traffic moving forward cyclically within a queue, or any process where arrivals are effectively metered. Of more practical significance is the case where arrivals form ‘green waves’ generated by coordinated service in a signalised network, and randomness of arrivals is reduced. Service may also be non-random, and the exact relationship of green waves to signal cycles will be important.

3.2.3 M/G/1 process with Erlang-*m* service

If the service time probability distribution is similar in form to (3.2.3):

$$b(t)=\frac{m\mu(m\mu t)^{m-1}e^{-m\mu t}}{(m-1)!} \qquad \text{so that} \quad \text{mean}[b]=1/q,\ \text{var}[b]=1/(m\mu^2) \qquad (3.2.5)$$

the probability that *j* units arrive in one service time, assuming exponentially distributed (random) arrivals, is:

$$\pi_j=\int_0^\infty \frac{e^{-qt}(qt)^j}{j!}\,\frac{m\mu(m\mu t)^{m-1}e^{-m\mu t}}{(m-1)!}\,dt=\frac{(j+m-1)!}{j!(m-1)!}\,\frac{\left(\rho/m\right)^j}{\left(1+\rho/m\right)^{j+m}} \qquad (3.2.6)$$

These probabilities can in theory be used to calculate queue size probabilities, using relationships between generating functions. This is practicable for m=1, M/M/1, and also for m=∞, which can be interpreted as M/D/1, but the general case looks like heavy going.

3.2.4 Generalised notation for deriving moments of G/G/1

An alternative approach using expectation values of arrival and service moments, used by Kleinrock (1975) and Bunday (1996) to derive the Pollaczek-Khinchin mean queue formula, appears more user-friendly. This can be used to derive the mean steady-state queue including a modification for dispersion of arrivals I_a following Heydecker (unpublished), with extension to variance. In addition to the variables in the Definitions in Chapter 2, define[26]:

ς_n = the number of units arriving during the service time of unit n

q_n = the number of units in the system at the end of the service of n

$U(q_n)$ = 1 if a unit remains in the system i.e. q_n> 1, otherwise 0, so if $U(q_n)$ is 1 then service will occur in service interval n+1.

[26] This notation used by Bunday is believed to be due originally to Kleinrock, or possibly Kendall before him.

The number of units in the system evolves with each service event according to: (number-in-system after service event *n*+1) equals (number-in-system after service event *n*) minus (unit serviced) plus (arrivals during service of *n*+1), i.e:

$$q_{n+1} = q_n - U(q_n) + \varsigma_{n+1} \tag{3.2.7}$$

Useful intermediate results for the steady-state include:

$$[U(q_n)]^m = U(q_n) \forall m \neq 0 \qquad q_n U(q_n) = q_n \qquad E(q_n) \equiv L_e$$

$$E(\varsigma_n) = \rho \qquad E(U(q_n)) = \rho \qquad E[b] = 1/\mu \qquad \lambda E[b] = \rho \tag{3.2.8}$$

$$E(\varsigma_n^2) = \int_0^\infty \left[I_a \lambda s + (\lambda s)^2\right] b(s) ds = I_a \lambda E[b] + \lambda^2 \left[E[b]^2 + \mathrm{var}[b]\right]$$

$$= I_a \rho + \rho^2 + \lambda^2 \mathrm{var}[b] = I_a \rho + \rho^2 \left(1 + c_b^2\right) = I_a \rho + 2\rho^2 C \tag{3.2.9}$$

where C is the randomness coefficient: $C \equiv \frac{1}{2}\left(1 + c_b^2\right)$ (3.2.10)

3.2.5 G/M/1 with compound or batched arrivals

Heydecker's method of accounting for dispersion of arrivals assumes a compound Poisson process, that can be interpreted as arrivals in batches where the arrival rate of the *batches* accords with a Poisson process. This leads to a relatively simple closed-form modification to the steady-state mean queue as derived by Kleinrock (1975). While more general than the Erlang distribution approach of sections 3.2.1-2, it cannot cover all possible arrival processes.

After squaring (3.2.7) and simplifying some terms using (3.2.8):

$$q_{n+1}^2 = q_n^2 + U(q_n)^2 + \varsigma_{n+1}^2 + 2q_n\varsigma_{n+1} - 2\varsigma_{n+1}U(q_n) - 2q_n U(q_n) \tag{3.2.11}$$

$$= q_n^2 + U(q_n) + \varsigma_{n+1}^2 + 2q_n\varsigma_{n+1} - 2\varsigma_{n+1}U(q_n) - 2q_n$$

The next step is to take expectations, cancelling the first two terms in the steady state, and noting that ς_{n+1} is independent of q_n and $U(q_n)$ so $E(q_n\varsigma_{n+1}) = E(q_n)E(\varsigma_{n+1})$ etc.

Using equations (3.2.8-10), equation (3.2.11) becomes, term by term:

$$\rho + \left(I_a \rho + 2\rho^2 C\right) + 2\rho L_e - 2\rho^2 - 2L_e = 0 \tag{3.2.12}$$

Rearranging terms then gives a variant of the Pollaczek-Khinchin mean value formula:

$$L_e = \frac{(I_a+1)\rho + 2(C-1)\rho^2}{2(1-\rho)} \cong I\rho + \frac{(I_a-1)\rho}{2(1-\rho)} + \frac{C\rho^2}{(1-\rho)} \qquad (3.2.13)$$

In the RH expression, a 'unit-in-service' parameter *I* is substituted for the 1 in the term (*C*-1). Following the approach of Kimber and Hollis (1979), this is considered to reflect unavoidable mean time in service as when, having left the formal queue, waiting for a gap at a priority junction. Its validity is discussed in the next sub-section. Equation (3.2.13) is equivalent to the form quoted earlier as equation (2.3.50):

$$L_e = \frac{I^*\rho + (C-I)\rho^2}{(1-\rho)} \qquad (3.2.14)$$

$$\text{where } I^* \equiv I + \tfrac{1}{2}(I_a - 1) \qquad (3.2.15)$$

Variations from M/M/1 (I_a=1, I=1, c_b=1, C=1) can be represented either by the coefficients in (3.2.13-15) or by the Erlang factors *r* from (3.2.2), which is the inverse of the dispersion of arrivals r=1/I_a, and *m* from (3.2.5) that is related to *C*:

$$L_e = \frac{\rho}{1-\rho}\left[I(1-\rho) + \tfrac{1}{2}(I_a-1) + C\rho\right] = \frac{\rho}{1-\rho}\left[I(1-\rho) + \frac{1-r}{2r} + \left(\frac{1+m}{2m}\right)\rho\right] \qquad (3.2.16)$$

Equation (3.2.16) could in principle allow for the effect of signal green waves and to some extent coordination, although the dominant effect to be accounted for is the overlap of the arriving green waves with green phases that strongly affects the red/green phase component.

3.2.6 Analysis of the P-K derivation and source of the unit-in-service factor

The divisor (1-ρ) is common to all equilibrium queue formulae, and ensures realistically that no finite equilibrium queue can form unless ρ<1. It arises from the last two terms on the RHS of (3.2.7), representing the expected net gain during the next service interval. Individually these terms have the same expectation ρ, but in (3.2.11) they are multiplied by the first term, which creates an asymmetry. Since ς_{n+1} is independent of q_n while $U(q_n)$ is not, and the expectation of ς_{n+1} is ρ while $U(q_n)$=1 for all non-zero queue states, the result is L_e (1-ρ). Expectations of higher powers of q_n can be expected similarly to include a term involving the appropriate moment multiplied by factors including (1-ρ), which will then appear as a divisor in the expression for the moment.

In the earlier analysis of M/M/1 versus M/D/1, the presence of the unit-in-service is associated with whether a queue can be described using infinitesimal service intervals, or requires finite service periods. For G/M/1, service intervals are infinitesimal by default, so this route to deriving the parameter is not available. In equation (3.2.7), q_n *includes* the unit-in-service, and service *will* occur in period n+1 if $q_n > 0$. This suggests that the queue without unit-in-service, say $\tilde{q}_n$, can be represented (superficially at least) by $q_n - U(q_n)$. Substituting this on both sides of (3.2.7), and moving a term from left to right results in:

$$\tilde{q}_{n+1} = \tilde{q}_n - U(q_{n+1}) + \varsigma_{n+1} \qquad (3.2.17)$$

Now $U(q_n)$ is zero if and only if q_n=0, but $\tilde{q}_n$ is zero if q_n is either 0 or 1. However, this is not a concern since $\tilde{q}_n$ can be decoupled from $U(q_n)$. Its expectation is then L_e if time in service is not normally included, although this is numerically different from the expectation of q_n. There is no need to replace $U(q_{n+1})$, provided it is recognised that it is no longer independent of ς_{n+1}.

Table 3.2.1 expands the calculation. The boxes represent the terms of interest in equation (3.2.17), where the repeated block for q_n = 0 has been excluded since it is 'invisible' to the equation. It can be seen directly that ς_{n+1} is independent of $U(q_n)$ but not of $U(q_{n+1})$, as a result of the two cells shaded that are 1 instead of zero.

Table 3.2.1 Expansion of terms in equation (3.2.17) (*m* represents any integer > 1)

q_n	$U(q_n)$	$\tilde{q}_n$	ς_{n+1}	q_{n+1}	$U(q_{n+1})$	$\tilde{q}_{n+1}$
.	-		+	=		
0	0	0	0	0	0	0
0	0	0	1	1	1	0
0	0	0	*m*	*m*	1	*m*-1
1	1	0	0	0	0	0
1	1	0	1	0	1	0
1	1	0	*m*	*m*	1	*m*-1
2	1	1	0	1	1	0
2	1	1	1	2	1	1
2	1	1	*m*	*m*+1	1	*m*
3	1	2	0	2	1	0
3	1	2	1	3	1	1
3	1	2	*m*	*m*+2	1	*m*
etc...		...			...	

The contribution to the expectation of $\varsigma_{n+1}U(q_{n+1})$ is ρ from the ς_{n+1} column, but this applies only when $\tilde{q}_n$=0, hence the expectation is ρ(1-ρ). Since $E(\tilde{q}_n) = L_e$ by definition,

$E(U(q_n)) \equiv E(U(q_{n+1}))$, and all other terms are unchanged, the only effect on (3.2.12) is to replace the $2\rho^2$ term by 2ρ, hence:

$$\rho + \left(I_a\rho + 2\rho^2 C\right) + 2\rho L_e - 2\rho - 2L_e = 0 \qquad (3.2.18)$$

which rearranges to:

$$L_e = \frac{(I_a - 1)\rho + 2C\rho^2}{2(1-\rho)} = \frac{(I_a - 1)\rho}{2(1-\rho)} + \frac{C\rho^2}{(1-\rho)} \qquad (3.2.19)$$

Thus removing the unit-in-service has indeed eliminated the $I\rho$ term in (3.2.13).

The RH form of equation (3.2.16) is probably of limited practical use, but reveals that there is only partial symmetry between arrival and service processes. In the case I=C=1 (M/M/1-like), the effect of dispersion in the arrivals is to multiply the steady-state mean queue by a constant factor, whereas in the case I=0, I_a=1 (M/D/1-like), the randomness factor C has this effect. According to (3.2.16), if arrivals are uniform, I_a=0 or D/M/1, the smallest queue is half the M/M/1 value. If *both* arrivals and service are uniform (I_a=0, C=½), then the purely deterministic case D/D/1 should apply and (3.2.16) reduces to:

$$L_e = \left(I - \tfrac{1}{2}\right)\rho \qquad (3.2.20)$$

With a unit-in-service, equation (3.2.20) says that under completely uniform conditions a mean queue of ρ/2 should still form, so the average waiting time is half the average service time, 1/(2μ). This seems reasonable for a random 'snapshot', since arrivals and service need not be synchronised. Without a unit-in-service (I=0), equation (3.2.20) appears to predict a 'negative queue'. As pointed out earlier, the introduction of arrivals dispersion I_a hinges on an assumption about the type of arrival process, so if this were relaxed the anomaly might be avoided. However I is not a statistical parameter in the normal sense, and I=0 is associated with finite service intervals, which is a limiting case in Erlang terms, so the analysis may no longer be sound. Since uniform arrivals at low intensity are unlikely to be of major interest, the peculiar result can reasonably be neglected for present purposes. However, Kouvatsos (1988) has the discussed the equilibrium form of G/G/1 queues from the Maximum Entropy viewpoint, which may be worthwhile pursuing in further research.

3.3. GENERALISING EQUILIBRIUM VARIANCE

3.3.1 Motivation and approach

This Section extends the method of the previous Section to get a corresponding expression for the equilibrium variance of queue size. It then considers an alternative formulation based on the general Pollaczek-Khinchin method and gets the same result. However, when uniform service is assumed, the result is not that for the usual M/D/1 queue since it includes a unit-in-service component that arises naturally as in the similar procedure for the mean queue. This result is exploited to suggest how the unit-in-service coefficient can be accounted for also in the equilibrium variance.

3.3.2 Variance of the G/G/1 process calculated from expectations

To calculate the variance of q_n the product of (3.2.7) and (3.2.11) is evaluated:

$$\begin{aligned} q_{n+1}^3 &= q_n^3 + q_n U(q_n) + q_n \varsigma_{n+1}^2 + 2q_n^2 \varsigma_{n+1} - 2\varsigma_{n+1} q_n U(q_n) - 2q_n^2 \\ &- q_n^2 U(q_n) - U(q_n)^2 - \varsigma_{n+1}^2 U(q_n) - 2q_n U(q_n)\varsigma_{n+1} + 2\varsigma_{n+1} U(q_n)^2 + 2q_n U(q_n) \\ &+ q_n^2 \varsigma_{n+1} + U(q_n)\varsigma_{n+1} + \varsigma_{n+1}^3 + 2q_n \varsigma_{n+1}^2 - 2\varsigma_{n+1}^2 U(q_n) - 2q_n \varsigma_{n+1} \end{aligned}$$

$$\begin{aligned} &= q_n^3 + q_n + q_n \varsigma_{n+1}^2 + 2q_n^2 \varsigma_{n+1} - 2\varsigma_{n+1} q_n - 2q_n^2 \\ &- q_n^2 - U(q_n) - \varsigma_{n+1}^2 U(q_n) - 2q_n \varsigma_{n+1} + 2\varsigma_{n+1} U(q_n) + 2q_n \\ &+ q_n^2 \varsigma_{n+1} + U(q_n)\varsigma_{n+1} + \varsigma_{n+1}^3 + 2q_n \varsigma_{n+1}^2 - 2\varsigma_{n+1}^2 U(q_n) - 2q_n \varsigma_{n+1} \end{aligned} \qquad (3.3.1)$$

Taking expectations as in the previous Section:

$$\begin{aligned} 0 &= L_e + L_e E\left(\varsigma_{n+1}^2\right) + 2\rho E\left(q_n^2\right) - 2L_e\rho - 2E\left(q_n^2\right) \\ &- E\left(q_n^2\right) - \rho - \rho E\left(\varsigma_{n+1}^2\right) - 2L_e\rho + 2\rho^2 + 2L_e \\ &+ \rho E\left(q_n^2\right) + \rho^2 + E\left(\varsigma_{n+1}^3\right) + 2L_e E\left(\varsigma_{n+1}^2\right) - 2\rho E\left(\varsigma_{n+1}^2\right) - 2L_e\rho \end{aligned} \qquad (3.3.2)$$

So

$$3(1-\rho)E\left(q_n^2\right) = -\rho + 3\rho^2 + 3L_e(1-2\rho) + 3(L_e - \rho)E\left(\varsigma_{n+1}^2\right) + E\left(\varsigma_{n+1}^3\right)$$

(3.3.3)

To evaluate this, in addition to (3.2.11) the value of the following is needed:

$$E\left(\varsigma_n^3\right) = M_3\left(\varsigma_n\right) + 3E\left(\varsigma_n^2\right)E\left(\varsigma_n\right) - 2E\left(\varsigma_n\right)^3 = \int_0^\infty \left[J_a\lambda s + 3I_a(\lambda s)^2 + (\lambda s)^3\right]b(s)ds \tag{3.3.4}$$

In (3.3.4), J_a (for arrivals) is derived by using a relationship between skewness, variance and mean for the compound Poisson quoted by Willmot (1986), where M_3 is the third moment:

$$skewness \equiv \frac{M_3}{\sigma_a^3} = \frac{\left\{3\sigma_a^2 - 2\lambda + 3\left(\sigma_a^2 - \lambda\right)^2/\lambda\right\}}{\sigma_a^3} \tag{3.3.5}$$

For ordinary Poisson, where $\sigma_a^2=\lambda$ (mean arrival rate), this reduces to λ/σ_a^3, i.e. $M_3=\lambda$, while for compound Poisson, where $\sigma_a^2=I_a\lambda$, it evaluates to $J_a\lambda/\sigma_a^3$ with the new definition of σ_a, whence:

$$J_a = I_a^2 + I_a - 1 \tag{3.3.6}$$

The first two terms of (3.3.4) are got using (3.2.7-8), and the third term is equal to:

$$\alpha^3\left(M_3[b] + 3E[b]\text{var}[b] + E[b]^3\right) \approx \rho^3\left(1 + 3c_b^2 + 2\psi c_b^3\right) \tag{3.3.7}$$

where $M_3(b)$ is the third (central) moment of service and the third term arises because for the exponential distribution skewness is 2. This is just a convenient approximation, as the service time distribution need not always be assumed to be exponential, so the factor ψ (default=1) allows for adjustment of relative skewness, with J_a defined by (3.3.6):

$$E\left(\varsigma_n^3\right) = J_a\rho + 6I_aC\rho^2 + 6\rho^3\left(C - J\right) \qquad \text{where} \quad J \equiv \tfrac{1}{3}\left(1 - \psi c_b^3\right) \tag{3.3.8}$$

Evaluating (3.3.3) with (3.2.8) and (3.3.8):

$$E\left(q_n^2\right) = \frac{3\rho - 6\rho^2 + 3\rho^3 + \left(9\rho^2 - 3\rho^3 - 6\rho^4\right)C + 6\rho^4C^2 - 6\rho^3(1-\rho)J}{3(1-\rho)^2}$$
$$+ \frac{\left(L_e - \rho\right)\left(I_a - 1\right)}{(1-\rho)} + \frac{\left[\left(I_a^2 + I_a - 1\right)\rho + 6\left(I_a - 1\right)C\rho^2\right]}{3(1-\rho)} \qquad \text{hence}$$

$$E\left(q_n^2\right) = \rho + \frac{\rho^2\left(3-\rho-2\rho^2\right)C + 2\rho^4 C^2}{(1-\rho)^2} + \frac{\left(L_e + \tfrac{1}{3}(I_a-1)\rho + 2C\rho^2\right)(I_a-1)}{(1-\rho)} - \frac{2\rho^3 J}{(1-\rho)}$$

(3.3.9)

Here J_a, being a function of I_a, has been absorbed, while J remains distinct. Rearranging:

$$V_e = \rho(1-\rho) + \frac{3\rho^2(1-\rho)C + \rho^4 C^2}{(1-\rho)^2} + \frac{\left(L_e + \tfrac{1}{3}(I_a-1)\rho + 2C\rho^2\right)(I_a-1)}{(1-\rho)} - \frac{2\rho^3 J}{(1-\rho)}$$

(3.3.10)

When I_a=1, c_b=1, whence C=1, J=0, this reduces to the M/M/1 result:

$$V_e = \frac{\rho}{(1-\rho)^2} \qquad \text{(M/M/1)} \qquad (3.3.11)$$

When I_a=1, and C=$\frac{1}{2}$ (c_b=0, J_b=$\frac{1}{3}$), it gives a result for Deterministic service:

$$V_e = \frac{\rho\left(12-18\rho+10\rho^2-\rho^3\right)}{12(1-\rho)^2} \qquad \text{(M/D/1 ‘with u.i.s’}^{27}\text{)} \qquad (3.3.12)$$

Comparing (3.3.12) with (2.4.9) (with c_b=0 for Deterministic service) shows that the former exceeds the latter by exactly ρ. This can be interpreted as the contribution of the unit in service. The derivation of (2.4.9) assumed no waiting time, and hence no unit in service. At a light signal this is realistic because the saturation flow usually substantially exceeds the average junction capacity, leading to a disproportionate reduction in any random queue component. Assuming I to be independent of c_b, (3.3.10-12) are then consistent with:

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

(3.3.13)

Equation (3.3.13) has been arranged so that each parameter apart from C appears in only one term, analogous to (3.2.16). This is convenient for computation though not necessarily the most instructive form when coefficient values are limited to typical values, since parts of the different terms tend to cancel, simplifying the expression.

[27] By analogy with the concept of unit in service in M/M/1.

3.3.3 Variance of M/G/1 process calculated from a generating function

An alternative to the preceding is to start with a probability generating function. The Pollaczek-Khinchin transform formula (Medhi 2003) with somewhat modified notation[28] is:

$$P(s) = \sum_{0}^{\infty} p_i s^i = \frac{(1-\rho)(1-s)B^*(\lambda(1-s))}{B^*(\lambda(1-s)) - s} \qquad B^*(s) \equiv \int_0^{\infty} b(t)e^{-st}dt \qquad (3.3.14)$$

where $\{p_i\}$ are the steady state queue size probabilities, s is a dummy variable, and the function B^* is the Laplace-Stieltjes transform of the service time probability density function $b(t)$. From the LHS of (3.3.14), moments can be calculated as follows:

$$P(0) = (1-\rho) = \bar{p}_0, \; P(1) \equiv 1, P'(1) = L_e, \; (sP''+P')(1) = V_e + L_e^2 \qquad (3.3.15)$$

The form of B^* in two common cases is (Medhi 2003):

$$B^*_{M/M/1} = \frac{1}{1+(1-s)\rho}, \qquad B^*_{M/D/1} = e^{-(1-s)\rho} \qquad (3.3.16)$$

Equation (3.3.14) does not allow any value for p_0 other than (1-ρ), so is incompatible with M/D/1 as defined in Chapter 2, except possibly if B^* were to vanish when s=1. Therefore $\bar{p}_0$ is identified with the average over the service period (hence the bar) and the complement of utilisation. To evaluate moments it is necessary to take the limit as $s \rightarrow 1$, because direct evaluation causes several lower order terms to vanish giving 0/0. This is dealt with by l'Hôpital's Rule, worked out explicitly here. As it is easier to take the limit as a variable approaches 0, first define ε=1-s and use $d/ds = -d/d\varepsilon$; then in the case of M/M/1:

$$P(s) = \frac{1-\rho}{1-s\rho} = P^*(\varepsilon) = \frac{1}{1+\left(\frac{\rho}{1-\rho}\right)\varepsilon} \qquad (3.3.17)$$

$$L_e = -P^{*\prime}(\varepsilon)\Big|_{\varepsilon\to 0} = \frac{\left(\frac{\rho}{1-\rho}\right)}{\left[1+\left(\frac{\rho}{1-\rho}\right)\varepsilon\right]^2} \quad \rightarrow \quad \frac{\rho}{1-\rho} \qquad (3.3.18)$$

[28] Although different sources share some similarity of notation there is no complete consistency, so familiar symbols for variables commonly met are used here.

$$V_e = \frac{2(1-\varepsilon)\left(\frac{\rho}{1-\rho}\right)^2}{[\ ]^3} + \frac{\left(\frac{\rho}{1-\rho}\right)}{[\ ]^2} - \left(\frac{\rho}{1-\rho}\right)^2 \qquad \rightarrow \qquad \frac{\rho}{(1-\rho)^2} \qquad (3.3.19)$$

where [ ] represents the bracket in the denominator of (3.3.18).

For M/D/1, according to Medhi (2003):

$$P(s) = \frac{(1-\rho)(1-s)}{1-se^{\rho(1-s)}} = P^*(\varepsilon) = \frac{(1-\rho)\varepsilon}{1-(1-\varepsilon)e^{\rho\varepsilon}} \qquad \text{so} \qquad (3.3.20)$$

$$L_e = -P^{*'}(\varepsilon)\Big|_{\varepsilon\to 0} = \frac{(1-\rho)\left[\left(1-\rho\varepsilon+\rho\varepsilon^2\right)e^{\rho\varepsilon}-1\right]}{[\ ]^2} \qquad (3.3.21)$$

Approximating $e^{\rho\varepsilon}$ to 2nd order in ε:

$$\begin{aligned} Num(L_e) &= (1-\rho)\left[1+\rho\varepsilon+\tfrac{1}{2}\rho^2\varepsilon^2-\rho\varepsilon-\rho^2\varepsilon^2-\tfrac{1}{2}\rho^3\varepsilon^3+\rho\varepsilon^2+\rho^2\varepsilon^3+\tfrac{1}{2}\rho^3\varepsilon^4-1\right] \\ &= (1-\rho)\left[\left(\rho-\tfrac{1}{2}\rho^2\right)\varepsilon^2+O\left(\varepsilon^3\right)\right] \\ Den(L_e) &= \left[1-\rho+\left(p-\tfrac{1}{2}\rho^2\right)\varepsilon+O\left(\varepsilon^2\right)\right]^2\varepsilon^2 = (1-\rho)^2\varepsilon^2+O\left(\varepsilon^3\right), \qquad \text{so when } \varepsilon\to 0 \end{aligned}$$

$$L_e = \frac{\rho-\frac{1}{2}\rho^2}{1-\rho} = \rho+\frac{\rho^2}{2(1-\rho)} \qquad (3.3.22)$$

Medhi (2003) goes on to point out (using different notation) that:

$$p_{0e} = 1-\rho \qquad (3.3.23)$$

consistent with steady-state utilisation but differing from the M/D/1 value derived earlier. Comparing (3.3.22) with (2.4.6), this appears to represent M/D/1 *with* unit-in-service, which is not normally considered to apply at a signalised junction since the time in the green phase, being the inverse of saturation flow, is much less than the average service time[29,30].

The nature of the approximation means that truncating the exponential in (3.3.20) to finite powers ≥2 makes no difference to the M/D/1 result, while truncating it to the first two terms just gives the M/M/1 form, so effectively there are only two possible results.

[29] The parameter *C* is sometimes modified empirically from 0.5 to account for deviations from the ideal.

[30] Including or excluding an ‘in-service’ component could be less important than neglecting the effect of the upstream-moving discharge wave, i.e. treating the queue as ‘vertical’ rather than ‘horizontal’.

To calculate the variance V_e in this way, the last of equations (3.3.15) must be evaluated with (3.3.20). Judging by the foregoing, this promises to be tedious, although terms of order ε^2 and below may well cancel so only terms in ε^3 need be evaluated. Alternatively, equation (3.3.12) can be checked by evaluating the moments of (3.3.14) numerically. To reduce error, this calculation has been done symmetrically, taking the average of results from *s* slightly <1 and slightly >1. Figure 3.3.1 shows results for ρ values in the range 0.1-0.9, using a logarithmic scale and broken white lines for the M/D/1 'with unit in service' as calculated from (3.2.13) and (3.3.12). The results confirm that the 'without unit in service' results calculated from (2.4.8) and (2.4.9) differ consistently by ρ from 'with unit in service'.

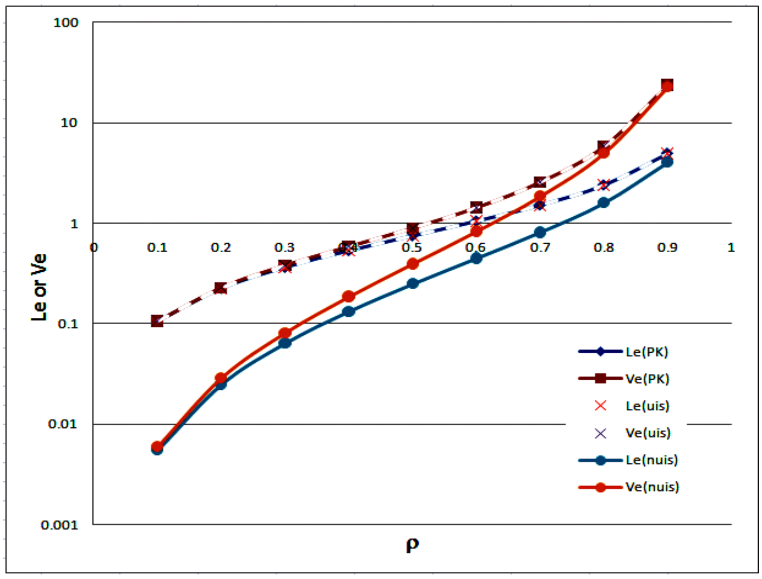


Figure 3.3.1 M/D/1 queue mean and variance calculated by different methods

In conclusion, the results of section 3.3.1 are valid for M/M/1 and M/D/1 'with unit-in-service', and by implication to M/D/1 'without unit-in-service', allowing equations (3.2.13) and (3.3.11) or (3.3.13) to be applied with a degree of confidence.

3.3.4 Horizontal versus vertical queuing and the effect of mixed traffic

As pointed out in the Introduction, real queues occupy finite space, which affects both their growth and discharge behaviour, leading to an increase in delay and maximum extent (Taylor 2005a, 2009). Conversely, when traffic is only slightly slowed and compressed, a 'queue' as normally understood may not be discernible. This is most relevant where traffic remains in motion, and less so where it comes to a halt in a dense queue, as occurs in urban networks. A mixture of traffic types will also affect randomness and effective capacity. As pointed out earlier in Chapter 2, hyper-exponential is a more realistic distribution of arrival and service headways of mixed streams. Investigation of these issues is left for future research.

3.4. CONSISTENT USE OF STATISTICAL PARAMETERS

This Section reviews some features of the statistical coefficients in the P-K formula and alternatives that have been proposed, none of which is entirely satisfactory.

3.4.1 Coefficient of variation of arrivals as an alternative to dispersion index

With the unit in service *I*, after the manner of Kimber and Hollis (1979), along with the index of dispersion I_a, as suggested by Heydecker (2009 unpublished), the Pollaczek-Khinchin mean queue (3.2.16) shows only partial symmetry between the statistics of arrivals and service. Kimber, Summersgill and Burrow (1986) (in their Appendix 2) sought to generalise the randomness coefficient by including the coefficient of variation of arrivals c_a, here defining:

$$C^* = \tfrac{1}{2}\left(c_a^2 + c_b^2\right) \qquad \text{in place of } C = \tfrac{1}{2}\left(1 + c_b^2\right) \qquad (3.4.1)$$

Formula (3.4.1) has a long history, being quoted also by Sakasegawa (1977), and Sakasegawa and Yamazaki (1977) with several variations. Newell (1982) derived an equivalent on the assumption of 'heavy traffic', i.e. $\rho \approx 1$. Gross *et al* (2008) quote it *for any* ρ. However, referring back to the derivation (3.2.7-12), the origin of the '1' is in dividing $E[b]^2$ through by μ^2, leaving no gap into which c_a^2 can sneak. To be consistent with (3.2.16), c_a ought to satisfy:

$$c_a = \sqrt{1 + \frac{I_a - 1}{\rho}} \qquad (3.4.2)$$

However, this is inconsistent with accounting for the degree of non-randomness in the arrivals either through dispersion I_a or by modifying ρ. If $\rho \approx 1$ then $c_a^2 \approx I_a$, so the two are equivalent in the 'heavy traffic' case. However, (3.4.2) can be undefined if $I_a < 1$ and ρ is sufficiently small, and becomes unbounded as ρ tends to 0.

In an informal attempt to shed light on this, the following idealisation is considered. If a variate *X*, consisting of observations $\{X_i\}$, represents the number of arrivals in a unit time period, then its 'inverse' $X^I = 1/X$, consisting of observations $\{1/X_i\}$, represents the intervals between arrivals, provided that the arrival rate is neither so large that there is a high probability of multiple arrivals in a unit interval, nor so small that the interval between arrivals has to span multiple unit time periods. In this case, the relationship between the variate and its inverse is necessarily symmetrical. If ς represents the mean arrival rate $E(X_i)$, and ς^I the mean arrival interval, σ s.d., *v* variance, the simplest possible relationships satisfying this symmetry principle are:

$$\frac{\sigma^{\mathrm{I}}}{\varsigma^{\mathrm{I}}} = \frac{\sigma}{\varsigma} \quad \text{or} \quad \frac{v^{\mathrm{I}}}{\varsigma^{\mathrm{I}2}} = \frac{v}{\varsigma^2}, \quad \varsigma^{\mathrm{I}} = \frac{1}{\varsigma} + \varsigma v^{\mathrm{I}} \qquad \text{and} \qquad \varsigma = \frac{1}{\varsigma^{\mathrm{I}}} + \varsigma^{\mathrm{I}} v \qquad (3.4.3)$$

In Heydecker's derivation of the index of dispersion I_a, unit time is taken to be the mean service time, so the mean arrival rate ς is numerically equal to ρ, not λ=ρμ. So, from (3.4.3):

$$c_a^2 = \frac{\sigma_a^2}{\tau_a^2} \approx \frac{\sigma_\varsigma^2}{\varsigma^2} = \frac{I_a}{\varsigma} = \frac{I_a}{\rho} \qquad \text{hence} \qquad c_a \approx \sqrt{\frac{I_a}{\rho}} \qquad (3.4.4)$$

This is only trivially consistent with (3.4.2), when ρ≈1, and suffers from the same unboundedness when ρ→0. For values of ρ much different from 1 the assumption about the time unit may fail. Equation (3.4.4) is more consistent than (3.4.2) with the definition of I_a as the index of dispersion, but whichever of c_a or I_a is taken as constant for a given type of process, the other becomes dependent on ρ.

It *is* possible to rewrite (3.2.16) in a form that uses (3.4.1), in accordance with (3.4.4):

$$L_e = I\rho + \frac{\rho^2}{2(1-\rho)} + \frac{C^*\rho^2}{(1-\rho)} \qquad (3.4.5)$$

There is no difference where the only processes considered have I_a=1, namely M/M/1 for yield-type processes and M/D/1 for signal-type processes, but the issue will arise later in the context of diffusion approximations. Which form is preferable may depend on identifying a type of process with non-Markov arrivals that can be characterised in a physically meaningful way independent of ρ. The factor ρ in the I_a term in (3.2.13) or (3.2.16) seems unavoidable.

The form of (3.2.16) suggests that it is I_a that is independent of ρ for an identifiable process such as batched arrivals with a given Erlang factor, and for this reason it seems preferable to c_a. In the perfectly deterministic case C^*=0, while the minimum value of C is ½. Since the I_a term is dropped completely, this results in a simpler value of L_e compared with (3.2.20), which avoids the 'negative queue size' prediction, but appears to be saying in effect that the arrival and service processes must always be out of synchronisation.

$$L_e = I\rho \qquad (3.4.6)$$

For example, if ρ=1, so that arrival and service rates are matched, there is always a customer awaiting service, whereas intuively it would be given a 50:50 chance of ‘catching’ the service without having to wait. There is an interesting analogy here with the ‘rule’ that if passengers come to a bus stop hoping to catch a bus that arrives randomly at a certain mean rate, then the expected waiting time is equal to the mean interval between buses, not half the mean interval as would be expected if buses arrive regularly. These arguments tend to favour (3.2.20) over (3.4.6), hence I_a over c_a.

3.4.2 Consistent use of the unit-in-service between mean and variance

A way to modify the M/M/1 probability distribution to exclude the unit-in-service is to define:

$$\begin{aligned}&\tilde{p}_0 = 1-\rho^2\\ &\tilde{p}_i = (1-\rho)\rho^{i+1} \qquad i > 0\end{aligned} \tag{3.4.7}$$

so, relative to M/M/1:

$$\begin{aligned}&p_0 - \tilde{p}_0 = -\rho(1-\rho)\\ &p_i - \tilde{p}_i = \rho^i(1-\rho)^2 \qquad i > 0\end{aligned} \tag{3.4.8}$$

The distribution (3.4.7) satisfies:

$$\sum_{i=0}^{\infty} \tilde{p}_i = 1 \qquad \sum_{i=0}^{\infty} i\tilde{p}_i = \tilde{L}_e = \frac{\rho^2}{1-\rho} \qquad L_e - \tilde{L}_e = \rho \tag{3.4.9}$$

and it follows directly from (3.4.7-9) that:

$$\sum_{i=0}^{\infty} i^2\tilde{p}_i = \frac{\rho^2(1+\rho)}{(1-\rho)^2} \qquad \tilde{V}_e = \frac{\rho^2\left(1+\rho-\rho^2\right)}{(1-\rho)^2} \qquad V_e - \tilde{V}_e = \rho(1+\rho) \tag{3.4.10}$$

There is no change in the dispersion of arrivals between M/M/1 and M/D/1, but the value of C changes from 1 to ½ as c_b goes from 1 to 0. Taking this into account and comparing (2.4.9) with (3.3.13), the equivalents of (3.4.9-10) for M/D/1 are:

$$\left(L_e - \tilde{L}_e\right)_{M/D/1} = \rho \qquad \left(V_e - \tilde{V}_e\right)_{M/D/1} = \rho \tag{3.4.11}$$

3.5. EQUILIBRIUM DISTRIBUTIONS OF SOME QUEUING PROCESSES

3.5.1 Motivation and approach

This Section addresses a wider range of equilibrium queues found in standard works, including with Erlang arrival or service distributions, and queueing with multiple service channels. Expressions for recurrence relations and equilibrium state probabilities are given where possible, and distributions calculated or simulated using an extended version of the 'Qsim' Markov simulation software program originally developed for M/M/1 and M/D/1. The relationships between the equilibrium moments the Pollaczek-Khinchin formula with different statistical parameters are explored, with the aim of bringing making the queue processes amenable to the time-dependent approximation framework. First, properties of M/M/1 and M/D/1 are recalled for completeness and comparison.

Note: in several Figures in this Section, smoothed continuous charts have been retained in preference to histograms which would be more 'correct' but have been found difficult to interpret in practice, because it is difficult to separate multiple graphs particularly where they are non-monotonic or multi-modal. While technically less accurate, smoothed graphs are felt to give a general impression of the form of the probability distributions.

3.5.2 M/M/1

The M/M/1 process is the archetypal queue process and straightforward in principle, despite the complexity of the exact solution of Morse (1958) in Section 2.3. Its properties lead quickly to the deterministic queue formula repeated here as (3.5.1), which also follows logically from conservation, where L_0 is the queue size at time t=0, and ρ is the demand intensity or the ratio of average demand to average capacity μ:

$$L(t) \equiv L_0 + (\rho - x)\mu t \qquad (3.5.1)$$

As pointed out earlier in Chapter 2, a critical feature of (3.5.1) is that, if the mean queue is to tend to a steady state and its size is to remain finite, the term in brackets must approach zero as time $t\to\infty$, whence $x\to\rho$ as $t\to\infty$. As x is the average utilisation of the service, if the possible states of service are either occupied or unoccupied, unoccupied means the absence of a queue, and the probability p_0 of the queue being zero is identified with this, then in the steady state:

$$p_{0e} = 1 - \rho \qquad (3.5.2)$$

If (3.5.2) is interpreted as the *average* probability of zero queue during the service interval (which in the case of M/M/1 can be treated as infinitesimal so that average = instantaneous), so that by definition of utilisation $p_0(t) \equiv 1 - u(t)$ and $p_0(t) \leftrightarrow 1 - x(t)$ as $t \rightarrow \infty$, then (3.5.2) is the condition of (3.5.1) remaining finite and *any* equilibrating queue process must satisfy both equations (3.5.1) and (3.5.2).

3.5.3 M/D/1

Using the Pollaczek-Khinchin transform formula, Medhi (2003) shows that (3.5.2) applies to the M/D/1 queue *with unit-in-service*, as derived in the previous Section, equation (3.3.22). However, the derivation in Chapter 2 of M/D/1 *without unit-in-service* from recurrence relations, gives the different formula (2.4.7), repeated here as:

$$p_{0(e)} = e^{\rho}(1 - \rho) \qquad (3.5.3)$$

The explanation given is that this refers to the queue size at the *end* of a finite service period (e.g. a signal green period), whereas service can actually occur throughout the period, resulting in a smaller *average* value of p_0, say $\bar{p}_0$, which has been shown to satisfy (3.5.2) and hence is consistent with the deterministic interpretation of utilisation embodied in (3.5.1).

3.5.4 Erlang-*m* service

Erlang service is usually described before Erlang arrivals because it is easier to understand. Kleinrock (1975) and Medhi (2003) interpret M/E*m*/1 as a multi-stage service process where an effective capacity of μ is achieved by *m* exponential service processes each of capacity *m*μ in series. The process is no longer exponential, having a service time distribution given by (3.2.5). This has a simple set of differential recurrence relations resembling M/M/1, but referring to *stages* not customers:

$$\begin{aligned}
&\frac{1}{\mu}\frac{d\vec{p}_0}{dt} = m\vec{p}_1 - \rho\vec{p}_0 \\
&\frac{1}{\mu}\frac{d\vec{p}_i}{dt} = m\vec{p}_{i+1} - (m+\rho)\vec{p}_i \qquad (i<m) \qquad (3.5.5) \\
&\frac{1}{\mu}\frac{d\vec{p}_i}{dt} = m\vec{p}_{i+1} - (m+\rho)\vec{p}_i + \rho\vec{p}_{i-m} \qquad (i\geq m)
\end{aligned}$$

Combined demand intensity is ρ=λ/μ, delivering equilibrium recurrence relations for ρ<1:

$$\vec{p}_1 = \frac{\rho}{m}\vec{p}_0$$

$$\vec{p}_{i+1} = \left(1+\frac{\rho}{m}\right)\vec{p}_i \qquad (i<m) \qquad (3.5.6)$$

$$\vec{p}_{i+1} = \left(1+\frac{\rho}{m}\right)\vec{p}_i - \frac{\rho}{m}\vec{p}_{i-m} \qquad (i\geq m)$$

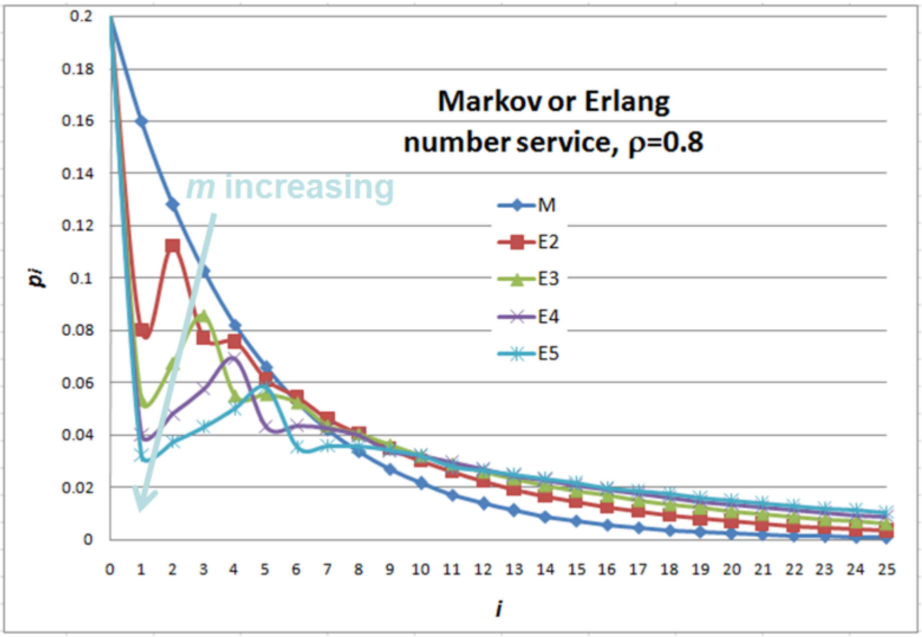


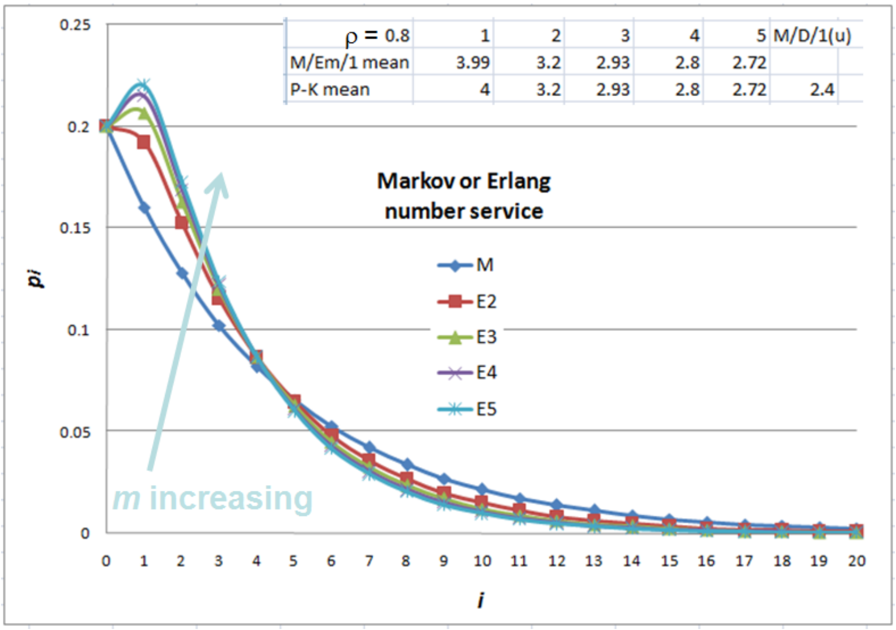
Figure 3.5.1 Markov simulated M/E*m*/1 ‘raw’ probability distributions for several *m*

In Figure 3.5.1, the steady-state distributions of stages, as simulated by the Markov program Qsim, exhibit an oscillatory tendency with period *m*. This is a consequence of the recurrence relation where the *i*+1 element depends on the *i-m* element but not on those between it and element *i*. However, these are not the final distributions in terms of *customers*.

The ‘rule’ governing the system says that only one of the *m* servers can be in use at any one time. Therefore their utilisations are correlated even though their service time distributions are not. Thus the probability of the system being empty is not $(1-\rho/m)^m$. Instead there are *m* ways in which one server can be occupied with a probability of ρ/*m*, so the probability of the system being empty is correctly (1-ρ), which is smaller. The mean of the stage distribution must be divided by *m* to get the mean number of *customers* in the system, which is what is meant here by ‘queue size’ in the traffic context. Accordingly, the customer queue size distribution is given by summing the stage distribution in groups of *m* terms, except for p_0 which is unchanged:

$$p_i = \sum_{j=\max[m(i-1)+1,0]}^{mi} \vec{p}_j \qquad (\, p_0 = \vec{p}_0 = 1-\rho \text{ in the steady state}) \qquad (3.5.7)$$

The resulting distributions are shown in Figure 3.5.2, along with their simulated mean values and the values of the corresponding P-K mean queue formula (3.2.13).



ρ = 0.8	1	2	3	4	5	M/D/1(u)
M/Em/1 mean	3.99	3.2	2.93	2.8	2.72	
P-K mean	4	3.2	2.93	2.8	2.72	2.4

Figure 3.5.2 Markov simulated M/E*m*/1 final probability distributions for several *m*

Griffiths *et al* (2008) show how this particular queue process can be applied to a motorway facility with limited capacity (the Severn Bridge) where service is not perfectly random (*m* = 2 or 3). At finite values of *m*, each service interval admits a Poisson count of arrivals, but it is hard to interpret this in traffic terms, while (3.2.5) has straightforward interpretation as a *service* distribution more uniform than exponential. In the limiting case *m*→∞, all arrivals take place in one uniform service interval, as at a signal. The results confirm the P-K formula and support the idea that M/D/1 *with unit-in-service* can be identified with M/E∞/1. Staged arrivals are considered dual to bulk arrivals (q.v.).

3.5.6 Erlang-*r* arrivals

Kleinrock (1975) and Medhi (2003) interpret E*r*/M/1 as a multi-stage arrival process where each stage involves an arrival rate *r*λ. Although it is hard to interpret this in traffic terms, this system is considered to be dual to Erlang-*m* service, and both Erlang parameters can be combined into one code for calculation, though this fails if both are simultaneously >1:

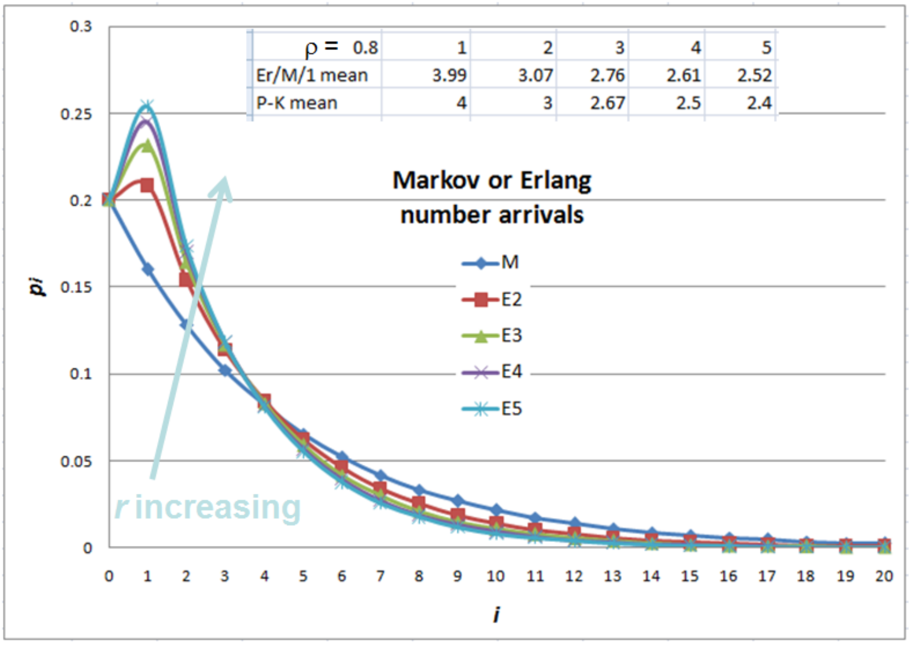
$$\frac{1}{\mu}\frac{d\vec{p}_0}{dt} = \vec{p}_r - r\rho\vec{p}_0$$

$$\frac{1}{\mu}\frac{d\vec{p}_i}{dt} = \vec{p}_{i+r} - r\rho\vec{p}_i + r\rho\vec{p}_{i-1} \qquad \text{(if } i<r\text{)} \qquad (3.5.8)$$

$$\frac{1}{\mu}\frac{d\vec{p}_i}{dt} = \vec{p}_{i+r} - (1+r\rho)\vec{p}_i + r\rho\vec{p}_{i-1} \qquad \text{(if } i\geq r\text{)}$$

The ‘customer’ probabilities are calculated by summing stage probabilities in a different way from Erlang-*m* service, giving the superficially similar distributions in Figure 3.5.3:

$$p_i = \sum_{j=ri}^{r(i+1)-1} \vec{p}_j \qquad \text{(also gives } p_0 = 1-\rho \text{ in the steady state)} \qquad (3.5.9)$$



ρ = 0.8	1	2	3	4	5
Er/M/1 mean	3.99	3.07	2.76	2.61	2.52
P-K mean	4	3	2.67	2.5	2.4

Figure 3.5.3 Markov simulated E*r*/M/1 final probability distributions for several *r*

3.5.6 An alternative interpretation of Erlang-*r* arrivals

As Kleinrock (1975) shows, the following steady-state probabilities apply to E*r*/M/1:

$$p_0 = 1-\rho \qquad\qquad p_i = \rho(1-\hat{\rho})\hat{\rho}^{\,i-1} \qquad (3.5.10)$$

This ‘Russian dolls’ nested structure also applies when $i\geq n$ in M/M/*n* and may be expected once exponential behaviour takes over beyond the scale-dependent region of the probability distribution. In the case of (3.5.10) there is no more scale-dependence than with M/M/1, and the parameter *r* defines the next least complex family of distributions. Equations (3.5.10) lead to simple formulae (3.5.11-12) for the equilibrium mean and variance. Matching (3.2.16) and (3.5.11 left) relates $\hat{\rho}$ to arrivals dispersion or the parameter r. While necessarily correct for I_a=1, r=1, this leads to a singularity in the extended P-K formula when I_a=0 and r=∞, since $\hat{\rho}$ is undefined for $\rho<0.5$.

$$\hat{L}_e = \frac{\rho}{1-\hat{\rho}} \qquad \hat{V}_e = \frac{\rho(1+\hat{\rho}-\rho)}{(1-\hat{\rho})^2} \tag{3.5.11}$$

$$\hat{\rho} = \frac{2\rho + I_a - 1}{I_a + 1} = \frac{(2\rho-1)r+1}{r+1} \tag{3.5.12}$$

3.5.7 Performance of equilibrium moment formulae with Erlang processes

Equilibrium variance can be calculated either directly from the distributions or from equation (3.3.10), or (3.3.13) with I=1 in these cases. Figure 3.5.4 compares calculated and simulated of L_e and V_e, showing a good fit between the two.

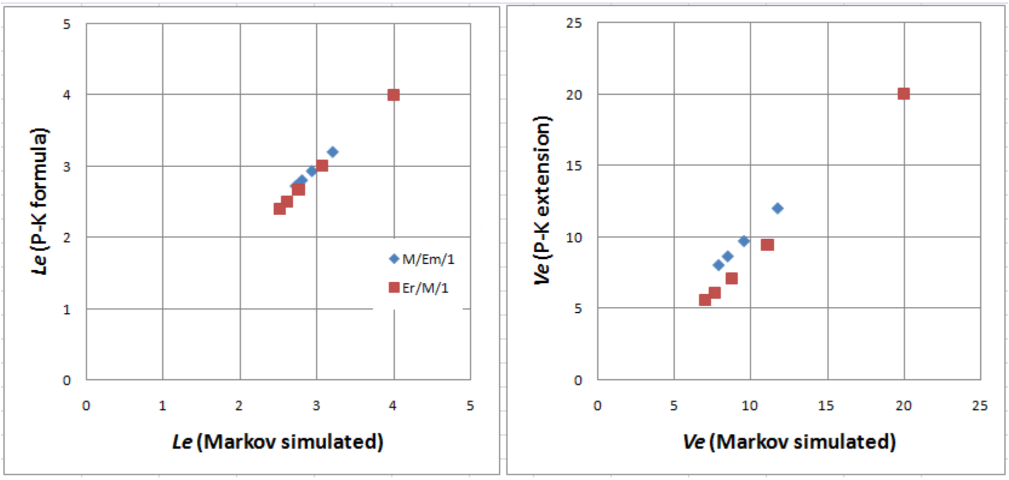


Figure 3.5.4 Calculated versus Markov simulated Erlang distribution moments

3.5.8 Generalising Erlang processes

It is not obvious how a description could include both Er>1 and Em>1, and none has been found in standard works. The following combined method can be programmed conveniently by equations (3.5.13) provided only one of r, m is >1 on any one occasion.

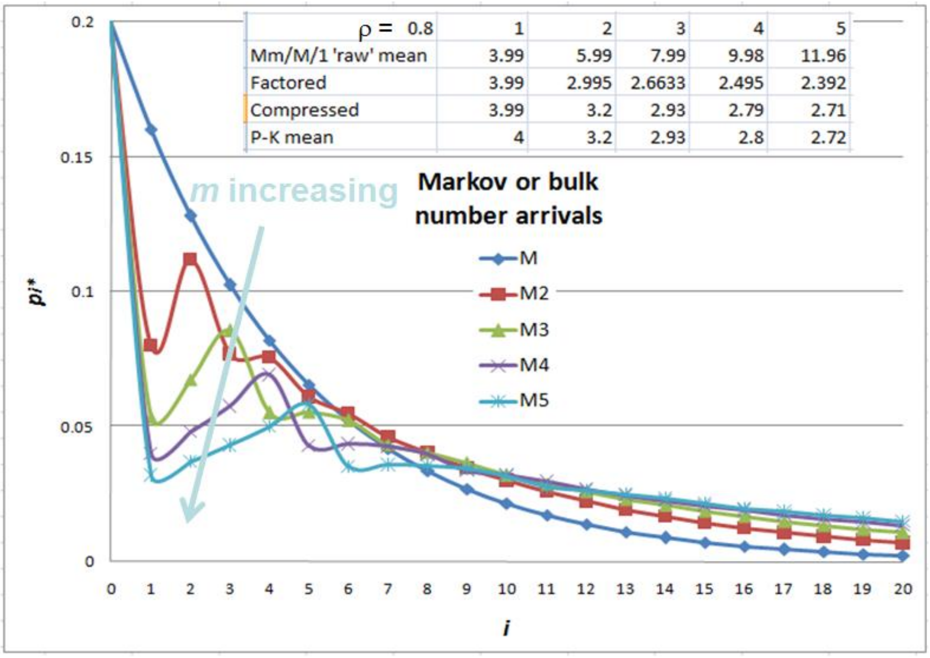
$$\frac{1}{\mu}\frac{d\vec{p}_i}{dt} = m\vec{p}_{i+r} - r\rho\vec{p}_i \qquad + r\rho\vec{p}_{i-m} \ \text{(if } i \geq m\text{)} \qquad - m\vec{p}_i \quad \text{(if } i \geq r\text{)} \tag{3.5.13}$$

There is noticeable asymmetry between the Erlang service and arrival processes, as is reflected in the asymmetry in the generalised P-K mean queue formula (3.2.16). This does encourage combining arrivals and service statistics in one parameter C^*.

3.5.9 Bulk arrivals M*m*/M/1

Bulk arrivals, where arrivals come in lots of *m*, is theoretically congruent to Erlang-*m* service, apart from physical interpretation (Kleinrock 1975). It is simulated by similar but not identical recurrence relations that describe the probability distribution of the number of customers not stages. The Bulk arrivals parameter can be a random variate, usually taken to be Poisson distributed, but the point is sufficiently made by fixed values. The resulting 'raw' distributions are not compressed after the manner of equations (3.5.7) and (3.5.9), and they have a distorted shape because of the low probability of ending up with fewer in the queue than the Bulk value. As anticipated, the results in Figure 3.5.5 are identical to Figure 3.5.1, but a table of means has been added (also in some later sections) where:

- 'Raw' mean is obtained by taking the simulated queue states at face value
- Factored mean is the 'raw' mean divided by the Erlang parameter
- Compressed mean is after application of the appropropriate range summing
- P-K mean is the value of the extended P-K formula with Erlang parameter



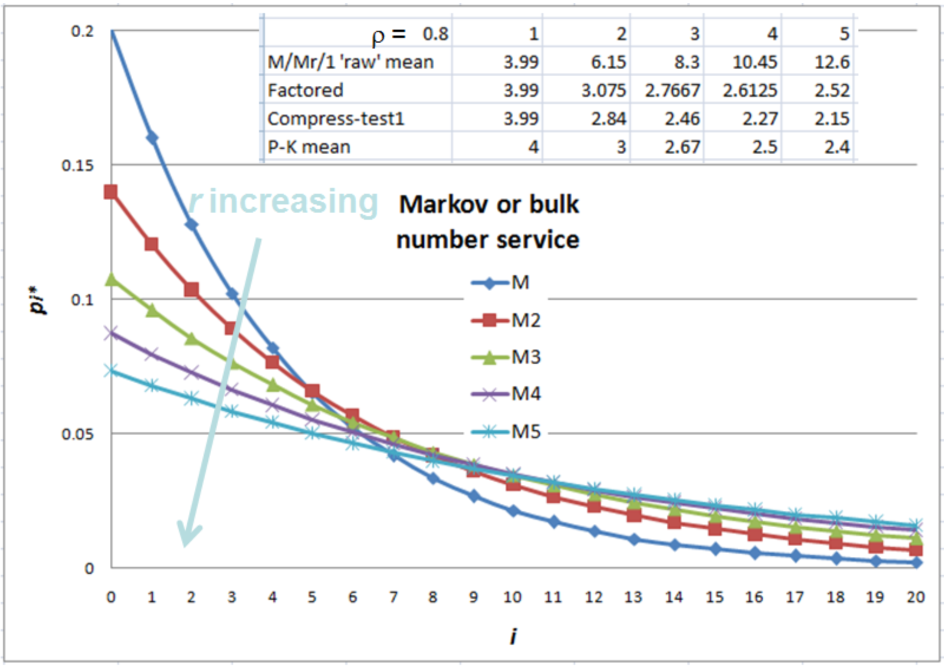
ρ = 0.8	1	2	3	4	5
Mm/M/1 'raw' mean	3.99	5.99	7.99	9.98	11.96
Factored	3.99	2.995	2.6633	2.495	2.392
Compressed	3.99	3.2	2.93	2.79	2.71
P-K mean	4	3.2	2.93	2.8	2.72

Figure 3.5.5 Markov simulated 'raw' Bulk arrivals probability distributions

The compressed and P-K means match very closely

3.5.10 Bulk service M/M*r*/1

This is said to be analogous to Erlang arrivals, and it appears that as for Bulk arrivals the parameterisation of the P-K formula must be ‘reversed’, so that in this case I_a is changed. Simulating this process gives the ‘raw’ distributions graphed in Figure 3.5.6.



ρ = 0.8	1	2	3	4	5
M/Mr/1 'raw' mean	3.99	6.15	8.3	10.45	12.6
Factored	3.99	3.075	2.7667	2.6125	2.52
Compress-test1	3.99	2.84	2.46	2.27	2.15
P-K mean	4	3	2.67	2.5	2.4

Figure 3.5.6 Markov simulated ‘raw’ Bulk service probability distributions

These are geometric (see also Kleinrock 1975) but for *r*>1 the ratio between terms is *not* ρ. The distributions have a particular property involving the first *r*+1 terms enshrined in the recurrence relations used to generate them, plus the usual geometric property:

$$r\rho p_0^* = \sum_{j=1}^{r} p_j^* \qquad (3.5.15)$$

$$p_{i+1}^* = \zeta p_i^* \qquad (3.5.16)$$

Together these lead to the relationship:

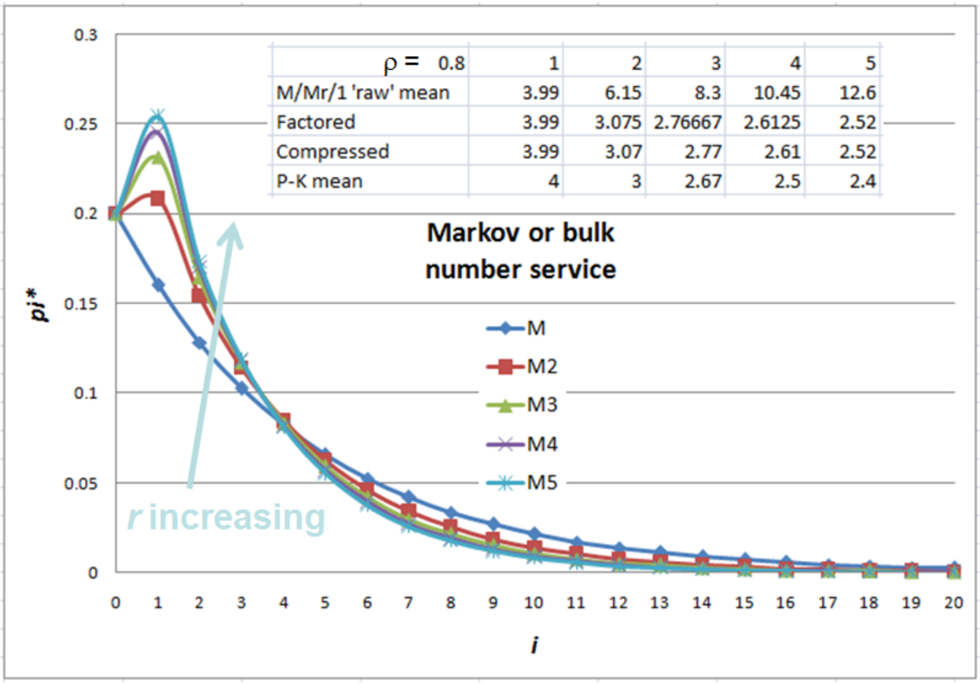
$$\rho = \frac{\zeta\left(1-\zeta^r\right)}{r(1-\zeta)} \qquad (3.5.17)$$

This is soluble for ζ though it is hard to see what it means physically, compared for example to the very natural interpretation of the difference between average and end-of-green p_0 in the M/D/1[G] queue.

Despite the supposed ‘duality’ with Erlang arrivals, compression of the distribution according to equation (3.5.9) does not work in this case, and does not yield the correct values for p_0. Equations (3.5.15-17) allow the following expression for p_0, on eliminating ζ:

$$p_0 = 1-\rho = 1-\frac{\rho}{1-p_0^*}+\left(\rho+\frac{1}{r}\right)p_0^* - \frac{1}{r}p_r^* \qquad (3.5.18)$$

On the *assumption* that the compressed distribution has the nested form (3.5.10), translating to just one element of the compressed distribution, the parameter $\hat{\rho}$ is equal to ζ^r. This is reasonable in the sense that any irregularity is likely to extend only as far as around the first r elements of the ‘raw’ distribution. This necessarily ‘works’ in the sense that any value of $\hat{\rho} < 1$ will generate a valid distribution. The distributions are shown in Figure 3.5.7.



ρ = 0.8	1	2	3	4	5
M/Mr/1 'raw' mean	3.99	6.15	8.3	10.45	12.6
Factored	3.99	3.075	2.76667	2.6125	2.52
Compressed	3.99	3.07	2.77	2.61	2.52
P-K mean	4	3	2.67	2.5	2.4

Figure 3.5.7 Final Markov simulated Bulk service probability distributions

The means of the compressed distributions are now essentially the same as those factored values from the ‘raw’ distributions, i.e. (raw_mean)/r, and now match those for Er/M/1, although as in that case they do not match the P-K means exactly.

3.5.11 Multi-channel queue M/M/*n*

Standard queuing theory for multiple servers centres on the M/M/*n* or G/G/*n* processes, where arrivals choose an idle server if one is available and otherwise choose a channel randomly, there being no interaction between channels' service. This describes something like a supermarket checkout line, airport checkin hall, or toll plaza. Where interaction between channels can occur the process will be more like queuing on multiple lanes sharing a common service process, which is considered later in Chapter 6.

If ρ is the combined demand intensity on the ensemble relative to the total capacity of all the channels, the steady-state queue and queue size probability distribution are given by the following formulae (e.g. Medhi 2003, with modified notation[31]):

$$L_e^{(n)} = n\rho + C(n,\rho)\frac{\rho^2}{1-\rho} \qquad \text{where} \quad C(n,\rho) = \frac{(n\rho)^n p_0^{(n)}}{n!\rho(1-\rho)} \qquad (3.5.19)$$

$$\text{and} \qquad p_0^{(n)} = \left[\sum_{i=0}^{n-1}\frac{(n\rho)^i}{i!} + \frac{(n\rho)^n}{n!(1-\rho)}\right]^{-1} \qquad p_i^{(n)} = \frac{n}{\min(i,n)}\rho p_{i-1}^{(n)} \quad (i>0) \qquad (3.5.20)$$

Equation (3.5.19) has an obvious similarity to the P-K mean queue formula, with *I* being replaced by *n* because each server behaves in the same way as the single server in M/M/1, and *C*(*n*,ρ) taking the place of the statistical parameter *C*, although the analogy is not exact because the coefficient of variation of service within each channel remains 1. Bunday (1996) considers the limiting case M/M/∞ also, a Poisson distribution that remains defined even when ρ>1, and explains it is identical to defection of a fraction *i*/(*i*+1) of arrivals when the queue size is *i*.

$$L_e^{(\infty)} = \rho \qquad V_e^{(\infty)} = \rho \qquad (3.5.21)$$

$$p_0^{(\infty)} = e^{-\rho} \qquad p_i^{(\infty)} = \frac{\rho}{i}p_{i-1}^{(\infty)} = \frac{\rho^i e^{-\rho}}{i!} \quad (i>0) \qquad (3.5.22)$$

M/M/*n* can also be described in terms of defection. This is embodied in the factor in $p_i^{(n)}$. The converse of defection is partial utilisation of the service facilities. If all servers are idle, with probability p_0, the contribution to utilisation is zero. If all servers are busy then the specific contribution to utilisation before factoring in the state probability is 1. Since an arrival always chooses a free server if one is available, all servers are busy when $i \geq n$. For intermediate values of *i*, the specific contribution to utilisation is *i*/*n*.

[31] The notation has been adjusted so that the function *C*(*n*,ρ) takes the place of *C* in the P-K formula.

Multiplying these by probabilities (3.5.20 right) gives absolute contributions of $\rho p_{i\text{-}1}$, summing which confirms that the proper utilisation is consistent with (3.5.1-2)[32]:

$$u = 0.p_0 + \sum_{i=1}^{\infty} \rho p_{i-1}^{(n)} = \rho \qquad \text{hence} \qquad \bar{p}_0 = 1 - \rho \qquad (3.5.23)$$

Probability distributions for n from 1 to 10 according to (3.5.20) are graphed in Figure 3.5.8 (resembling but different from Erlang distributions). Comparing these with equations (3.5.22) for n=∞ gives the impression that they describe two different systems, since as graphed p_0 decreases steadily with n, whereas at n=∞ it jumps to the value $e^{-0.8}$ = 0.4493. Instead of peaking at some i~O(n) the n=∞ distribution (3.5.22) declines steadily with i, provided ρ<1. If equations (3.5.21-22) are taken at face value, then to recover (3.5.23) the specific contributions to utilisation must equal i, so $u^{(\infty)}$ has the same form as $L_e^{(\infty)}$. Division by n becomes meaningless, but specific contributions do increase with i in the same manner as for finite n except that for n=∞ they do so without limit, so (3.5.23) still holds.

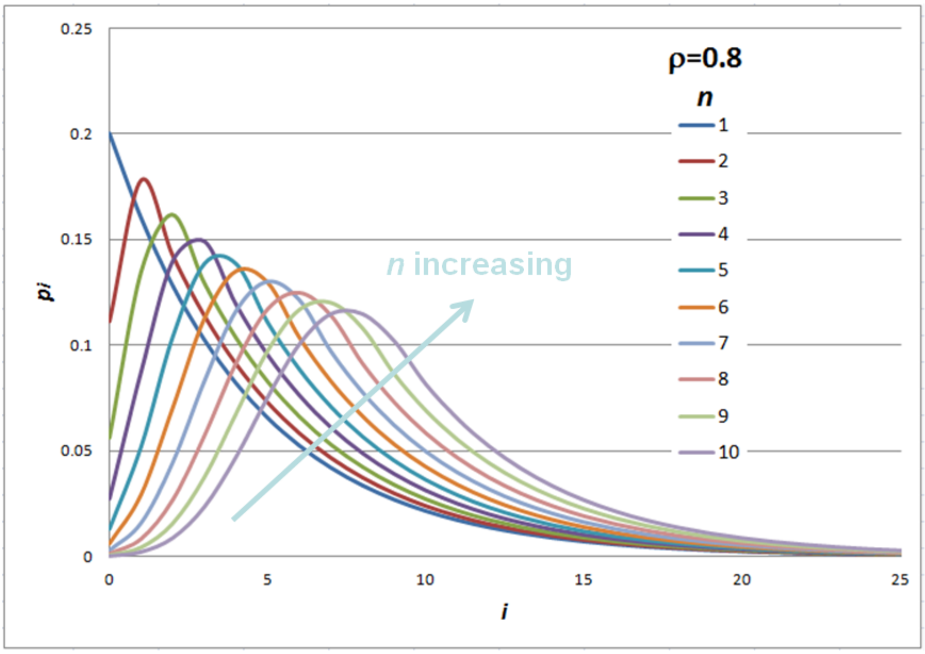


Figure 3.5.8 M/M/n equilibrium probability distributions for several values of n

The dependence of C on ρ means that approximations that rely on the particular form of the Pollaczek-Khinchin mean queue like the sheared solution of Kimber and Hollis (1979) cannot be used to translate from equilibrium to time-dependence, although numerical solutions should be possible if quasi-equilibrium is assumed to apply.

[32] Recall that u represents instantaneous utilisation and x is its time average, but these converge at t=∞.

3.6. USE OF NESTED GEOMETRIC PROBABILITY DISTRIBUTIONS

3.6.1 Singly-nested geometric distributions

Several steady-state probability distributions described in the previous Section become geometric, with a constant ratio between successive state probabilities, at sufficiently high states, although different ratios can apply in low states. The simplest model of this is given by equation (3.6.1), where ρ^* and $\hat{\rho}$ are first and second level 'effective ρs', i.e:

$$p_{0e} = 1-\rho^* \qquad p_i = \rho^*(1-\hat{\rho})\hat{\rho}^{i-1} \ (i \geq 1) \tag{3.6.1}$$

The moments of this distribution are given by:

$$L_e = \frac{\rho^*}{1-\hat{\rho}}, \qquad V_e = \frac{\rho^*(1+\hat{\rho}-\rho^*)}{(1-\hat{\rho})^2} \qquad \text{or} \qquad \hat{\rho} = 1-\frac{\rho^*}{L_e} \tag{3.6.2}$$

In the first of equations (3.6.1), ρ^* could be defined so as to give the correct p_0 for the true ρ. In such cases p_0 would not equal the complement of the steady-state utilisation, but in principle M/D/1 could be accommodated. However, only two 'ρs' are available to fit three moments[33] p_{0e}, L_e and V_e, limiting the range of distributions that can be fitted.

Medhi (2003) quotes (this appears initially to have been misprinted[34]) a nested equilibrium distribution proposed by Kobayashi (1974a) to take on board arrival and service statistics, derived by considering a continuum approximation (see Chapter 5 later):

$$p_0 = 1-\rho \tag{3.6.3}$$

$$p_i = \rho(1-\hat{\rho})\hat{\rho}^{i-1} \tag{3.6.4}$$

$$\text{where} \quad \hat{\rho} = \exp\left\{-\frac{2(1-\rho)}{\mu^2(\rho^3\sigma_a^2+\sigma_b^2)}\right\} = \exp\left\{-\frac{2(1-\rho)}{(c_a^2\rho+c_b^2)}\right\} \tag{3.6.5}$$

$$\text{and} \quad L_e = \frac{\rho}{1-\hat{\rho}} \approx \frac{\rho(c_a^2\rho+c_b^2)}{2(1-\rho)} \qquad (\text{as } \rho \to 1) \tag{3.6.6}$$

[33] We grace p_0 with the style 'moment' here, which is legitimate if the corresponding multiplier of $\{p_i\}$ is taken to be 'no queue'=1 and 'any queue'=0.

[34] In equation (8.2.21), p389, $(\hat{\rho})$ is raised (incorrectly) to the power of n, then in equation (8.2.24b), p390, $(\hat{\rho})$ is raised (correctly) to the power of n-1.

The result is not exact, as can be seen if M/M/1 statistics are inserted in (3.6.5) giving $\hat{\rho} = \exp\{-(1-\rho)\}$ rather than $\hat{\rho} = \rho$. However, the approximation is known to be most accurate near ρ=1 ('heavy traffic') where these two functions converge. Given its approximate nature, including the fact that statistical parameters are themselves only an approximation to a real process, a practical approach to making this time-dependent might be to set $\hat{\rho}$ according to the P-K formula, then replace ρ in its expression by a function of time.

3.6.2 Doubly-nested geometric distributions

If a third 'effective ρ' is introduced, the simplest doubly-nested distribution is:

$$p_{0e} = 1-\rho^* \qquad p_1 = \rho^*(1-\hat{\rho}) \qquad p_i = \rho^*\hat{\rho}(1-\vec{\rho})\vec{\rho}^{\,i-2} \;\; (i\geq 2) \tag{3.6.7}$$

whose first and second moments are:

$$L_e = \frac{\rho^*(1+\hat{\rho}-\vec{\rho})}{1-\vec{\rho}} \qquad V_e + L_e^2 = \frac{\rho^*(1+3\hat{\rho}-\vec{\rho}(2+\hat{\rho}-\vec{\rho}))}{(1-\vec{\rho})^2} \tag{3.6.8}$$

Note that in this case ρ^* need not be the same as the actual demand intensity ρ, and accordingly p_{0e} = 1-ρ is not assumed, widening the range of processes that can be handled. In principle, all three moments can now be fitted. After some manipulation, in which $\hat{\rho}$ is first expressed in terms of ρ^* and $\vec{\rho}$, and the result of substitution into L_e solved for $\vec{\rho}$, the formulae for the three 'rhos' in terms of known steady-state moments are:

$$\rho^* = 1-p_{0e} \qquad \hat{\rho} = \frac{(V_e + L_e(L_e-1))(1-\vec{\rho})^2}{2\rho^*} \qquad \vec{\rho} = \frac{V_e + L_e(L_e-3)+2\rho^*}{V_e + L_e(L_e-1)} \tag{3.6.9}$$

By setting all the 'rhos' equal, it can be confirmed that the last two equations (3.6.9) are satisfied identically when the model reduces to M/M/1. In more general cases, explicit queue statistics are in principle unnecessary since they are subsumed by the moments. This is confirmed by Figure 3.6.1 where the fit between the true M/D/1 as calculated by equations (2.4.4-7) and the doubly-nested model, while not exact, is very close. M/M/1 distributions have been added for comparison, to emphasise the much greater step between p_0 and p_1 in the M/D/1 distribution.

This approximation method relies on the proposition that local symmetry under linear transformation of origin and appropriate scaling means all equilibrium distributions tend to geometric once sufficiently remote from the origin. The method enables the ultimate geometric ratio to be obtained as $\vec{\rho}$.

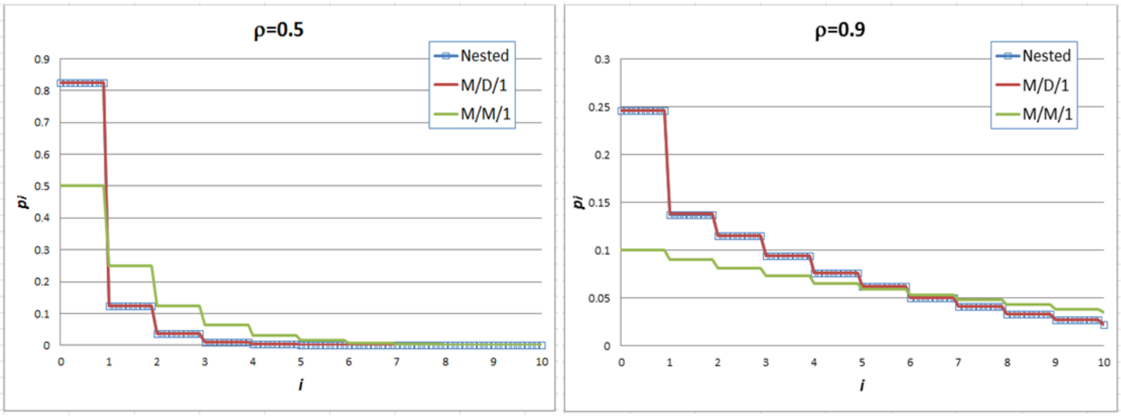


Figure 3.6.1 Fit between doubly-nested model and exact calculated M/D/1 distributions (‘flatter’ distributions are M/M/1 for comparison)

It gets more interesting when the method is applied to the M/D/1[G] distributions described earlier in Section 2.4. However, it is true that G has to be large to make a serious difference. While the fit shown in Figure 3.6.2 is not exact for p_0 and p_1, the relative magnitudes of the probability components are quite well represented. The method can be applied equally with moments calculated using empirical methods such as those to be described in the next Section.

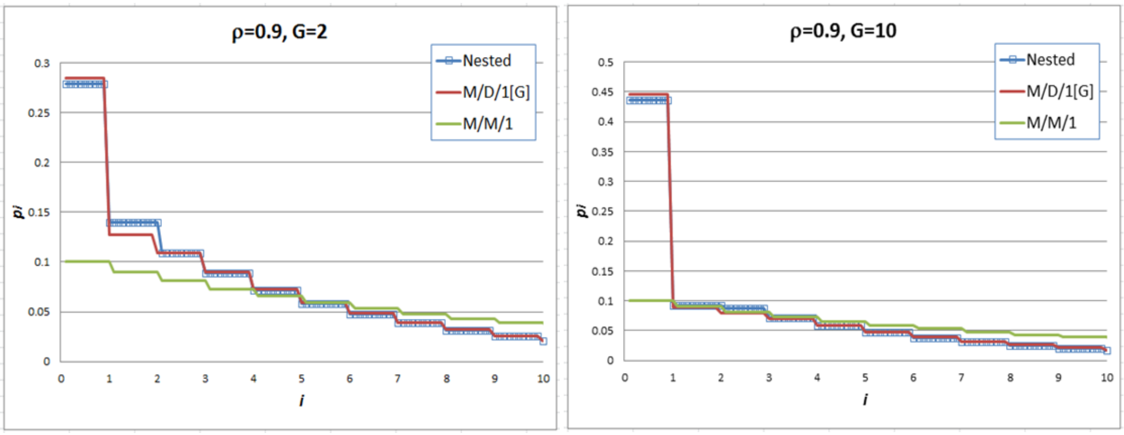


Figure 3.6.2 Fit between doubly-nested model and simulated M/D/1[G] distributions (‘flatter’ distributions are M/M/1 for comparison)

Furthermore, by suitable choice of L_e and V_e it is possible to generate a distribution that is non-monotonic, like those of Erlang arrivals or bulk service, Figures 3.5.4 and 3.5.8. The doubly-nested distribution in Figure 3.6.3 (below) approximates that for the M/M4/1 case.

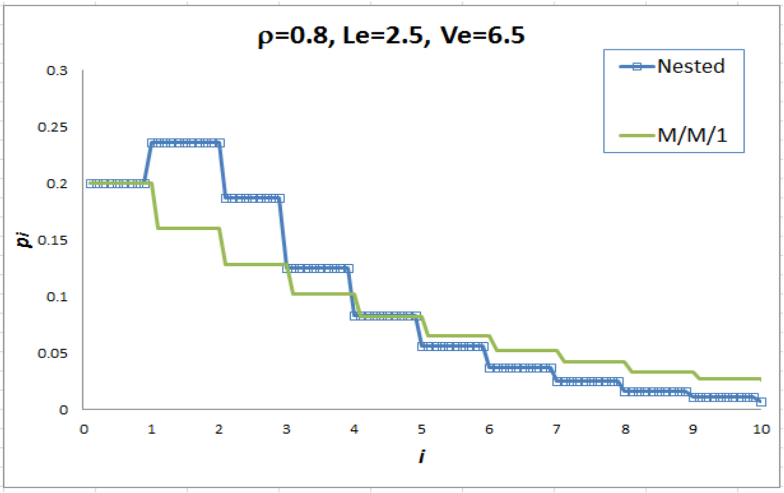


Figure 3.6.3 Doubly-nested model of M/M4/1 'bulk service' case ('flatter' distribution is M/M/1 for comparison)

3.6.4 Maximum Entropy distributions

The singly-nested geometric distribution in equation (3.6.1) is also the Maximum Entropy distribution as stated by Kouvatsos (1988), who expresses the probability distribution as either the exponential of a weighted sum of functions, or an equivalent product of powers of functions. If states are restricted to integer queue sizes, probabilities in terms of constraint functions are given by equation (3.6.10), and entropy can be expressed in terms of expectations of the constraint functions by equation (3.6.11), where the γ_k are Lagrange multipliers:

$$p_i = \exp\{-\gamma_0.1 - \gamma_1.Q - \gamma_2.i\} \qquad (3.6.10)$$

$$H = -\sum_0^{\infty} p_i \ln(p_i) = \gamma_0.1 + \gamma_1.(1 - p_0) + \gamma_2.L_e \qquad (3.6.11)$$

The γ_0 term represents the constraint (prior information) $\Sigma p_i=1$, Q is 0 when i=0 and 1 when i>0 corresponding to a known value of p_0 or utilisation, and the third term corresponds to the mean queue L_e. Equation (3.6.10) expressed as a product reproduces the nested geometric distribution, the Lagrange multipliers being explicit functions of the first three probabilities:

$$p_i = p_0 \cdot \left(\frac{p_1^2}{p_0 p_2}\right)^Q \cdot \left(\frac{p_2}{p_1}\right)^i \qquad (3.6.12)$$

By the same principles, a doubly-nested distribution can be expressed as:

$$p_i = \exp\{-\gamma_0.1 - \gamma_1.Q_1 - \gamma_2.Q_2 - \gamma_3.i\} \qquad (3.6.13)$$

where Q_1 is as Q before, and Q_2 is 0 when $i \leq 1$, and 1 otherwise. Equivalently:

$$p_i = p_0 \cdot \left(\frac{p_1 p_2}{p_0 p_3}\right)^{Q_1} \cdot \left(\frac{p_2^2}{p_1 p_3}\right)^{Q_2} \cdot \left(\frac{p_3}{p_2}\right)^{i} \tag{3.6.14}$$

At first sight this appears not to embody variance as a constraint, because equation (3.6.13) contains no term in i^2. This is misleading because such a term would cause p_i to 'explode'. In fact, equations (3.6.7-9) link the first four probabilities p_0 .. p_3 to the now four constraints. But is this still the Maximum Entropy distribution given the constraints? Kouvatsos' analysis places no limit on the number of constraints and requires only that probabilities have a form similar to (3.6.10/13), which ultimately forces the distribution to develop geometrically.

3.6.4 Limitations of nested distributions

An obvious limitation of nested distributions is that they are primarily usable only for equilibrium distributions, but when does a distribution get near enough to equilibrium to qualify? Equations (3.6.7) will work only if all the 'rhos' are ≤ 1. This is necessarily true for ρ^* but for $\vec{\rho}$, from the third of equations (3.6.9), the condition is $L_e \geq \rho^*$. It is certainly not a problem for M/M/1 and does not appear to cause a problem generally. However, the second of equations (3.6.9) can predict $\hat{\rho} > 1$, contradicting the equilibrium hypothesis. Rho values from the J2P4 peak case are plotted in Figure 3.6.4, showing that in Ts 3-9, the probability distributions cannot be matched by a doubly-nested geometric. This is taken to reflect their being far from equilibrium.

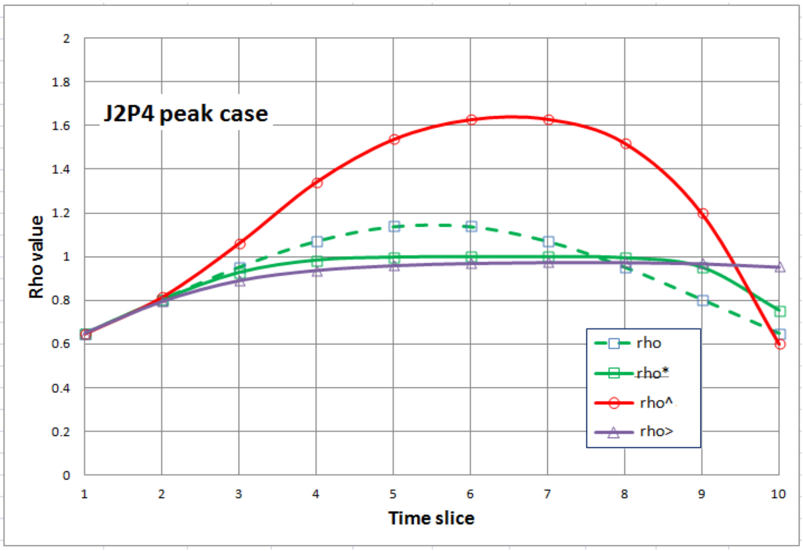


Figure 3.6.4 Doubly-nested parameters for time slices in J2P4 peak case

A further limitation is revealed by the M/M/*n* queue - see earlier in section 3.5.4. Theoretically the doubly-nested approximation cannot cope with any equilibrium distribution that peaks at *i*>2. Since the first few distributions peak at *i*=*n*-1, it should work up to *n*=3. In practice the estimated parameters are *feasible* up to *n*=4, as shown by Figure 3.6.5. However, the *shape* of the distributions is reproduced only up to *n*=3 as shown by Figure 3.6.6 (compare Figure 3.5.1). In principle *n*=4 could be approximated with *four* parameters, but these cannot be determined uniquely by only three moments (note also that equilibrium variances were got from calculated distributions, not explicit formulae which are unknown in this case).

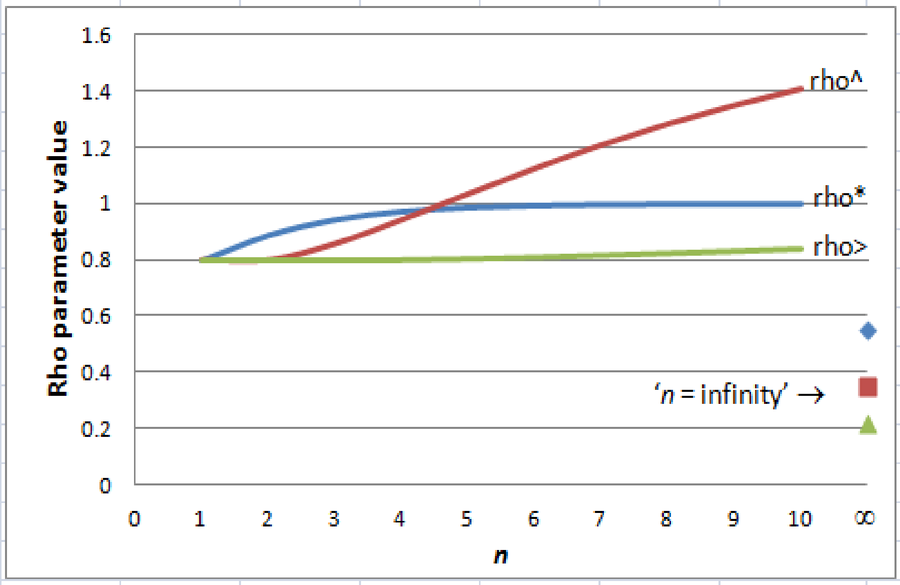


Figure 3.6.5 Doubly-nested parameters for M/M/*n* queue with ρ = 0.8

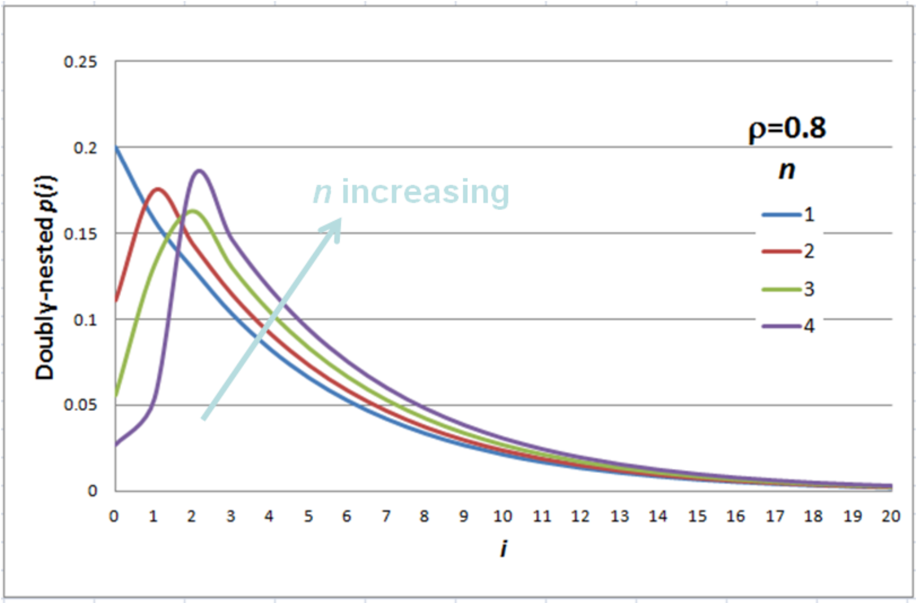


Figure 3.6.6 Doubly-nested approximations to M/M/*n* queue distributions for ρ = 0.8

3.7 EMPIRICAL METHODS TO ACCOUNT FOR SIGNAL GREEN CAPACITY

A summary of the contents of this Section was presented at the UTSG 2013 conference (Taylor 2013) and has been submitted for journal publication as Taylor and Heydecker (2013).

3.7.1 Historical approaches to signal queue modelling

The extended M/D/1[G] processes defined earlier in Section 2.4, that take account of signal green capacity, are difficult to analyse and exact closed-form moments are not available. Several authors since the 1960s have developed empirical approximations to the equilibrium queue. This Section reviews their work and develops an alternative approach to estimating all three moments: p_0, mean and variance; enabling probability distributions also to be estimated.

Analysis of signal queues can be traced back to A J H Clayton in 1940 (see Allsop / Hutchinson 1972), but signal queue models are often associated with F V Webster (1958) and Webster and Cobbe (1966) who developed a delay or equivalent queue formula (3.7.1). This contains a deterministic red phase or so-called ‘uniform’ queue, representing the average over the signal cycle of the queue produced by the red phase including its decay during the green phase, a stochastic term representing the average effect of oversaturation and random variations leading to transient excess of arrivals over capacity, leaving a queue at the end of the green phase, as sketched in Figure 3.7.1, and a third correction term. Here, the green/cycle ratio is represented by Λ to avoid confusion with arrival rate, c is cycle time.

$$L = L_P + L_V + L_W = \frac{x\mu c(1-\Lambda)^2}{2(1-x\Lambda)} + \frac{x^2}{2(1-x)} - 0.65(x\mu c)^{\frac{1}{3}} x^{(2+5\Lambda)} \qquad (3.7.1)$$

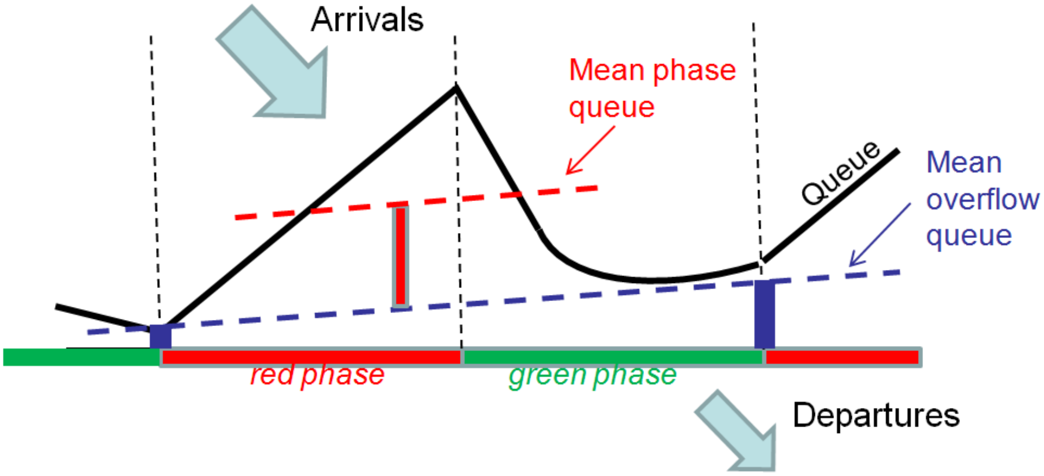


Figure 3.7.1 Sketch of a signal queue, showing phase and stochastic components

If demand exceeds capacity, (3.7.1) is adjusted by replacing the phase component by a simplified term obtained by limiting *x* to 1. Since the signal is at the stop line and any overflow is taken up by the last two terms in (3.7.1), the stochastic process should be treated as 'outside' the phase queue process, so technically should be subject to the demand intensity. If this exceeds 1, the stochastic term in (3.7.1) becomes undefined so should be replaced by a time-dependent formula. The inconsistency of (3.7.1) with time-dependence raises practical issues, which are alluded to by Taylor (2003). However, the concern in this research is only with the stochastic component.

Because the service at a signal is quite regular, even if intermittent, it is idealised as Deterministic, and the cyclic character is considered absorbed by the other components. The practical accuracy of the method, if not compensated for by the correction term in equation (3.7.1), can be improved by modifying the randomness parameter in the M/D/1 queue, which is often given a value in the range 0.5-0.6, rather than exactly 0.5 (Burrow 1987).

It is not known how fully these details are addressed in macroscopic traffic modelling. In the CONTRAM time-dependent assignment program (Taylor 1990, 2003), only the stochastic component is made time-dependent, and an *ad hoc* adjustment is used to 'relax' the phase queue component over a short time if its value falls between time slices - if its value rises it is simply stepped up. Bin Han (1996) combines phase and stochastic queues in a single continuously differentiable sheared formula, but there may still be issues because of the different timescales over which these components operate and their real discontinuities.

A comprehensive time-dependent formulation of the signal queue has been derived by Heidemann (1994), drawing on results by Meissl (1963). At equilibrium results are virtually indistinguishable from Webster and Cobbe's. Heidemann's formulation includes an ostensibly exact formula for the stochastic queue, requiring evaluation of complex roots of a function (see Appendix D). At equilibrium, however, it can be expressed in the same form as (3.7.1) with an identical phase component and a stochastic component now containing *G* explicitly, plus a new correction term that involves the stochastic queue but does not contain any empirical constants:

$$L = L_P + L_{V[G]} + L_H \qquad \text{where} \quad L_H = \frac{-\Lambda(1-x)\left(2L_{V[G]} + \Lambda x\right) + \Lambda x}{2(1-\Lambda x)} \qquad (3.7.2)$$

Olszewski (1990) has shown by simulation that the stochastic queue decreases as absolute green period capacity is increased. This cannot be accommodated by the M/D/1 formula.

Akçelik (1980, 1998a) has developed a time-dependent saturating solution resembling the sheared solution, which incorporates *G* in the form of *gs* in a parameter modifying the effective demand intensity. This is not considered further here as it is not clear how it can be incorporated in the present scheme, but several alternative approaches are considered in Appendix E. Empirical approximations to account for the effect of green period capacity on the mean stochastic equilibrium queue have been proposed by Newell (1960), Miller (1969) and Cronjé (1983a), although Miller considers Newell's model too complex for practical use. The rest of this Section considers these, and then proposes alternative approximations to the queue moments based on simulations of the M/D/1[G] process described in Section 2.4.

3.7.2 Results of simulations with a range of parameters

Using the results of Markov simulations based on M/D/1[G] recurrence relations (2.4.10), with maximum *i* of 10,000 for accuracy, equilibrium mean queues have been calculated for a range of values of ρ values, and green capacities *G* up to 100. These are plotted in Figure 3.7.2, confirming not only the behaviour observed by Olszewski (1990) and others, but also suggesting strongly that the queue tends to zero as the green period gets indefinitely long. This can be considered the deterministic limit, where only oversaturation results in queuing.

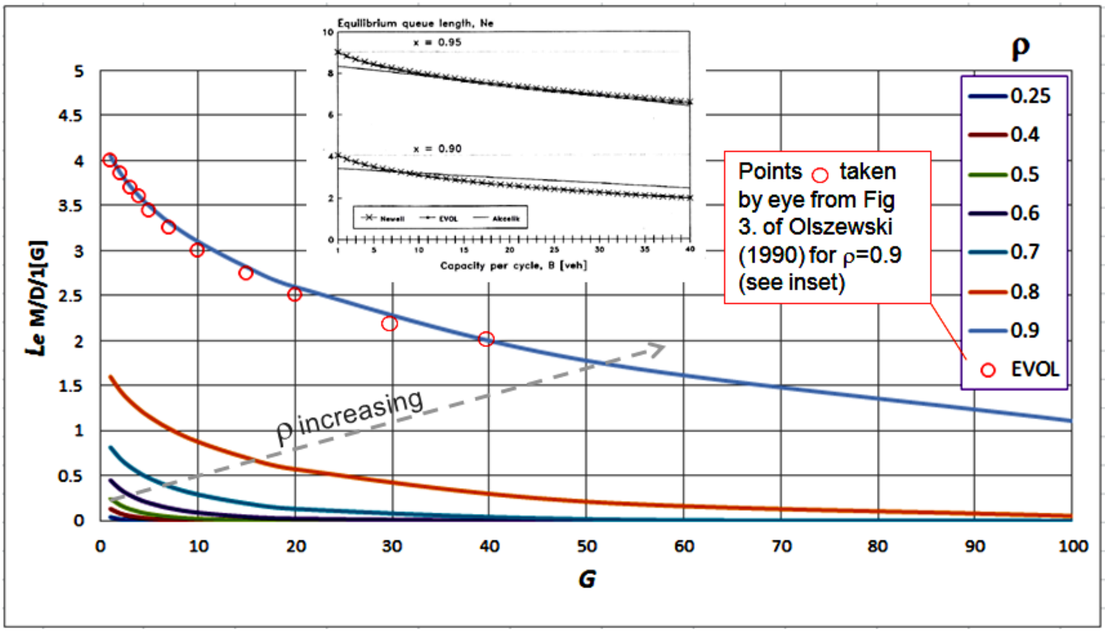


Figure 3.7.2 Dependence of M/D/1[G] equilibrium mean queue on green period *G*. The original Figure 3 of Olszewski (1990) is inset as evidence of the points marked.

The process is not absolutely independent of capacity, saturation flow, cycle time, etc, because these variables are related to ρ and *G* as in equations (3.7.3):

$$G = gs = c\mu \qquad \text{hence} \qquad \lambda = \rho\mu = \frac{\rho gs}{c} = \frac{\rho G}{c} \qquad (3.7.3)$$

There are four truly independent variables λ, g, c, s, that can be substituted for by ρ, G, c, μ through (3.7.3). Capacity μ can be ignored because equilibrium properties are independent of it. However, in order to maintain the capacity (rate), the ratio g/c must then stay constant, so the idealised M/D/1[G] processes are explicitly dependent only on ρ, G. It is as though time stops at the end of one green period and restarts at the beginning of the next, but is stretched in proportion so that its 'clock' keeps pace with reality. During the green periods, arrivals are assumed to have a headway distribution consistent with the demand intensity ρ. A *physical* explanation for the reduction in queue size with increasing G is that, depending on whether a queue is present at the moment a customer arrives within the green period, the arrival may effectively 'disappear' because it takes advantage of spare capacity that would otherwise be unused. This is not 'censoring' because it does not happen according to a rule, but it could lead to both a reduction in the effective arrival rate at service and apparent the bunching of arrivals later in the green period. However, as will be shown later, this effect could not be accommodated by adjusting the arrivals dispersion I_a in the P-K formula.

3.7.3 Behaviour and estimation of p_0

In Figure 3.7.3, graphs of simulated p_{0e} against G for different values of ρ appear to have similar shapes on different scales.

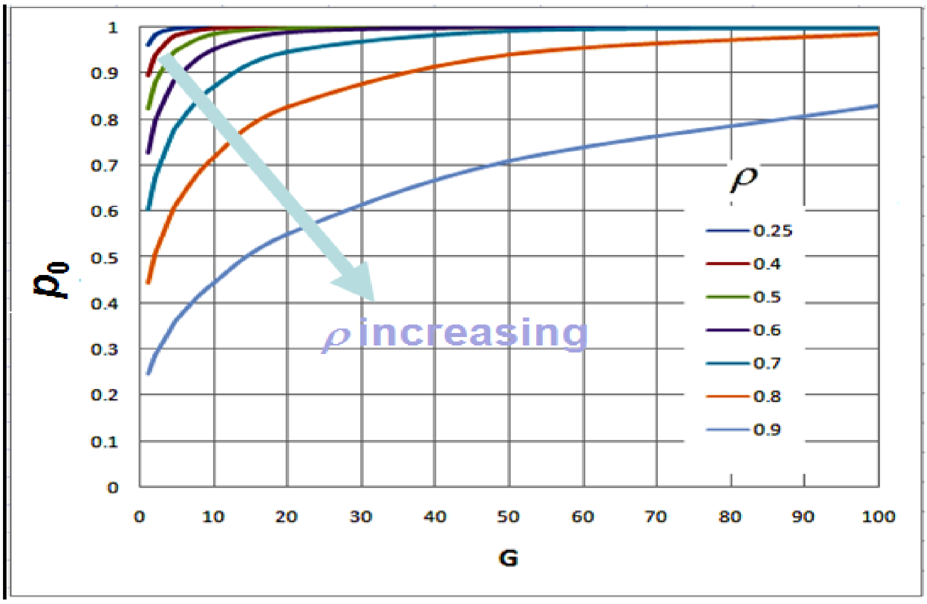


Figure 3.7.3 Markov simulated equilibrium p_{0e} for a range of M/D/1[G] processes

Experimentation with simple transformations shows that graphs of p_0 follow a common trajectory very closely when plotted against a dimensionless 'link function' z, defined generally by equation (3.7.4) (in what follows, the dependence on ρ and G is understood):

$$z(h) = \frac{G+h}{\tau_{re1}} \quad \text{where } \tau_{re1} = \left(1-\sqrt{\rho}\right)^{-2} \text{ and } h\text{=2 in this case}^{35} \qquad (3.7.4)$$

τ_{re1} is the stochastic relaxation time (2.3.38) for each ρ with μ set to 1. This result is evident in Figure 3.7.4, where left end of the LH plot is expanded in the RH plot. The idea of the 'link function' is that combinations of basic variables that give the same link function value lead to the same results in the target function. This reflects a 'symmetry' between ρ and G.

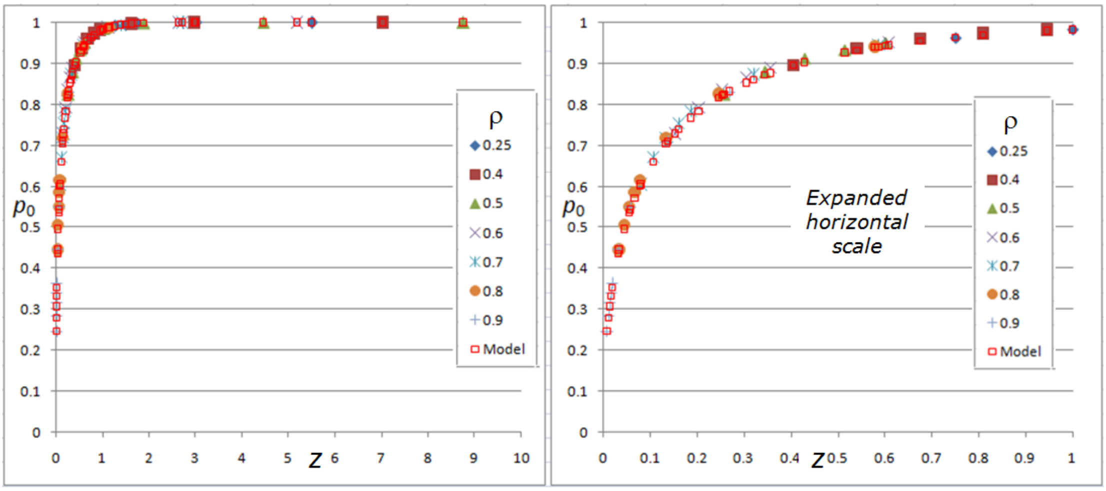


Figure 3.7.4 Markov-simulated p_0 against the function $z(2)$ (expanded scale at right)

Points for G=1 satisfy (2.4.7), and since the points for the other values of G lie on the same curve, interspersed with those of G=1, they can be calculated using the same formula together with an 'effective ρ', η_0, that needs to satisfy:

$$3\left(1-\sqrt{\eta_0}\right)^2 = (G+2)\left(1-\sqrt{\rho}\right)^2 \qquad \text{or}$$

$$\eta_0 = \max\left[\left(1-\sqrt{\frac{G+2}{3}}\max\left(1-\sqrt{\rho},0\right)\right)^2,0\right] \qquad (3.7.5)$$

The lower limit of zero reflects the vanishing (sub-resolution) of the Markov simulated mean for small ρ and large G. Then p_0 can be estimated for all G, as required, by:

[35] This can be written in terms of signal cycle time c as $z = \dfrac{c+\hat{h}}{\tau_{re}}$ where $\hat{h}$ a modified adjustment.

$$p_{0[G]est} \cong e^{\eta 0}(1-\eta_0) \tag{3.7.6}$$

In Figure 3.7.4 estimated values are indicated by red squares. The estimates are not perfect but very close, with RMS error = 0.0077 or 0.93% of average. Estimated and Markov simulated values are compared directly in Figure 3.7.5.

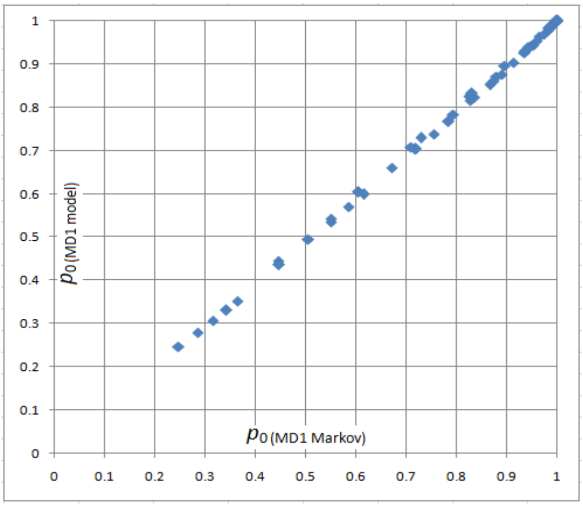


Figure 3.7.5 Performance of M/D/1[G] p_0 estimates compared to simulation

3.7.4 Attempts at direct approximation to mean queue size

Graphs of the Markov simulated M/D/1[G] queue sizes in Figure 3.7.2 also have a similar shape, suggesting that a link function may exist for them but, unlike p_0, they are not confined within the bounds of [0,1], so the treatment has to be somewhat different. This is broadly the approach used by previous authors. Miller's formula for the stochastic queue (with modified notation as used throughout this Section) is:

$$L_{e[G]M} = \frac{\exp\left(-\frac{4y}{3\rho}\right)}{2(1-\rho)} \quad \text{where} \quad y = (1-\rho)\sqrt{G} \tag{3.7.7}$$

Cronjé (1983a) offers, without further explanation, a "suggested modification to Newell" where an explicit expression involving the variable y replaces a factor in Newell's formula. Dispersion of arrivals I_a is also a factor, although in what follows it is dropped since this use would be incompatible with its role in the extended P-K formula, i.e. I_a=1 is assumed:

$$L_{e[G]C} = \frac{I_a \rho \exp\left(-y - \frac{1}{2}y^2\right)}{2(1-\rho)} \quad \text{where } y \text{ is as above} \tag{3.7.8}$$

Mean queue values fall approximately onto a common trajectory, as shown by Figure 3.7.6, if the link function (3.7.4) is chosen as:

$$z = z(-1.4) \text{ optimally,} \qquad \text{although } z = z(-1) \text{ works} \qquad (3.7.9)$$

Unlike for p_0, an ‘effective ρ’ does not arise so immediately in this case, since to achieve this fit the mean queue sizes have first to be normalised to the L_e values for the ‘standard’ M/D/1 values with G=1. Functions can however be fitted to the data, and then transformed back into an empirical approximation to L_e in terms of ρ and G of the form:

$$L_{e[G]est}(\rho, G) = \Re_{[G]}(z).L_{e[1]} \qquad (3.7.10)$$

Two forms of $\Re_{[G]}(z)$, ‘power’ and ‘exponential’, have been tried:

$$\Re_{[G]pwr}(z) = \min\left(\left(1 - \sqrt{0.9z - 1}\right)^2, 1\right) \qquad (3.7.11a)$$

$$\Re_{[G]\exp}(z) = \min\left(\exp\left(-3.5z^{0.7}\right), 1\right) \qquad (3.7.11b)$$

Figure 3.7.6 also plots the estimated values, and the average absolute and percentage accuracies of the methods are given in Table 3.7.1. Miller’s formula underestimates seriously, and the Cronjé-Newell formula appears to overestimate, although inspection of numerical results shows it is the most accurate for high values of ρ≥0.7. Of the functions in (3.7.11), the power function gives lower RMS error, around 0.025 over the data range of 0-1, compared to around 0.03 for the exponential function, but the exponential gives lower percentage error.

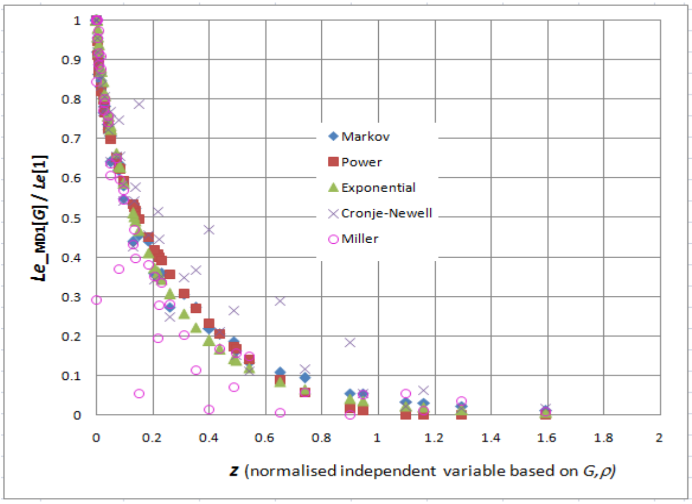


Figure 3.7.6 Relationship between normalised L_e values and the link function z

Table 3.7.1 Accuracies of the M/D/1[G] queue size models (limited to 1 x $L_{e[1]}$)

Methods	**Absolute RMS error**		**Average percent error**	
	All ρ	ρ≥0.7	All ρ	ρ≥0.7
Power	0.025	0.034	18.3	15.5
Exponential	0.030	0.039	10.7	9.7
Cronjé-Newell	0.067	0.009	39.0	4.8
Miller	0.124	0.027	31.5	26.4

Exponential functions have the advantage that their values do not fall off too rapidly at higher values of *G* and so can achieve better approximations. As is often the case, the Miller and Cronjé-Newell formulae are optimised for ‘heavy traffic’. Their problem is that the exponential terms, that ought to reduce to ρ when *G*=1, do so approximately only for large ρ. There appears to be no common ground between the methods that could be exploited to produce a better one. However, the good performance of the Cronjé-Newell formula at higher values of ρ makes it attractive as the basis of a method adjusted to give better results for smaller ρ values. This can be tested by calculating the error relative to Markov simulations in the factor:

$$f_C(\rho, G) = \frac{\exp\left(-y - \tfrac{1}{2}y^2\right)}{\rho} \qquad \text{where} \quad y = (1-\rho)\sqrt{G} \qquad (3.7.12)$$

This is (3.7.8) normalised to act as a multiplier of the M/D/1 mean queue (2.4.7). Replacing the contents of the bracket with a function of the form (1-exp(*y*)) does not avoid error. Figure 3.7.7 shows that the logarithm of the factor error is a fairly linear function of *G*, but its slope is extremely sensitive to ρ, so sensitive that no function of ρ normally met could describe it.

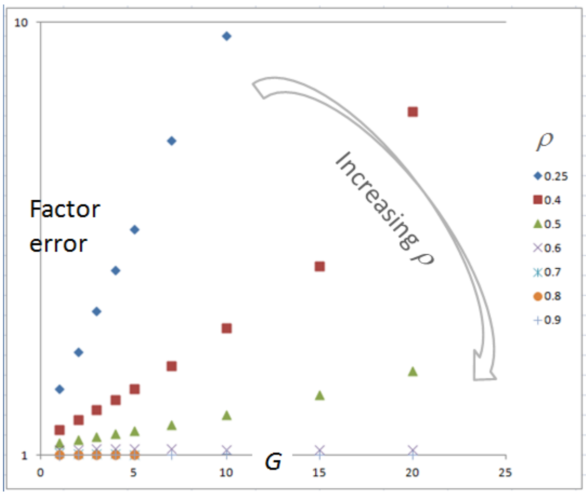


Figure 3.7.7 Cronjé log factor error dependence on ρ and *G*

Figure 3.7.8 shows that the slope of these factors has a ‘hockey stick’ shape in relation to ρ. It is not obvious whether the slightly negative values for ρ≥0.7 are real or the result of imprecision in the simulations. An alternative interpretation is that the slope falls linearly becoming zero somewhere between ρ=0.5 and 0.6. On that assumption, admitting that there are only three data points, the slopes themselves can be fitted by a linear relationship whose coefficients are slightly rounded from the actual regression values:

$$s = 0.4 - 0.75\rho \qquad (3.7.13)$$

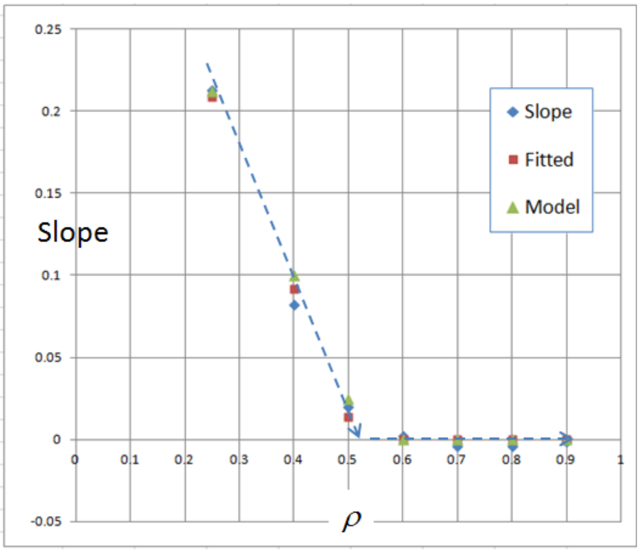


Figure 3.7.8 Dependence on ρ of average slopes in Figure 3.7.7

Figure 3.7.8 also plots the fitted and equation (3.7.12) values, that are assumed to be zero for ρ≥0.6. It is possible that the true relationship bends smoothly and goes to zero near where the value of L_e is 0.5 for G=1 (ρ≈0.6185). However, this affects so narrow a range of ρ, at factor error so close to 1, that it probably isn't worth pursuing. Consequently, the following composite formula is arrived at:

$$\tilde{L}_{e[G]} = \frac{\rho^2}{2(1-\rho)} \min\left[\exp(\max(0.4 - 0.75\rho, 0)G) \frac{\exp\left(-y - \frac{1}{2}y^2\right)}{\rho}, 1 \right] \qquad (3.7.14)$$

where $y = (1-\rho)\sqrt{G}$ as before, and the minimum is needed to overcome overprediction by the Cronjé-Newell function when G=1. Figure 3.7.9 shows that this adjusted Cronjé-Newell formula fits all the simulated data substantially better than the original. It is disappointing that an empirical and non-smooth function of ρ is involved, but it is not obvious what common function could reproduce the extreme sensitivity to smaller values of ρ in Figure 3.7.8.

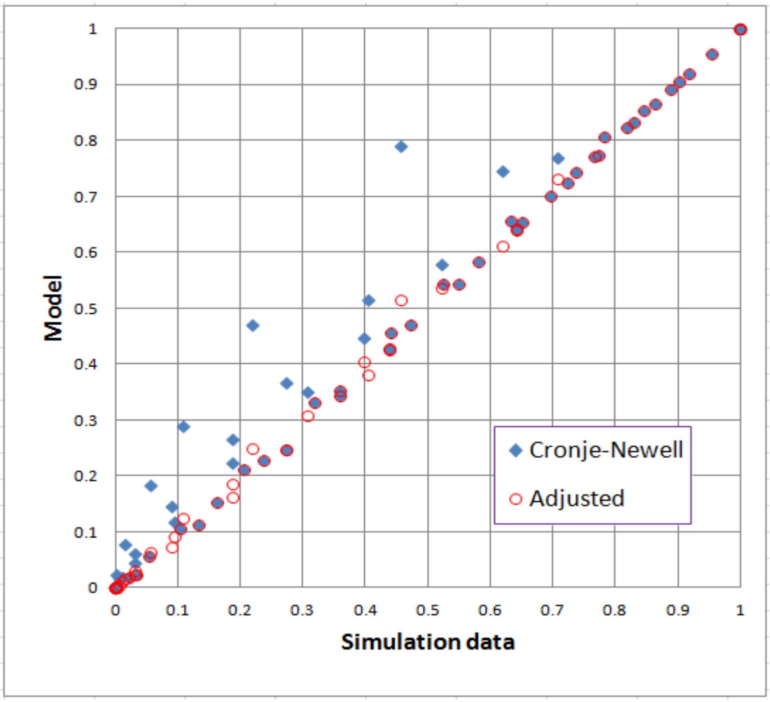


Figure 3.7.9 Performance of normalised Cronjé-Newell and adjusted queue models

3.7.5 Estimating mean queue using effective traffic intensity

An alternative to such *ad hoc* adjustment is to define an effective traffic intensity or 'effective ρ' for M/D/1[G] equilibrium queues by inverting equation (2.4.8), viz:

$$\eta_1 = \sqrt{L_{e[G]}\left(L_{e[G]} + 2\right)} - L_{e[G]} \qquad (3.7.15)$$

Earlier results including from queues with Erlang arrivals suggest that the probability distribution of any stochastic queue tends to become geometric as queue size increases, but no particular distribution has been assumed in defining (3.7.15), so its physical meaning is unclear.

However, Figure 3.7.10 reveals that a common trajectory exists for the ratio η_1/ρ when plotted against the link function $z(0)$, with the ratio limited to 1, although some points for lower values of ρ are less well fitted, leading to the 'effective ρ' formulation (3.7.16):

$$L_{e[G]} \cong \frac{\eta_1^2}{2\left(1 - \eta_1\right)} \qquad \text{where} \quad \eta_1 = \rho \min\left[\exp\left(\frac{-G}{\tau_{re1}}\right), 1\right] \qquad (3.7.16)$$

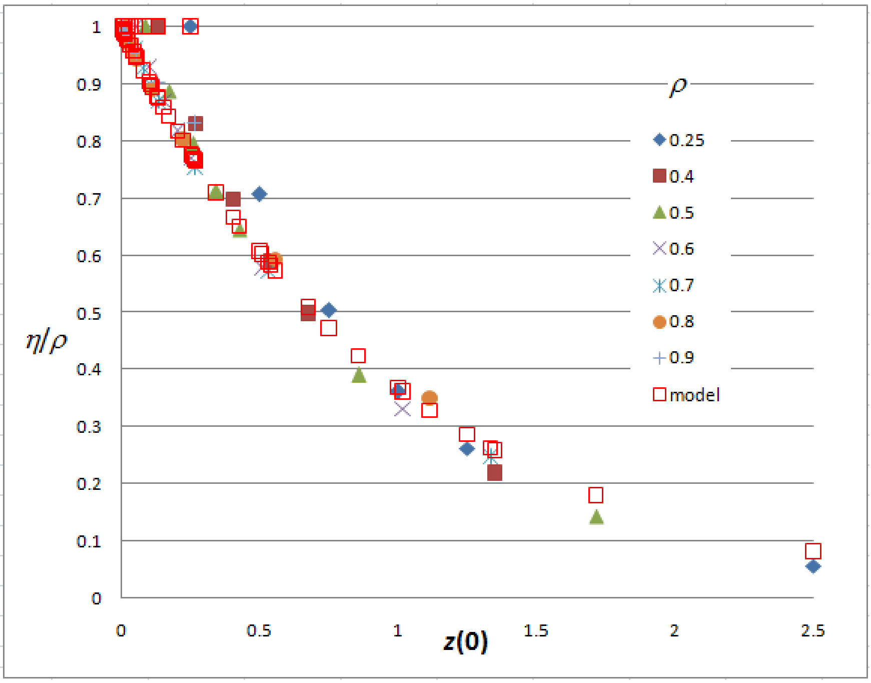


Figure 3.7.10 Best fit of η_1/ρ as function of z for estimating $L_{e[G]}$

$L_{e[G]}$ is estimated directly by the equivalent of (2.4.7). Table 3.7.2 compares the accuracies of all the models. This shows accuracy is fair, with RMS error of 0.024 in η_1 and 0.060 in $L_{e[G]}$, similar to Cronjé-Newell, but no extra constants are needed, indeed the accuracy deteriorates rapidly if additional factors or powers of the z function are introduced.

Table 3.7.2 Accuracy of the M/D/1[G] queue size estimates ($\geq$ 1 x $L_{e[1]}$)

Model	**Absolute RMS error**		**Average percent error**	
	All ρ	ρ≥0.7	All ρ	ρ≥0.7
Cronjé-Newell	0.067	0.009	39.0	4.8
Adjusted C-N	0.014	0.009	5.1	4.8
Effective Rho	0.060	0.091	14.4	5.9

The fit of the ratio η_1/ρ is also shown by the red squares in Figure 3.7.10. The link function fails for G=1, and the ratio has to be forced to 1 in that case, though that presents no practical problem for computations since G is assumed to be discrete. Although this method results in higher errors than the adjusted Cronjé-Newell method, it may be preferred for its structural transparency and computational simplicity where the highest accuracy is not essential. Its accuracy is best for smaller values of G and higher values of ρ, and it avoids Cronjé-Newell's extreme departures at smaller ρ values.

3.7.6 Estimating variance

An approximation to equilibrium M/D/1[G] variance arises directly from the observation that variance values lie close to a common trajectory when normalised to $V_e[1]$ using the exponential function (3.7.17) and link function (3.7.18). This is shown in Figure 3.7.11, where compliance with the trajectory appears, somewhat surprisingly, better than for $L_{e[G]}$.

$$V_{e[G]} \approx V_{e[1]} \min[\exp(-3z(1)),1] \quad \text{where} \qquad (3.7.17)$$

$$z(1) = \frac{G+1}{\tau_{re1}} \qquad \text{(from (3.7.4)} \qquad (3.7.18)$$

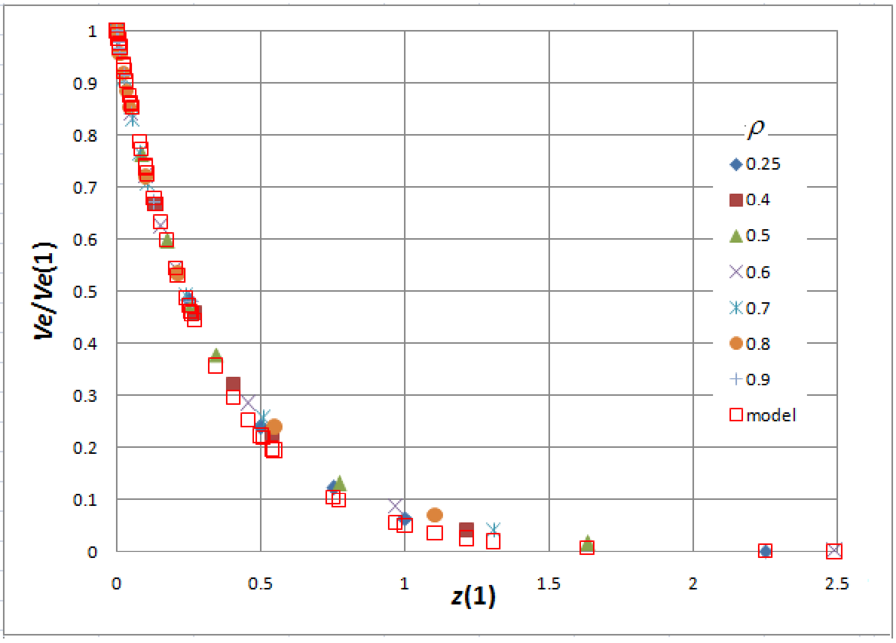


Figure 3.7.11 Exponential approximation to$V_{e[G]}/V_{e[1]}$ using link parameter ζ = -3

Absolute error is minimised at 0.125 by the factor -3 within the exponential function (3.7.17), with percentage error 11.1% and a tendency to underpredict smaller values. However, percentage error is minimised at 7.07% by a factor around -2.7, with absolute error 0.178 but greater overprediction of larger values. Large values are best fitted by a factor -4. Therefore the form of the link function is necessarily a compromise. Estimating variance values in this way is less satisfactory than defining an 'effective ρ'. There may possibly be a way this could be achieved, although the complexity of the expression for $V_{e[1]}$ rather argues against it. However, it is not critical since only the *value* of equilibrium variance is required as an asymptotic constraint. Variance does not come from a time-dependent solution like the sheared queue.

3.7.7 Accuracy of the results and discussion of the methods

Figure 3.7.12 shows that the approximations described in this Section give good results over the wide range of parameter values tested. Figure 3.7.13 compares estimated with simulated moments for extended M/D/1[G] distributions, i.e. where each origin is shifted by *G*.

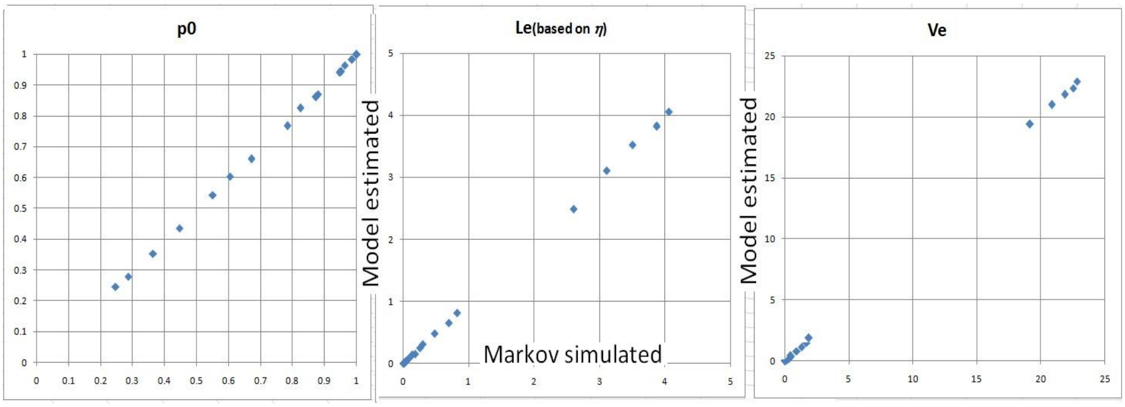


Figure 3.7.12 Results of M/D/1 estimations for 25 combinations of ρ and *G*

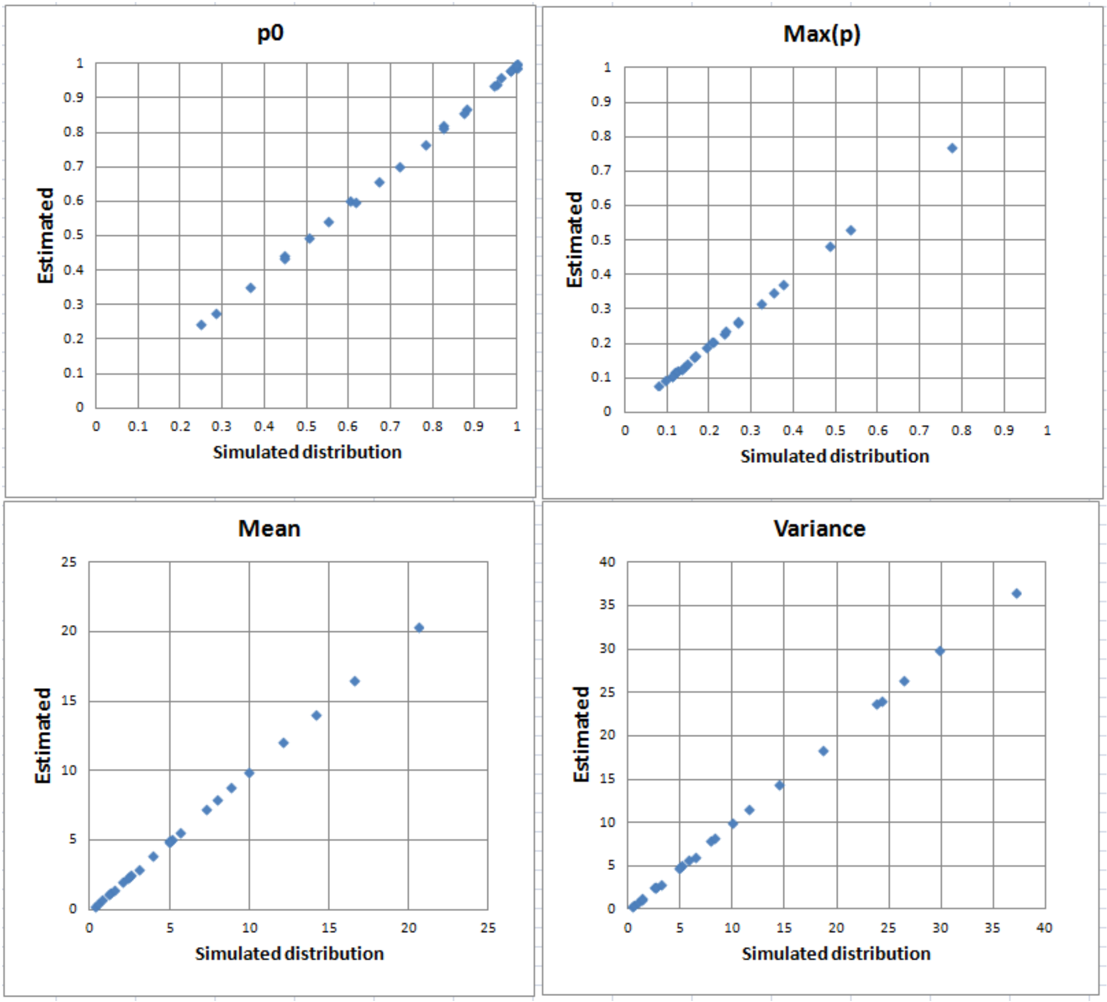


Figure 3.7.13 Fit between estimated and simulated extended M/D/1[G] moments

A common element in the formulae developed by the authors cited is the divisor (1-ρ), also invariably present in the Pollaczek-Khinchin mean formula. This suggests normalisation with respect to L_e[1]. Another is the presence of an expression in ρ and G involving (1-ρ)√G, which suggests a symmetry between ρ and G. The link-functions z(h) attempt to exploit what may be a deeper symmetry betwen G and the characteristic timescale set for each ρ by stochastic relaxation time. The corresponding physical interpretation is that arrivals in green periods of different lengths should on average have the same impact on the final stochastic queue if their z values are the same, although this may be stretching a point where the relaxation time is long compared to the duration of the green period.

3.7.8 Feasibility of estimating M/D/1[G] with modified statistical parameters

It should be asked whether, as an alternative to the preceding, modifying statistical parameters in the P-K mean queue formula alone could account for the effect of green period length. Structurally, three possible factors could contribute:

- Reduction in the effective arrival rate
- Bunching of arrivals
- Alteration of the service statistics

The preceding results show that adjusting the effective traffic intensity, which in practice means the effective arrival rate, can lead to useful approximations. Service is already assumed to be uniform, c_b=0, and the derivation of (3.2.13) allows no role for c_a, so any effect of bunching must act through I_a. Figure 3.7.14 shows how I_a has to vary to reproduce $L_{e[G]}$ as estimated.

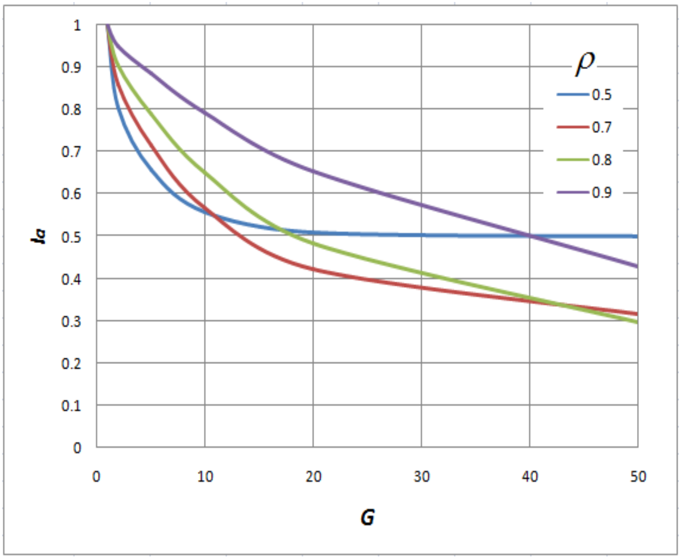


Figure 3.7.14 Arrivals dispersion I_a needed to give estimated $L_{e[G]}$

In Figure 3.7.15, I_a has been replaced by the function below and plotted against $z(0)$:

$$f(I_a) = 1 + \frac{I_{a(est1)} - 1}{\rho} \qquad z(0) = \frac{G}{\tau_{re1}} \qquad (3.7.19)$$

It is apparent from the form of (3.7.19) that the process of estimating I_a has merely been reversed. Figure 3.7.14 (above) confirms that the relationship between the quantities in (3.7.19) is weak. A rather poor exponential approximation of the 'link function' type is equation (3.7.20), but the scatter in Figure 3.7.15 shows that no function of this type can be satisfactory.

$$I_{a(est2)} = \exp(-3z(0)) = \exp\left(\frac{-3G}{\tau_{re1}}\right) \qquad (3.7.20)$$

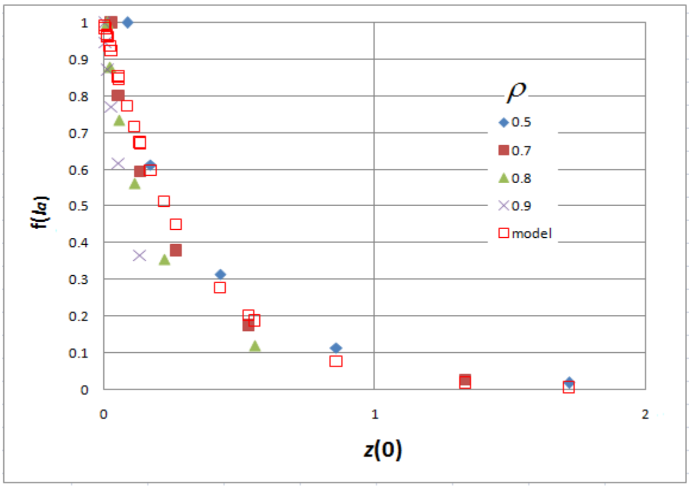


Figure 3.7.15 Function of dispersion I_a needed to give estimated $L_{e[G]}$

3.7.9 Properties of M/D/1[G] probability distributions

Figure 3.7.16 compares simulated probability distributions, extended to include notional state, for ρ=0.8 and five values of *G* (including 1) with Gamma distributions, which are convenient to work with because of their relatively simple continuous functional form[36,37].

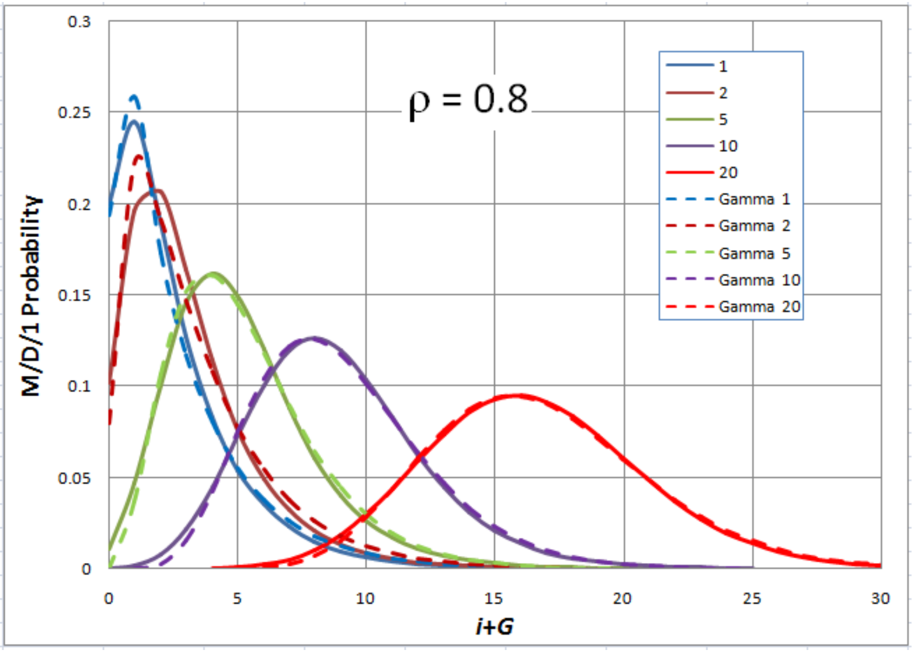


Figure 3.7.16 M/D/1[G] simulated extended distributions for ρ=0.8 and various *G*

Gamma being a continuous distribution, its parameters can be fitted, with varying degrees of accuracy, to simulated distributions using standard numerical methods, as e.g. embodied in Excel's Solver. LogNormal and Poisson distributions have also been tried but found less satisfactory. More detail on this is given later in Chapter 5 where the use of continuous functions to approximate dynamic probability distributions is also addressed.

Some results for the *extended* probability distributions are established. By direct calculation:

$$Mean = \sum_{-G}^{\infty}(i+G)p_i^{(*)} = \sum_{-G}^{0} ip_i^{(*)} + \sum_{0}^{\infty} ip_i + G\sum_{-G}^{\infty} p_i^{(*)} = L_e - L^* + G = L_e + G\rho \qquad (3.7.21)$$

$Mode \approx G\rho$ this being the mean number served in the green period or cycle. (3.7.22)

[36] The Gamma distribution, as a model of queue size probability distributions in over-saturated peaks, has been proposed by Olszewski (1990) and also investigated by Halcrow Fox and Associates under contract to the Transport Research Laboratory (unpub.). Discrete alternatives could be Erlang or Negative Binomial.

[37] The Negative Binomial is used in accident analysis where the accident count at any particular site or juncture is believed to be a Poisson variable, but the variance between sites produced by unspecified factors leads to the combined distribution being over-dispersed. When mean rates at sites are assumed to be Gamma distributed, the combined distribution is Negative Binomial (Heydecker and Wu 2001).

Using a method similar to (3.7.21) the variance of the notional probability terms can be calculated from the Variance of the extended distribution and the 'real' part:

$$V^{*}_{[G]} = Variance - V_{e[G]} - 2G(1-\rho)L_{e[G]} \qquad (3.7.23)$$

The variance of the notional distribution terms alone can now be calculated if an expression for $V^{*}_{[G]}$ can be found. The approach is similar to that used previously. Recalling the limiting value (2.4.25), $V^{*}_{[G]}$ is first assumed to depend on some function $Y(z)$ such that:

$$V^{*}_{[G]} = G\rho(1 - Y(z)\rho) \qquad \text{where} \qquad (3.7.24)$$

$$z \approx z(-1.5) = \frac{G-1.5}{\tau_{re1}} \qquad \text{and} \qquad Y(z) \in [0,1]$$

Figure 3.7.17 confirms that ρ and G can be linked through $Y(z)$ or its complement:

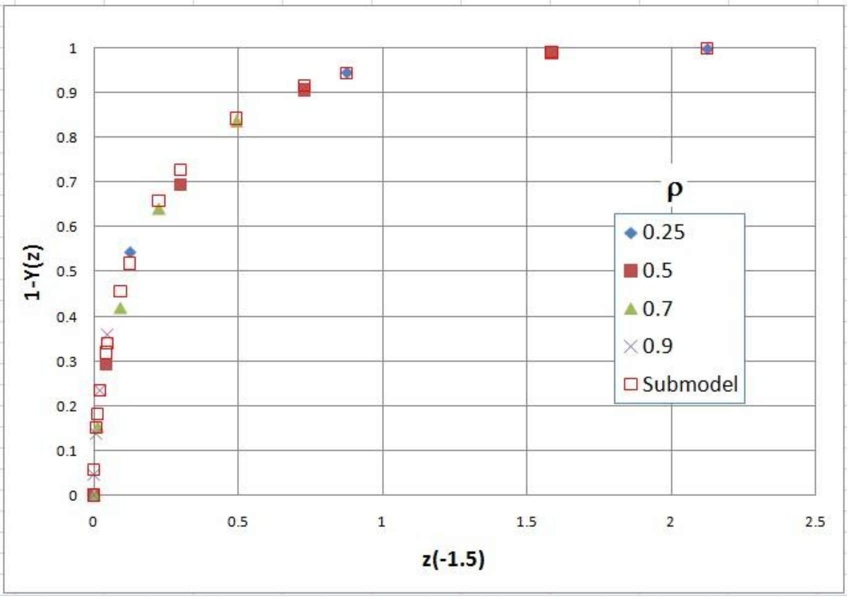


Figure 3.7.17 Common trend component for variances of 'notional' parts of distributions

Taking the logarithm of $Y(z)$ shows it is *nearly* exponential except at small z. (1-$Y(z)$) more resembles a saturating function like L or x in a growing queue. However, for the sake of simplicity an exponential sub-model is attempted. A reasonable fit, with RMSE=0.081 and average percentage error 1.65%, is got with the following approximation:

$$Y(z) \approx 1 - \sqrt{1 - \exp(-2.5z)} \qquad (3.7.25)$$

Figure 3.7.18 graphs ‘real’ probability distributions simulated for ρ=0.8 and several values of *G*, showing how p_0 becomes increasingly dominant.

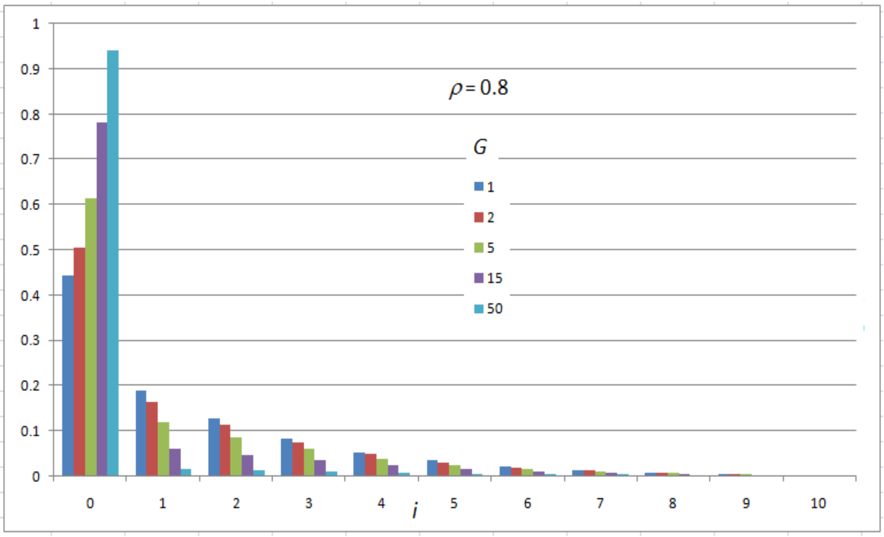


Figure 3.7.18 Simulated M/D/1[G] equilibrium probability distributions

3.7.10 An empirical approach to estimating M/D/1[G] probability distributions

This section includes for completeness initial investigations and some speculation about the distributions, now superseded by work described later in Chapter 6. Referring back to Figure 3.7.18, for sufficiently large *i* the ratios between successive p_i appear to approach a constant ratio, as expected when remote from the influence of the zero boundary. Furthermore, this ratio appears to be independent of *G*, an approximation being:

$$\lim\left(\frac{p_i}{p_{i-1}}\right) \approx R(\rho) = \frac{\rho(1+\rho)^2}{4} \qquad (3.7.26)$$

For *G* up to 5, based on the limiting form (3.7.26), the ratio between successive p_i in the middle range of values of *i* appears to follow:

$$r_i = \frac{p_i^*}{p_{i-1}^*} \approx R(\rho).\left(1 + Ge^{-(G+i-1)\xi}\right) \qquad (3.7.27)$$

where ξ is best set to ρ for larger *G*, and to 1 for smaller *G*, and for *i*<1 this applies to notional probabilities $i \in [-G,0]$, not the real p_0. The ratios of successive probabilities for one level of demand, ρ=0.9 and $i \geq 1$, normalised so that $p_1=1$, are broadly similar for different *G* values, as shown in Figure 3.7.19, with the ratio between terms in their middle ranges given approximately by (3.7.26-27).

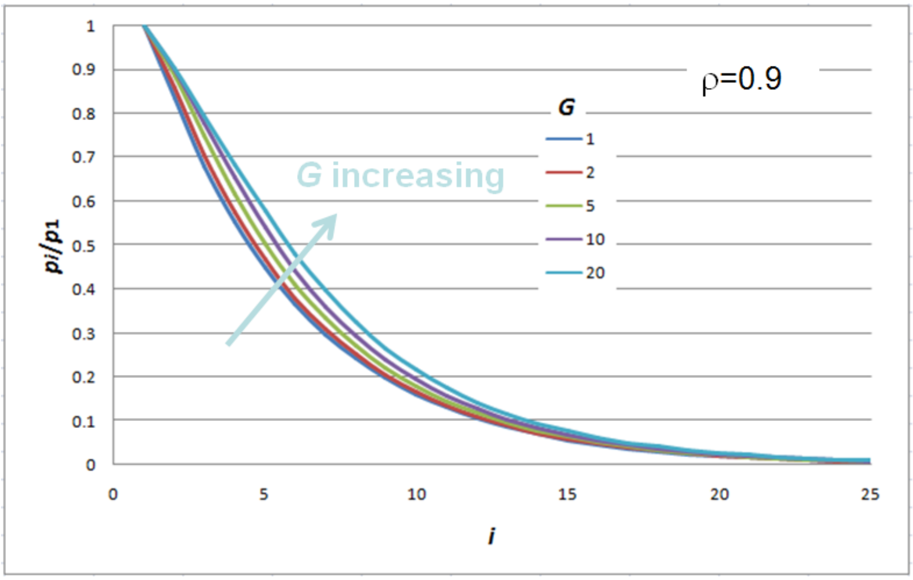


Figure 3.7.19 Normalised M/D/1[G] probability distributions

The fit against simulated values can be good for some values of ρ and G, as shown by Figure 3.7.20 for the particular case G=5. Unfortunately, calculation of the whole distribution is so sensitive to the values of the p_{-i}^{*} that even small errors render this model of little value.

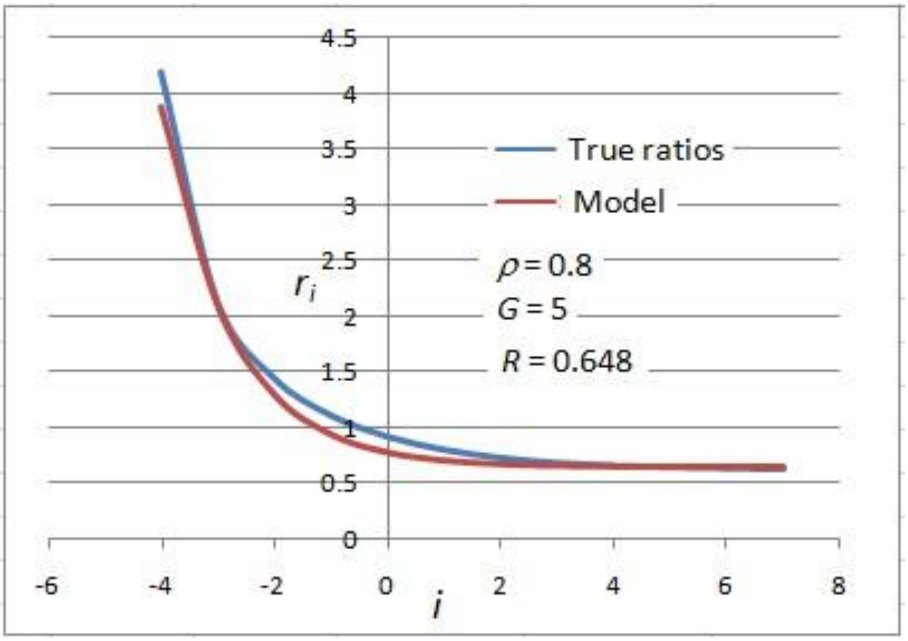


Figure 3.7.20 Simulated and estimated ratios between successive probabilities, G=5

While the extended probability distributions may be fitted best by Gamma distributions, it may be supposed that they are nearly geometric for i sufficiently greater than 0. Can this yield to an ‘effective ρ’ approach? In the steady state, p_0 is the complement of the utilisation, and according to deterministic queuing theory, the arrival rate must balance the utilisation.

Using equation (2.4.6), this appears to imply that, for example, the effective arrival rate for G=1 should be given by:

$$\eta = 1 - p_{0e[1]} = 1 - e^{\rho}(1-\rho) \qquad (3.7.28)$$

For ρ=0.9 this has the value 0.754 rather than the limiting value 0.812 from (3.7.26). While these functions are not close analytically they are not greatly different in absolute terms over the range of $\rho \in [0,1]$. So I_a can now be recalculated using (3.7.28) rather than (3.7.19/20), and $p_{0[G]}$ calculated from (3.7.5-6):

$$I_{a(est3)} = 1 + \frac{2(1-\eta)\left(L_{e[G]} - L_{[1]}(\eta)\right)}{\eta} \qquad (3.7.29)$$

However, this now gives values of I_a>1 in most cases, which is not the expected result. This happens because the η corresponding to p_0 from equations (3.7.5-6) is smaller than that corresponding to L_e from equation (3.7.16).

It seems therefore that this approach is not leading anywhere, but the method of doubly-nested distributions can still be applied.

3.8. CONCLUSIONS ON MORE GENERAL EQUILIBRIUM QUEUES

In this Chapter 3 the results of Chapter 2 have been extended to more general arrival and service statistics, using an event-based formalism found in standard reference works, to produce generalised formulae for the equilibrium mean and variance of queues. Equilibrium formulae for a number of specific queue processes are derived from recurrence relations, and are matched to statistical parameters of the Pollaczek-Khinchin mean formula using Markov simulations, enabling them to be brought into the time-dependent approximation framework that will be addressed later in Chapter 4.

The unit-in-service component in the equilibrium queue formula, whose presence seems to correlate with the possibility of formulating the queue process on infinitesimal time intervals, has been introduced into the equilibrium variance formula in a consistent way. An inconsistency in characterising the statistics of arrivals between the dispersion index of arrivals, as obtained naturally during the derivation of the Pollaczek-Khinchin mean formula, and the coefficient of variation of arrivals as employed by some authors, has been discussed but remains unresolved.

It is shown that at some equilibrium probability distributions, in particular those described, with mode ≤ 2, can be approximated using a doubly-nested geometric distribution expressed in terms of their three moments: utilisation or p_0, mean and variance. The extended M/D/1[G] queues, representing stochastic queues at signals with specific green period capacities, do not appear to have close form expressions for their moments. New approximations for these have been proposed, using a link-function approach, that are shown to be reasonably accurate over a range of green capacities and demand intensities, not just for ‘heavy traffic’ as in the case of earlier formulae of Miller and Newell. The doubly-nested geometric method can be used to estimate their probability distributions.

CHAPTER 4: TIME-DEPENDENT APPROXIMATION

4.1. INTRODUCTION

The need for simple and efficient calculation of time-dependent queuing in traffic modelling software, as well as enabling the properties of queues to be explored in a more unified and less algebraically demanding way than exact analysis, has encouraged approximate methods. ‘Shearing’, an heuristical merging of rigorous deterministic and equilibrium descriptions, is conceptually simple and does not require additional empirical parameters. The flexibility of the formulation is another reason for working with it here, where the primary purpose is to extend the range of applications including to variability and reliability. Exploration of alternative existing queue approximations and more fundamental improvements are left for future research.

Kimber and Hollis (1979) formalised shearing in terms of coordinate transformation, though a more revealing interpretation is that this treats the queue as quasi-static. Similar approaches were made by Doherty (1977) and Catling (1977), and hinted at by Newell (1971 or earlier). Rider (1976) also used a quasi-static approach. Kimber and Hollis quickly recognised that the sheared method could be inaccurate, especially during decay. They proposed replacing the initial queue by a shift in the time origin, except for initial values greater than twice the equilibrium queue, where they used a linear model, since the rate of change of the mean queue remains nearly constant as long as the queue is large enough to keep the service saturated.

The assumption that the queue is effectively in a constant state of quasi-equilibrium is a deep source of potential inaccuracy, especially in the middle stages of growth or during post-peak decay where actual queue size probability distributions tend to be far from equilibrium form. The new deterministic variance formula is generally at odds with the predictions of shearing not only during growth and decay but also at equilibrium. Although the sheared queue approximation converges reliably to the correct equilibrium value it never quite ‘forgets’ its initial state. This can be exploited to achieve a correction making it consistent with the variance formula at equilibrium. Using this asymptotic ‘anchor’ as well as that of expected initial behaviour, it should be possible to approximate the development of the queue more accurately than using the mean size constraints alone. Simultaneously, time-dependent variance can be estimated. Because of the form of the variance equation, this approach hinges on how the delay, the integral average of the queue function, is approximated along with the mean queue.

After reviewing the sheared queue method, this Chapter 4 looks at corrections and an alternative for the difficult decay regime, which wherever possible are directly related to initial and asymptotic behaviour, and tests these against simulations using the set of peak cases.

4.2. THE SHEARED QUEUE APPROXIMATION AND RELATED ISSUES

4.2.1 The shearing transformation

In the late 1970s, P D Whiting, working at the Transport Research Laboratory in the UK, proposed a ‘coordinate transformation’ to merge deterministic time-dependent and equilibrium queue formulae in order to produce an approximate time-dependent formula that could be applied seamlessly through and above saturation. This method, known as ‘shearing’, was applied first by Robertson and Gower (1977) in the traffic signal optimisation software TRANSYT, and further developed by Kimber and Hollis (1979) and Kimber and Daly (1986), while similar ideas of Doherty (1977) were developed by Catling (1977), who described a similar method with the emphasis on stochastic delay at signals. Rider (1976) investigated relaxation behaviour using an approach based on inverting the relationship between the steady-state queue and demand intensity, but restricted results to the time-dependent formulation of the probability of the queue being zero[38]. Newell (1982) considered queues developing through saturation, but appears not have been aware of this work and did not describe a transformation as such. An interpretation by Heydecker and Verlander (1998) will be discussed shortly.

The ‘coordinate transformation’ involved in shearing was originally posed diagrammatically, as shown in Figure 4.2.1, where different versions of traffic intensity are related by:

$$\rho_d - 1 = \rho - f_e^{-1}(L) \qquad (4.2.1)$$

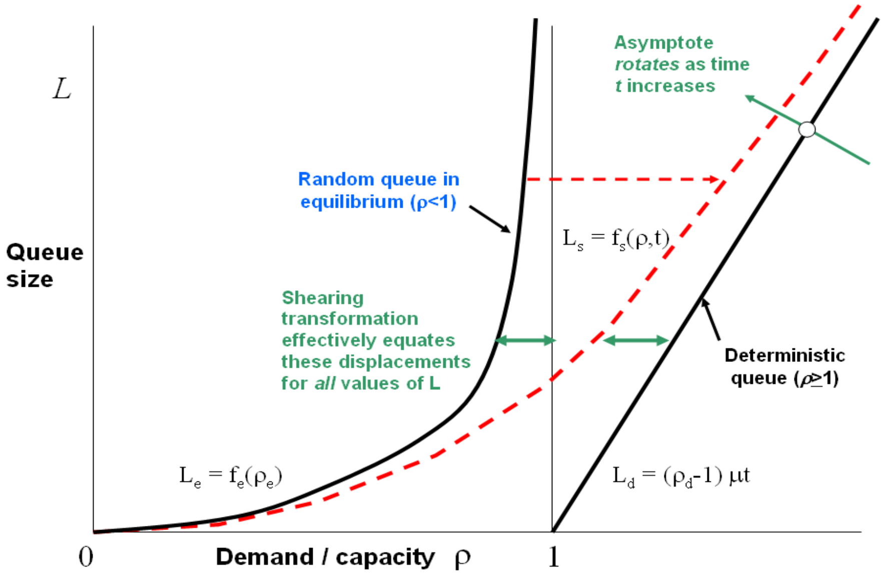

L
Queue size
Random queue in equilibrium (ρ<1)
Asymptote *rotates* as time *t* increases
$L_s = f_s(\rho, t)$
Shearing transformation effectively equates these displacements for *all* values of L
Deterministic queue ($\rho \geq 1$)
$L_e = f_e(\rho_e)$
$L_d = (\rho_d - 1)\,\mu t$
0
Demand / capacity ρ
1

Figure 4.2.1 Graphical interpretation of the shearing transformation

[38] Knowing this would be sufficient for defining a Geometric equilibrium probability distribution, but not any other kind of distribution.

In the transformation, ρ is the true demand intensity or ratio of demand to capacity, the function f_e is an equilibrium mean queue size function appropriate to the queue process, usually the Pollaczek-Khinchin mean value formula, and ρ_d is a synthetic variable that allows rewriting of the simple deterministic queue formula in the form of (2.3.4), assuming full utilisation of capacity - essentially (2.3.1) with constant parameters:

$$L_d(\rho,t) = L_0 + (\rho_d - 1)\mu t \equiv L_0 + \left(\rho - f_e^{-1}(L)\right)\mu t \qquad (4.2.2)$$

$$\text{where} \quad f_e^{-1}(L) \text{ is also the average utilisation } x(t) \qquad (4.2.3)$$

is the time-averaged utilisation or degree of saturation of *service* (at the stop line). Therefore (4.2.3) is equivalent to equating the time dependent queue to the equilibrium queue that *would be* generated at the service by traffic of intensity *x*, where it is convenient to use the forms of equations (2.3.50) or (3.2.14-15):

$$L(x,t) = \frac{I^* x + (C - I)x^2}{1 - x} \qquad \text{where} \quad I^* = I + \tfrac{1}{2}(I_a - 1) \qquad (4.2.4)$$

Since the system is *not* in steady-state equilbrium, this constitutes a *quasi-static* or *quasi-equilibrium* assumption, but unlike (2.3.50) which, as it stands, fails when $\rho \geq 1$, (4.2.4) is always defined since *x* is always less than 1. Therefore (4.2.2-4) can be solved for all *t*. The quasi-static approach is not restricted to this relatively simple and analytically soluble case. Holland and Griffiths (1999) show how it can be applied to the multi-channel queue (section 3.5.11 earlier), producing apparently very accurate results, but needing numerical solution.

4.2.2 Sheared queue solution with initial queue

The shearing transformation is equivalent to solving the equation[39]:

$$L(\rho, x, \mu, t) \equiv L_0 + (\rho - x)\mu t = \frac{I^* x + (C - I)x^2}{1 - x} = L_e(x) \qquad (4.2.5)$$

This is most simply solved for average utilisation *x* giving the quadratic solution:

[39] In queue formulae, capacity μ and time *t* always appear in combination, their product representing *throughput capacity*. Conventionally, however, the latter is not treated as an independent variable, and in any case it is sometimes desirable to refer to time in the absolute, though it always appears in a dimensionless expression.

$$x(t) = \frac{g - \sqrt{g^2 - 4fh}}{2f} \qquad \text{if } f \neq 0$$

$$x(t) = \frac{h}{g} \qquad \text{if } f = 0 \text{ and } g \neq 0 \qquad (4.2.6a)$$

where:

$$\begin{aligned} f &= \mu t - (C - I) \\ g &= L_0 + I^* + (\rho + 1)\mu t \\ h &= L_0 + \rho\mu t \end{aligned} \qquad (4.2.6b)$$

As an aid to differentiation note that this satisfies:

$$fx^2 - gx + h = 0 \qquad \text{so} \qquad (4.2.7a)$$

$$\frac{dx}{dt} = \frac{\mu\left(x^2 - (\rho + 1)x + \rho\right)}{g - 2fx} \qquad (4.2.7b)$$

$$\frac{dL}{dt} = \mu\left(\rho - x - t\frac{dx}{dt}\right) \qquad (4.2.7c)$$

The equivalent but more complicated solution for mean queue size L is as given by Kimber and Hollis (1979):

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

$$L_s(t) = \frac{H}{G} \qquad \text{if } F = 0 \text{ and } G \neq 0 \qquad (4.2.8a)$$

where:

$$\begin{aligned} F &= \mu t - (C - I) \\ G &= (L_0 - I^*)\mu t - 2(C - I)(L_0 + \rho\mu t) - (1 - \rho)(\mu t)^2 \\ H &= -\left[(C - I)(L_0 + \rho\mu t) + I^*\mu t\right](L_0 + \rho\mu t) \end{aligned} \qquad (4.2.8b)$$

$L_s(t)$, with subscript s, here specifically represents the sheared queue *function*. The reason for identifying the function in this way is that it is only an approximation to the mean queue, but there will be occasion later to employ the *function* L_s with modified parameters to achieve better accuracy or to modify its relaxation behaviour, while retaining its structural properties. Those properties make this quadratic function a useful alternative to the exponential for modelling saturating systems.

4.2.3 Sheared queue solution with origin time shift

Kimber and Hollis (1979) proposed an alternative formulation, eliminating L_0, in which the queue starts at zero at some time $-t_0$ so as to reach L_0 at t=0. They argued that the original method was inaccurate through not taking account of the history of queue development up to L_0. While substituting an hypothetical time t_0 for a real initial queue L_0 may be considered a drawback, since L_0 is likely to be available directly from calculation of earlier time periods. However, it may still be used as an alternative definition of the queue *function*. The sheared queue is reformulated as (4.2.9), where x^*, being the average utilisation over [-t_0,t] rather than [0,t], has the same form as (4.2.6a) but with L_0 suppressed and modified components:

$$L(t) \equiv (\rho - x^*)\mu t^* = Ix^* + \frac{Cx^{*2}}{1-x^*} = L_e(x^*) \qquad \text{where } t^* = t + t_0 \qquad (4.2.9)$$

$$\begin{aligned} f^* &= \mu t^* - (C - I) \\ g^* &= I^* + (\rho + 1)\mu t^* \\ h^* &= \rho\mu t^* \end{aligned} \qquad (4.2.10a)$$

To find t_0 in terms of L_0 it is easiest to invert the queue formula:

$$L_t(t) = \frac{G + \sqrt{G^2 - 4FH}}{2F} \qquad \text{if } F \neq 0$$

$$L_t(t) = \frac{H}{G} \qquad \text{if } F=0 \text{ and } G\neq 0 \qquad (4.2.10b)$$

where, on setting L_0=0 in (4.2.8b) and substituting shifted t^* for t:

$$\begin{aligned} F^* &= \mu t^* - (C - I) \\ G^* &= -(1-\rho)(\mu t^*)^2 - [I^* + 2\rho(C - I)]\mu t^* \\ H^* &= -\rho[I^* + \rho(C - I)](\mu t^*)^2 \end{aligned}$$

and by definition: (4.2.10c)

$$F^* L^2 - G^* L + H^* = 0 \qquad (4.2.11)$$

Setting t=0, L=L_0 and rearranging forms a quadratic in t_0:

$$t_0 = -\frac{1}{\mu}\left(\frac{R + \sqrt{R^2 - 4QS}}{2Q}\right) \qquad \text{(if } Q\neq 0) \qquad (4.2.12a)$$

where:

$$\begin{aligned}
Q &= (1-\rho)L_0 - \rho\left[I^* + \rho(C-I)\right] = (1-\rho)(L_0 - L_e) \\
R &= \left[L_0 + I^* + 2\rho(C-I)\right]L_0 \\
S &= -(C-I)L_0^2
\end{aligned} \qquad (4.2.12b)$$

Technically this should work only if the queue is growing, so there is a finite time at which it is zero. For decaying queues, Kimber and Hollis (1979) calculate the time at which the queue should equal twice the equilibrium value, assuming that it decays linearly. Thereafter they assume it decays according to a mirror image of the growth function.

Superficially the constructions appear equivalent, but in reality they are not. Referring back to section 2.5.8, the time t_0 corresponds to a particular point in the development of the queue from zero, at which the queue size probability distribution will differ from an equilibrium distribution in some specific (even if unknown) way. In the construction with L_0, however, the initial queue size distribution at t=0 is unspecified, although (4.2.5) means that quasi-equilibrium is assumed implicitly. If a decaying queue is assumed to fall to $2L_e$ at t=t_c, a yet different distribution ought to apply. Hence the use of subscript *t* instead of *s* to emphasise that they represent different model processes.

It is not the purpose here to compare the basic and origin-shifted sheared methods, for two reasons. First, corrections to obtain the correct asymptotic variance will inevitably alter the behaviour of the functions, making such a comparison of academic interest. Second, since both are approximations, which one works best may depend on circumstance, such as whichever happens by chance to reflect most closely the true probability distribution at a particular stage in queue growth. Figure 1.1.3 earlier suggests that origin-shifting does not necessarily result in greater accuracy, but this could be a topic for further investigation.

4.2.4 Relationship between queue size and delay and derived queue functions

In principle the sheared *function*, with time replaced by any monotonically increasing function of time, could represent an instantaneous queue size or its average ‘delay’, because the initial and final (equilibrium) values would be the same. Kimber and Hollis (1979) approximated the time-averaged queue, or delay per unit time, by the following on the basis that the queue function is nearly linear at the beginning and end of the time range:

$$D_s(t) \approx L_s\left(\frac{t}{2}\right) \qquad (4.2.13)$$

This delay function can be calculated as in section 4.2.2, but x will be different from that for the sheared queue, and the approximation will be poorest at some intermediate point where the function is most non-linear. Time transformation is given added impetus by the fact that as shown later, in particular M/M/1 cases where L_0=0 and ρ<1, as $t \rightarrow \infty$, the sheared queue *function* $L_s(t)$ is a better model of D in the variance formula (2.3.27) than is $D_s(t)$ itself.

Given the integral relationship between L and D (2.3.28), there may be potential benefit in deriving L from D because integrating L to get D is intractable, especially if time is replaced by a function of time, while differentiating D is more straightforward:

$$L(t)=\frac{d}{dt}(tD)=D(t)+t\frac{dD}{dt} \qquad (4.2.14)$$

Assuming D_s has the functional form of L_s, with the factor 2 in (4.2.13) replaced by a general function of time $\omega(t)$, where $L_s^{/}$ has the functional form of d$L_s(t)$/dt, the derived queue is:

$$L_d(t,\omega)=\frac{d}{dt}\left(tL_s\left(\frac{t}{\omega}\right)\right)=L_s\left(\frac{t}{\omega}\right)+\frac{t}{\omega}L_s^{/}\left(\frac{t}{\omega}\right)\left(1-\frac{t\dot{\omega}}{\omega}\right) \qquad (4.2.15)$$

For simplicity, L_s is now treated as just one of a range of possible *functions*. If ω is present, it can be accommodated through nested differentiation. Differentiating (4.2.5):

$$L_s^{/}(t)=\mu\left[(\rho-x_s)-tx_s^{/}\right]=\frac{dL_s}{dx_s}x_s^{/}=\left(\frac{I^*+(C-I)x_s(2-x_s)}{(1-x_s)^2}\right)x_s^{/} \qquad (4.2.16)$$

Hence the time derivative of x_s is expressible in terms of x_s:

$$x_s^{/}(t)=\frac{(\rho-x_s)\mu}{\mu t+\left(\frac{I^*+(C-I)x_s(2-x_s)}{(1-x_s)^2}\right)} \qquad (4.2.17)$$

Using (2.3.24), containing instantaneous utilisation, if L_s is interpreted as queue size, (4.2.16) can be rearranged into an alternative form suggested by Prof. B G Heydecker, in which p_0 appears as the weighted sum of time-averaged and equilibrium quantities:

$$p_0(t)=(1-\rho)+\frac{(\rho-x_s)}{1+(dx_s/dL_s)\mu t}=\frac{(1-x_s)+(1-\rho)(dx_s/dL_s)\mu t}{1+(dx_s/dL_s)\mu t} \qquad (4.2.18)$$

If L_s is used to estimate *delay*, then utilisation requires taking the second derivative:

$$L_s^{\prime\prime}(t) = -\mu\left[\left(2x_s^{\prime}(t)\right) + t x_s^{\prime\prime}(t)\right] \qquad (4.2.19)$$

$$x_s^{\prime\prime} = \frac{2x_s^{\prime 2}\left[(2x_s - \rho - 1)\mu + (\mu t - (C - I))x_s^{\prime}\right]}{\mu(\rho - x_s)(1 - x_s)} \qquad (4.2.20)$$

If $L_s(t)$ is used to represent *queue size* then:

$$u_s = \rho - \frac{L_s^{\prime}(t)}{\mu} \qquad (4.2.21)$$

If $L_s(t/\omega)$ is used to model *delay* then:

$$u_d = \rho - \frac{L_d^{\prime}(t,\omega)}{\mu} \qquad \text{where} \qquad (4.2.22)$$

$$L_d^{\prime} = \frac{2}{\omega}\left(1 - \frac{t\dot{\omega}}{\omega} - \frac{t^2}{2\omega}\left(\ddot{\omega} - \frac{\dot{\omega}^2}{\omega}\right)\right)L_s^{\prime} + \frac{t}{\omega^2}\left(1 - \frac{t\dot{\omega}}{\omega}\right)L_s^{\prime\prime} \qquad (4.2.23)$$

These formulae are available for calculating properties of variations of the sheared queue as and when necessary.

4.2.5 Effect on estimation accuracy of assumptions about the initial state

This sub-section comments on some issues of accuracy and interpretation associated with the sheared approximation and time slicing, in the light of the new result for variance. Kimber and Hollis (1979) argue that origin-shifting gives more accurate results for all values of time, because the shape of the queue development is no longer dependent on the initial queue size. The results are certainly more consistent in that sense. However, as pointed out earlier, a queue which grows from zero for a time t_0, reaching mean size L_0, is not the same as a queue that has exact size L_0 at t_0, because the former will have developed a probability distribution around L_0, whereas an initial L_0 could represent a ‘pure state’. It is sometimes argued that an initial L_0 represents the mean of a distribution, indeed Heydecker and Verlander (1998) suggest that it represents an *equilibrated* distribution. However, the distribution it actually represents cannot be controlled, and the lack of information about it within the method is a source of error. In general, it will not be possible to find a single time t_0 at which both initial queue and initial variance are matched, let alone three moments if p_0 is included.

Suppose (4.2.9) is accepted and a queue starting from zero at t=-t_0 grows to a mean value of L_0 at t=0, related to the initial queue by the following, assuming an M/M/1 process:

$$\mu t_0 = \frac{L_0}{\rho - \frac{L_0}{L_0 + 1}} \tag{4.2.24}$$

According to (4.2.13), the average delay per unit time over the growth period, say D_0, is approximately equal to the queue size at -t_0/2. While L_0 is small, (4.2.9) shows that the queue grows almost linearly, so initially $D_0 \approx L_0/2$.

Now introduce initial variance V_0. Bearing in mind that L_0, V_0 represent results after a time t_0, and the queue size and variance to put into the equation at t = -t_0 are both zero, (2.3.27) leads to the following *estimate* of variance at t = 0:

$$V_0^* = \left(2\rho - (1-\rho)L_0\right)\mu t_0 - L_0^2 - L_0 \tag{4.2.25}$$

On eliminating t_0 in (4.2.25) using (4.2.24), several terms cancel giving the result:

$$V_0^* = \frac{\rho L_0}{\rho - \frac{L_0}{L_0 + 1}} \tag{4.2.26}$$

If the queue *is* equilibrated at t=0 then the mean and variance should be related in the same way as (2.3.32) and (2.3.33), i.e. $V_0^*=L_0(L_0+1)$. Substituting this in (4.2.26) and multiplying out by the RHS denominator results in the following, but this is satisfied only if ρ=1:

$$\rho(L_0 + 1) - L_0 = \rho \qquad \Rightarrow \qquad \rho=1 \tag{4.2.27}$$

So, thanks to (2.3.27), it has been shown that in general the queue *cannot* be equilibrated at t_0. This does not matter greatly since the means to calculate p_0, L and V at moderate values of t are now available, and in principle the queue size probability distribution can be inferred from them. However, this only works for a queue starting at zero, and at large values of t equation (2.3.27) becomes increasingly sensitive since it involves the product of a number that is converging on zero and another that is increasing without bound. This applies also to the deterministic queue formula, but the sheared transformation takes care of it automatically.

Variance cannot be sheared because its time-dependent graph falls below the deterministic graph (see Figure 1.1.1 in the Introduction earlier). In any case, shearing the variance would beg the question of the accuracy of the result, whereas it would be preferable to use the variance formula to improve the accuracy of the whole approximation. Therefore other methods are required, which are the subject of much of what follows.

4.2.6 Effect of time slicing on accuracy

Equation (4.2.5) effectively equates the *instantaneous* time-dependent queue at time *t*, given by the middle formula, with the *equilibrium* queue that would result from an *average* degree of saturation *x* of service over [0,*t*], given by the RHS. This quasi-steady-state assumption is necessarily an approximation and a potential source of error.

Brilon (2007) looks at several variants of time-dependent equation, mostly describing delay rather than queue size, and points to problems that arise when like is equated with not exactly like. However, the approximation is convenient, avoiding the “disheartening”[40] complexity of exact transient solutions (e.g. Morse 1958), which in any case exist for few processes.

Heydecker and Verlander (1998) argue that better results could be got by integrating the queue size over very short time steps. This would certainly be advisable if ρ and μ vary significantly on short time scales, but it could be computationally demanding. They derive their main result for the rate of change of queue size by a fairly complicated procedure reinterpreted here. The rate of change of the mean queue in terms of instantaneous utilisation *u* or average degree of saturation *x* is:

$$\frac{\partial L}{\partial t} = (\rho - u)\mu = \left(\rho - x - \frac{\partial x}{\partial t} t\right)\mu \qquad (4.2.28)$$

Now if the queue is assumed to be quasi-equilibrated, treating L_e as a *function*:

$$\frac{\partial L}{\partial t} = \frac{\partial L_e}{\partial t} = \frac{\partial L_e}{\partial x}\frac{\partial x}{\partial t} \qquad (4.2.29)$$

Substituting for $\partial x/\partial t$ in (4.2.29) using (4.2.28) and rearranging, noting from that (4.2.5) $L_e(x)$ is an invertible function of *x*:

[40] Kleinrock (1975)

$$\frac{\partial L}{\partial t} = \frac{(\rho - x)\mu}{1 + \frac{\partial x}{\partial L_e}\mu t} \tag{4.2.30}$$

This is similar to the result got by Heydecker and Verlander (1998), and amounts to a way of converting between instantaneous utilisation and average utilisation. This in turn means that the queue development can be handled in a sequence of short quasi-static time slices, but this can also be done through (4.2.5-7).

Here, arrival and service rates can be generalised to time-dependent functions only if the quasi-static assumption remains consistent. There is no reason why it shouldn't, since it contains no assumptions about the behaviour of arrivals or service. However x and ρ then lose their meaning unless defined in terms of *average* throughput capacity and demand. If x is *defined* more generally by (4.2.31), then (4.2.30) is replaced by (4.2.32):

$$L(t) = \int_0^t \lambda(y)dy - x(t)\int_0^t \mu(y)dy \tag{4.2.31}$$

$$\frac{\partial L}{\partial t} = \frac{\lambda - x\mu}{1 + \frac{\partial x}{\partial L_e}\bar{\mu} t} \tag{4.2.32}$$

Apart from the arrivals λ now being a function of time, the essential difference is that the capacity μ in the numerator is instantaneous while that in the denominator is averaged. Instantaneous arrivals and capacity can change discontinuously (step functions) but the degree of saturation x is averaged over $[0,t]$ and therefore must be continuous, even if not everywhere differentiable. However, when the history of arrivals and service is allowed total freedom, the assumption of quasi-equilibrium seems increasingly artificial.

The approach adopted in this research retains where possible the quasi-static basis of time-dependent queue modelling, which has proved extremely successful, robust and applicable to a wide range of problems, while steering it towards greater accuracy through enhancements linked wherever possible to established queue properties. Developing *inherently* more accurate efficient queue approximations, that could involve dropping or modifying the quasi-static assumption, is not considered part of this research, but could be a topic of future research that could include reviewing the alternative approximate time-dependent methods which have variously been proposed.

4.2.7 Practical issues of predicting queues at junctions

Queue and delay measurements can be explained *post hoc* by measuring arrival rate profiles and measuring or inferring capacity profiles, but reliably predicting either ahead of time is more problematic. Kimber and Daly (1986) demonstrated that predicted queue sizes at a give-way junction, using the time-dependent method, could match observation roughly within a range of a factor of two either way, consistent with the exponential distribution of M/M/1 arrival and service times.

The futility of *over-precise* predictions is clear from this, but *accuracy* can be said to be achieved by methods that both deliver mean predictions within the expected range of variability or uncertainty, and respond realistically to *changes* in their data, and so are useful for predicting the effects of different designs or policies. This is more likely if the methods embody essential structural features or at least realistic constraints. It is these properties to which the sheared approximation owes its success.

However, queue variance and probability distributions have been left out of time-dependent modelling until now because they are not as easily measured as mean queue size, nor have analytical descriptions been available.The next Sections of this Chapter aim to address this.

4.3. CORRECTING THE SHEARED QUEUE APPROXIMATION FOR M/M/1

4.3.1 Extended sheared formula

The shearing approximation including dispersion of arrivals is, from (3.2.13):

$$L(t) \equiv L_0 + (\rho - x)\mu t = Ix + \frac{(I_a - 1)x}{2(1-x)} + \frac{Cx^2}{1-x} \qquad (4.3.1)$$

As in equations (3.2.14-15) this can be rewritten combining *I* and I_a into one coefficient:

$$L(t) \equiv L_0 + (\rho - x)\mu t = \frac{I^* x + (C - I)x^2}{1-x} \qquad (4.3.2)$$

where $$I^* = I + \tfrac{1}{2}(I_a - 1) \qquad (4.3.3)$$

This is convenient because *C* appears in the sheared solution only in the form (*C-I*), and I^* substitutes directly for *I* wherever it appears alone, and is identical to *I* as long as I_a=1. This makes one wonder whether I^* might have some physical significance, but difficulties arise because *I* appears to have only two possible values, 0 and 1, at least for single-channel queues.

4.3.2 Relationship between queue size and delay

Shearing the variance appears not to be possible, since the time-dependent graph falls below the deterministic line (see Newell (1968a) and Figure 1.1.1d in the Introduction) (it would also require solving a quartic not a quadratic). A stronger argument is that parallel approximations to mean and variance would not be expected to be mutually consistent, so additional correction would be required. Instead a way is sought to adjust the mean queue formula so that the variance equation (2.3.27) is satisfied. This involves both mean queue *L* and its time-average *D*, which is the same as mean total delay per unit time, thus constraining the behaviour of both functions. Conditions can be imposed initially and at equilibrium where true values are known. *D* can be calculated from the integral of *L* in accordance with (2.3.28), but the sheared queue function is awkward to integrate. Kimber and Hollis (1979) make use of the approximation:

$$D_s(t) \approx L_s(t/2) \qquad (4.3.4)$$

This delay is accurate at $t \to 0$ and $t \to \infty$ because the functions are almost linear there, but is inevitably less accurate at intermediate values where their behaviour is more dynamic. In effect the P-K formula is now sheared to a deterministic model of delay, where t is replaced by $t/2$ in (4.3.1-2), which can be considered exact if x is given a modified interpretation. This may seem artificial, but the original interpretation of x is also artificial and approximate In Figure 4.3.1, for $\rho=0.8$, it can be seen how the sheared queue function starts close to simulated L but as time progresses moves closer to simulated D. Note also that neither $L_s(t)$ nor $L_s(t/2)$ is close to the true L, at least in this simple case. So to satisfy the variance equation, L_s is the best estimate of *delay* D as $t \to \infty$, and while $L_s(t)$ gives the correct mean *queue* as $t \to 0$, $L_s(2t)$ is better as $t \to \infty$. Since V depends on both L and D, the greatest benefit is likely if they are compatible.

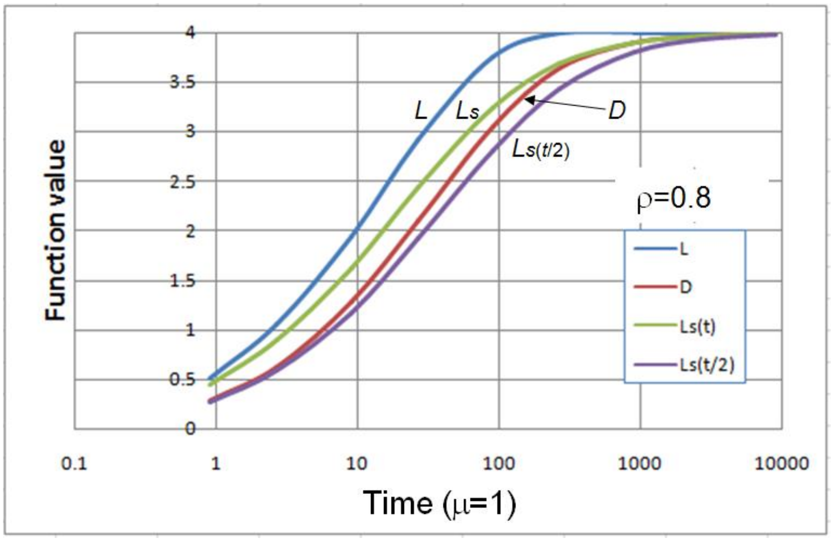


Figure 4.3.1 Markov simulated and sheared queue growth functions

To investigate the asymptotic behaviour of the sheared formula it is convenient to recast the sheared formula as a function of $z=1/t$, knowing that the Taylor expansion is convergent:

$$L(z) \equiv L_0 + \frac{(\rho - x)\mu}{z} = \frac{I^* x + (C - I)x^2}{1 - x} = L_e(x) \qquad (4.3.5)$$

This is equivalent to solving:

$$\alpha x^2 + \beta x + \chi = 0 \qquad \text{so that} \qquad (4.3.6a)$$

$$x(z) = \frac{-\beta + \sqrt{\beta^2 - 4\alpha\chi}}{2\alpha} \qquad \text{if } \alpha \neq 0$$

$$x(z) = -\frac{\chi}{\beta} \qquad \text{if } \alpha = 0 \text{ and } \beta \neq 0 \qquad (4.3.6b)$$

where:

$$\begin{aligned}\alpha &= (C-I)z-\mu \\ \beta &= \left(L_0+I^*\right)z+(\rho+1)\mu \\ \chi &= -\left(L_0 z+\rho\,\mu\right)\end{aligned} \qquad (4.3.6c)$$

From (4.3.6a) and LHS of (4.3.5), where $^{/}$ means differentiation with respect to z:

$$x^{/} = -\frac{(C-I)x^2+\left(L_0+I^*\right)x-L_0}{2\alpha x+\beta} \qquad (4.3.7a)$$

$$L^{/} = \frac{\left(I^*+(C-I)x(2-x)\right)x^{/}}{(1-x)^2} \qquad (4.3.7b)$$

And for the record the second derivatives are:

$$x^{//} = -\frac{2\left(\alpha x^{/}\left(x+x^{/}\right)+\left((C-I)x+L_0+I^*\right)x^{/}\right)}{2\alpha x+\beta} \qquad (4.3.8a)$$

$$L^{//} = \frac{2\left((C-I)x^{/}+L^{/}\right)x^{/}}{1-x}+\frac{L^{/}x^{//}}{x^{/}} \qquad (4.3.8b)$$

With these terms, the Taylor expansion around z=0 can be used to calculate limiting values. In the limit, with x=ρ and defining a factor ω dividing time, the terms become:

$$\begin{aligned}\alpha &= -\mu \\ \beta &= (\rho+1)\mu \\ \chi &= -\rho\,\mu \\ 2\alpha x+\beta &= (1-\rho)\mu\end{aligned} \qquad (4.3.9a)$$

$$\left. x^{/}\right|_{z\to 0} = -\frac{\left(L_0+I^*\right)\rho+(C-I)\rho^2-L_0}{(1-\rho)\mu} = -\frac{\left(L_e-L_0\right)}{\mu} \quad \text{so} \qquad (4.3.9b)$$

$$L_s\left(\frac{\omega}{t}\right)_{t\to\infty} = L_e-\left(L_e-L_0\right)\left[\frac{I^*+\rho(2-\rho)(C-I)}{(1-\rho)^2}\right]\frac{\omega_\infty}{\mu t} \qquad (4.3.10)$$

Since near z=0, $x=\rho+x^{/}z=\rho+x^{/}/t$, equation (4.3.9b) is trivially equivalent to the deterministic queue formula. To be compatible with the deterministic variance (2.3.27) the delay function must satisfy the following, regardless of the asymptotic behaviour of terms in L whose time-dependent components become negligible relative to that of D, which is multiplied by μt:

$$D(t \to \infty) = L_e - \frac{W_e - W_0}{2(1-\rho)} \frac{1}{\mu t} \tag{4.3.11}$$

Therefore to make the $L_s(.)$ *function* asymptotically compatible with simulated D:

$$\omega_\infty = \frac{(1-\rho)(W_e - W_0)}{2(L_e - L_0)(I^* + \rho(2-\rho)(C-I))} \tag{4.3.12}$$

In the simplest M/M/1 case (L_0=0, V_0=0, C=1, I^*=I=1, etc), this equals 1 for all ρ. This demonstrates that the basic sheared *queue* function is the better approximation to *delay* as $t \to \infty$ in this particular case, though the sheared delay (4.3.4) is necessarily better near t=0. Generally, ω_∞ can take different values. In the case of M/D/1, with unit-in-service it stays a little above 1, but without unit-in-service is close to 1/ρ, so can be moderate or very large. For example, for ρ=0.8, L_0=0, ω_∞ is around 1.1197 with unit-in-service and 1.1944 without unit-in-service, so no longer has a simple value.

The second order terms in D give no clue to the first order terms in whatever ought to serve as L, but only lead to a relationship involving the first order term of V. To investigate the behaviour of L its integral relationship to D (2.3.28) must be invoked. Since recovering L involves differentiating D, and the time divisors ω at t=0 and t=∞ are different, a time-variable time-scaling factor ω(t) must be introduced. In summary if, regardless of its parameters, delay D is defined using the *function* L_s then (where the subscript d means ‘derived’):

$$D(t) = \frac{1}{t} \int_0^t L(y) = L_s\left(\frac{y}{\omega}\right) dy \tag{4.3.13a}$$

$$L_d(t) = L_s\left(\frac{t}{\omega}\right) + L_s^{/}\left(\frac{t}{\omega}\right)\frac{\dot{\omega} t}{\omega} \tag{4.3.13b}$$

4.3.3 Basic corrections to delay

At t=0, linearity in the limit means ω must start with the value 2. In the simplest case of M/M/1 starting from zero, it falls to 1 at t=∞. To investigate its behaviour Markov simulation is used to generate queue growth profiles to compare with the sheared function for several values of ρ, then time is factored to get the sheared function L_s to match the simulated delay D. In order to compare the results on the same basis a time transformation is preferred that effectively makes the queue function independent of ρ (link function again), and also (possibly) independent of queue statistics, assuming that queue development is largely[41] monotonic:

$$L_0^* + \left(\rho^* - x^*\right)\mu t^* \equiv \frac{L_0 + (\rho - x)\mu t}{L_e(\rho)} \tag{4.3.14}$$

It is clear that the transformed equilibrium queue size is 1, so ρ^*=0.5 if the (*) function is assumed to be M/M/1-like. If L_0=0, and hence L_0^*=0, and μ is not transformed, a normalised time variable is defined by:

$$t^* = \frac{L_s(t)}{L_e\left(0.5 - x^*\right)} \qquad \text{where} \tag{4.3.15a}$$

$$x^* = \frac{L_s(t)}{L_e + L_s(t)} \tag{4.3.15b}$$

It is convenient to make the values of t used in the simulations and calculations a logarithmic sequence in multiples of τ_{re}. When ω is plotted against t^*, points for different values of ρ tend to cluster together, suggesting that a single function could indeed represent them all. While they do not converge exactly to the value 2 as $t^* \to 0$, but rather to around 1.8, the half-way value of ω occurs close to the ρ=0.5 relaxation time ($\tau_{re(0.5)}$) 11.66, suggesting that time or a function of it be factored by relaxation time. Now three possible functional forms of ω(t) are considered:

$$\omega_{[1]}(t) = \omega_0 + \left(\omega_\infty - \omega_0\right)\frac{\left(t^*/\tau_{re(0.5)}\right)^m}{1 + \left(t^*/\tau_{re(0.5)}\right)^m} \tag{4.3.16a}$$

$$\omega_{[2]}(t) = \omega_0\left(\frac{\omega_\infty}{\omega_0}\right)\wedge\left(\frac{\left(t^*/\tau_{re(0.5)}\right)^m}{1 + \left(t^*/\tau_{re(0.5)}\right)^m}\right) \tag{4.3.16b}$$

$$\omega_{[3]}(t) = \omega_0 + \left(\omega_\infty - \omega_0\right)\frac{\left(t/\tau_{re}\right)^m}{1 + \left(t/\tau_{re}\right)^m} \tag{4.3.16c}$$

[41] A queue growing from an initial pure or narrowly distributed state other than zero may sink somewhat before its probability distribution relaxes and the mean queue starts to increase.

In practice, nothing appears to be gained by using power *m* other than 1, leaving ω_0 as the only calibration parameter. Figure 4.3.2 shows RMS errors and optimum values of ω_0 as a function of the *minimum* value allowed to the time function, t^* or t/τ_{re}.

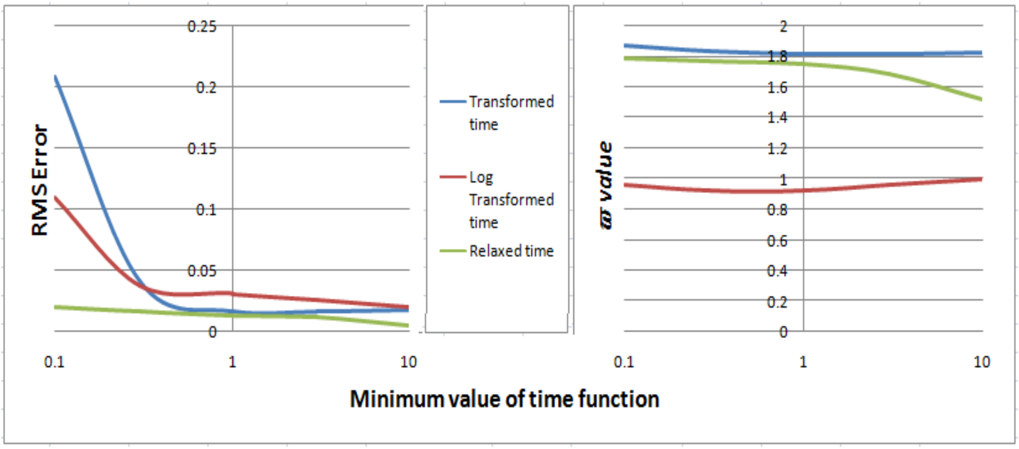


Figure 4.3.2 Performance of various candidates for the ω function

This shows that (4.3.16c) offers the most consistent performance, though this could leave unresolved the issue of what to do if the queue statistics are not M/M/1 and an appropriate formula for relaxation time is not available. Figure 4.3.3 plots $\omega_{[3]}$ against ‘relaxed’ time. Although RMS error remains low, points diverge at the lowest *t* values. Note that possible future use of time scaling based on the sheared function as in (4.3.15a) is not discounted. In fact this was the kind of approach tried first in the different form of the ratio of D_s to L_s.

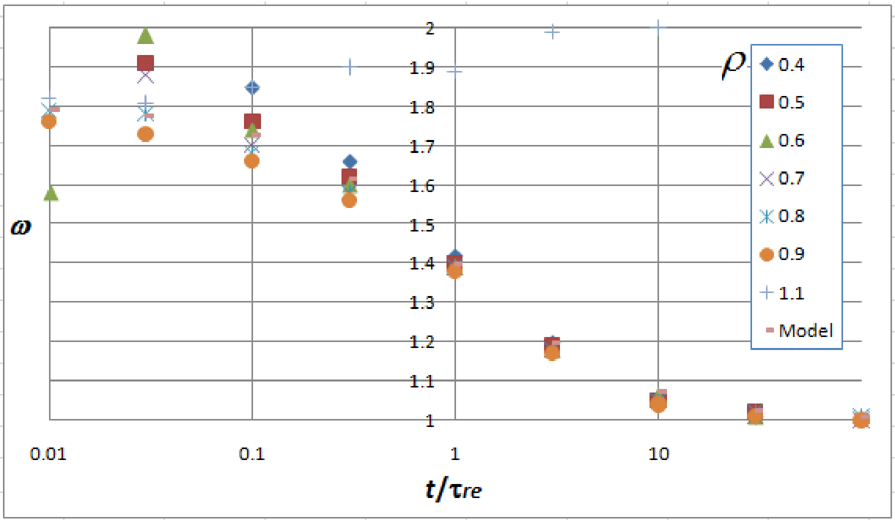


Figure 4.3.3 $\omega_{[3]}$ function against ‘relaxed time’ (two points with ω>2 omitted)

The formulae (4.3.16a,b) produce similar graphs but with substantially more divergence of points at low values of the time function (it is not clear why any divergence occurs).

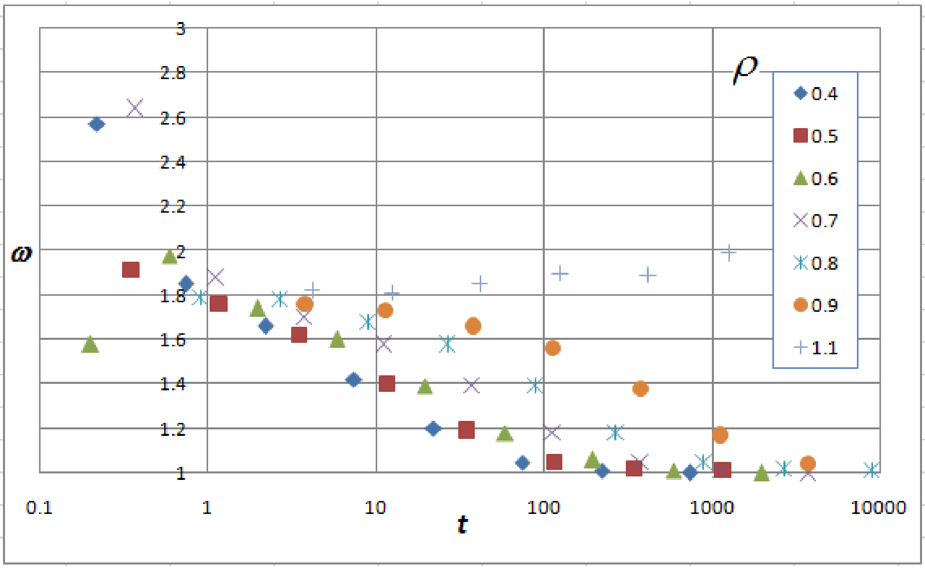


Figure 4.3.4 $\omega_{[3]}$ function against absolute time (μ=1)

In Figure 4.3.4, if a few rogue points resulting from inaccuracy at the extreme left are ignored, as well as points for ρ=1.1 (>1) which must behave differently, the impression is that data points tend towards a common trend as a function of low values of *actual* time (bearing in mind that τ_{re} ranges from 7.4 to 379.7, so *t*=1 in Figure 4.3.4 corresponds to values well below 1 in Figure 4.3.3). This trend might indeed eventually converge to a common value, but only at times so short they are of little practical interest. To get a handle on this it is necessary to simulate short growth times in fine detail. In doing so each time point is recalculated from a zero initial state to avoid inaccuracies in *D* to which ω is very sensitive.

.

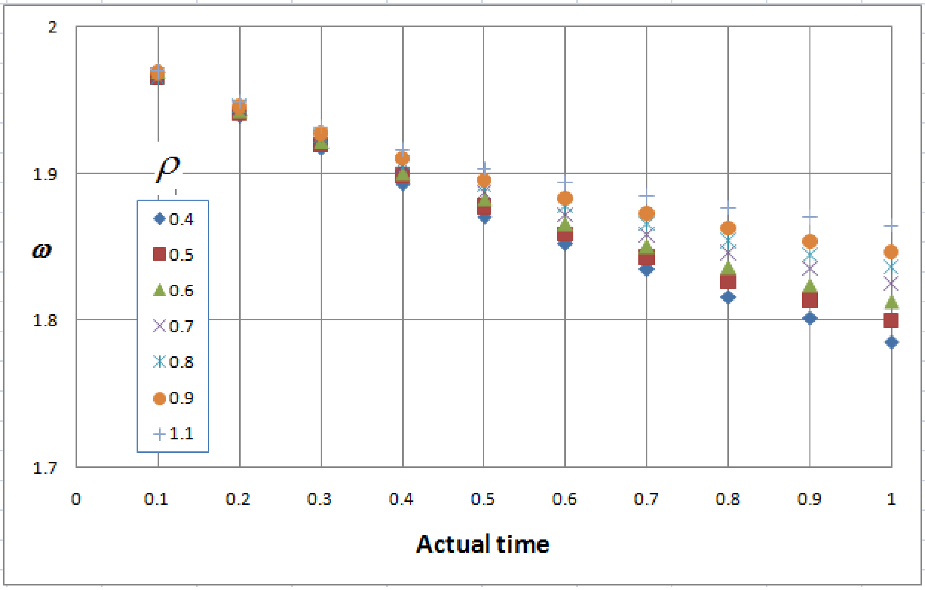


Figure 4.3.5 $\omega_{[3]}$ function against short actual times

In Figure 4.3.5, the behaviour of ω is seen to be converging smoothly towards 2 as $t \rightarrow 0$, but there appears to be no easy way to transform t so as to merge the graphs for different ρ. Various functions having been tried, probably the most exotic of which is to multiply t by $(1-\rho^2)^{1/4}$, but without a structural basis such functions are unlikely to extended to more general cases.

Returning to Figure 4.3.1, it is noticeable how smoothly D moves between $L_s(t/2)$ at small t and $L_s(t)$ at large t, and that the half-way point occurs near the relaxation time, 89.72 for this case (ρ=0.8), suggesting that a function weighting queue functions rather than time, say $\Omega(t)$, might be able to predict D. Figure 4.3.6 shows that when $\Omega(t)$ is defined and applied by equations (4.3.17), this behaves similarly to ω, including divergence at small times:

$$\Omega(t) = \frac{L_s(t) - D}{L_s(t) - L_s(t/2)} \qquad \text{or} \qquad (4.3.17a)$$

$$D = (1-\Omega(t))L_s(t) + \Omega(t)L_s(t/2) \qquad (4.3.17b)$$

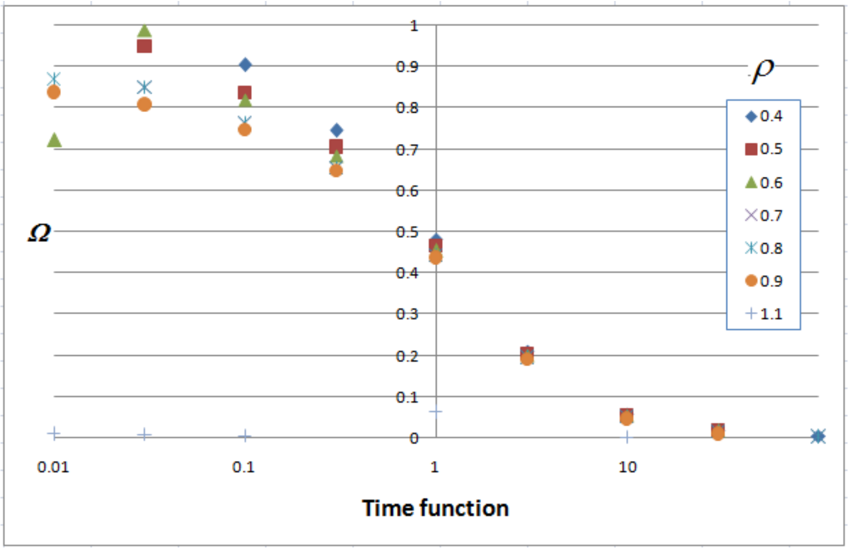


Figure 4.3.6 Ω function against ‘relaxed time’ (compare with Figure 4.3.3)

Absolute times less than 1 unit, representing 1/μ, are practically meaningless. Allowing that estimates become unstable at very small values of t, Figure 4.3.4 suggests that $\omega_{[3]}$ values plotted for various ρ converge around absolute time t=1, representing 1/μ since μ=1 for these tests. If instead of t/τ_{re} the time function is defined by ‘shifted relaxed time’, equation (4.3.18), much of the divergence at small values of t disappears, as shown in Figures 4.3.7 and 4.3.8 respectively (note that, as before, data for ρ>1 are not expected to comply):

$$t_1 = \frac{\mu t - 1}{\mu \tau_{re} - 1} \qquad (4.3.18)$$

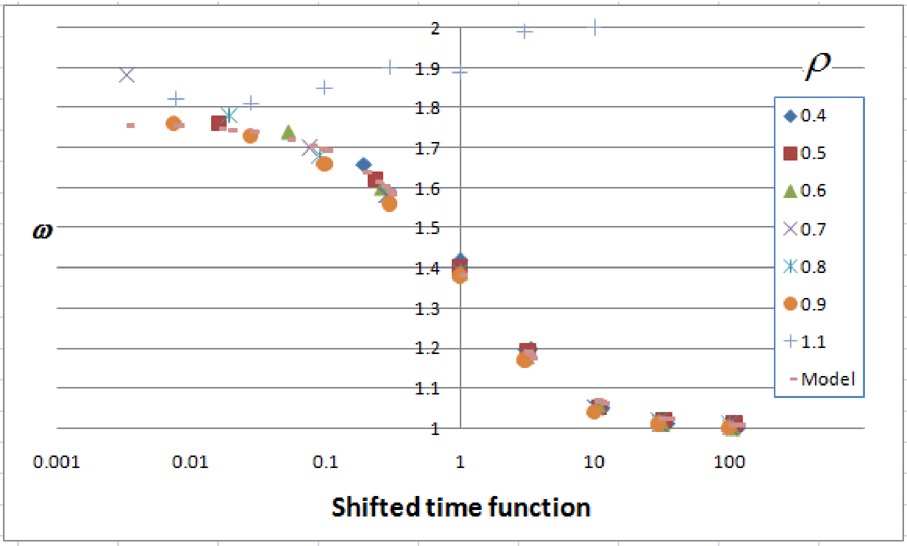


Figure 4.3.7 ω function v. shifted 'relaxed time', model factor = 1.77, RMSE = 0.024

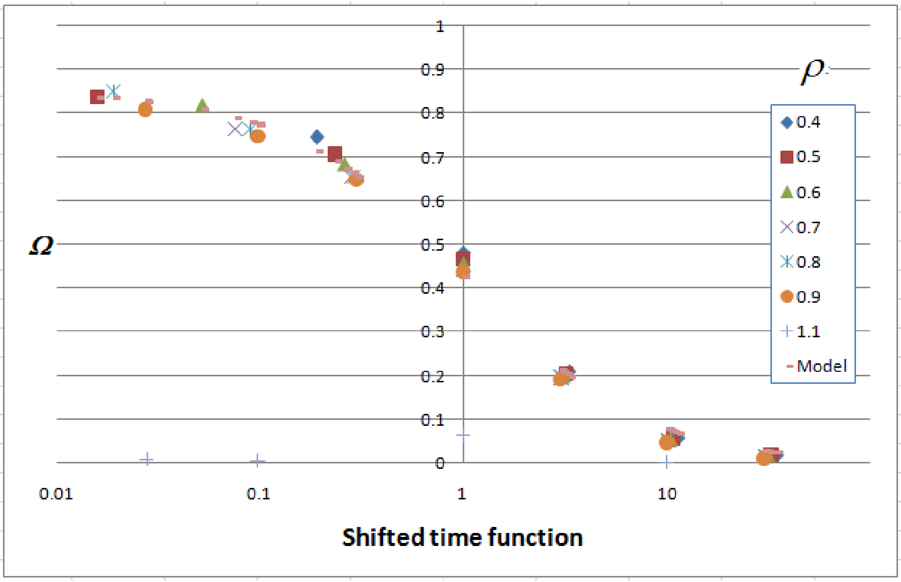


Figure 4.3.6 Ω function v. 'shifted relaxed time', model factor = 0.85, RMSE = 0.018

In both cases, the unfactored models overpredict, and consequently the factored models, that minimise error, are unable to achieve the theoretical result at extremely short times. While this may not matter in itself, it may matter that a factor must be introduced that could vary depending on the starting condition. The 'external' Ω function appears to give cleaner results than the 'internal' ω function, as well as being easier to work with as discussed earlier.

Regardless of initial conditions, $\omega_0 \approx 2$, and ω_∞ is given by (4.3.12). Equation (4.3.17b), together with the fact that all queue and delay functions start equal to L_0, implies that $\Omega_0 \approx 1$, and using (4.3.10-11) it can be shown that (4.3.19a) applies in general, and since the same difference of 1 applies at t=0, it may be *assumed* (if required) that it applies for all t.

Inspired by the Logistic form of Figure 4.3.6, Ω(*t*) is *defined* by (4.3.19b), using (4.3.18):

$$\Omega_\infty = \omega_\infty - 1 \qquad (4.3.19a)$$

$$\Omega(t) = \Omega_0 + (\Omega_\infty - \Omega_0)\left(\frac{t_1}{t_1 + 1}\right) \qquad (4.3.19b)$$

4.3.5 Interpolative correction in basic cases

Figure 4.3.7 plots RMS errors for the range of ρ<1 values tested.

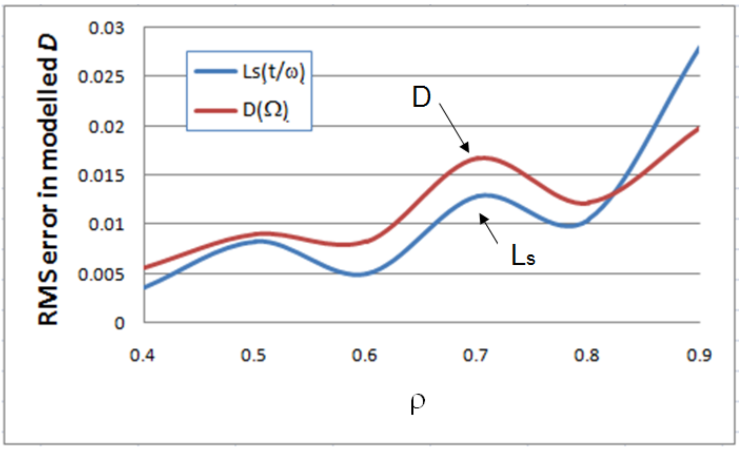


Figure 4.3.7 RMS errors for *D* corrected by ω and Ω methods relative to Markov simulated

Figure 4.3.8 shows that both the methods give very accurate results for delay growing from zero with ρ=0.8 (it is practically impossible to separate the graphs). Similar results are got for ρ=0.5, 0.7 and 0.9. For ρ=1.1, where Ω is forced to 1 (ω=2), the fit is also good as shown by Figure 4.3.9, with RMS error .3335 relative to *D* values ranging from 1.25 to 219.

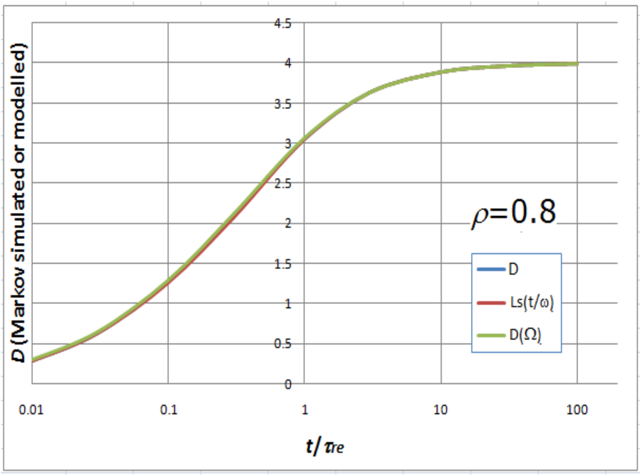


Figure 4.3.8 Comparison of Markov and ω/Ω-corrected *D* equilibrating example

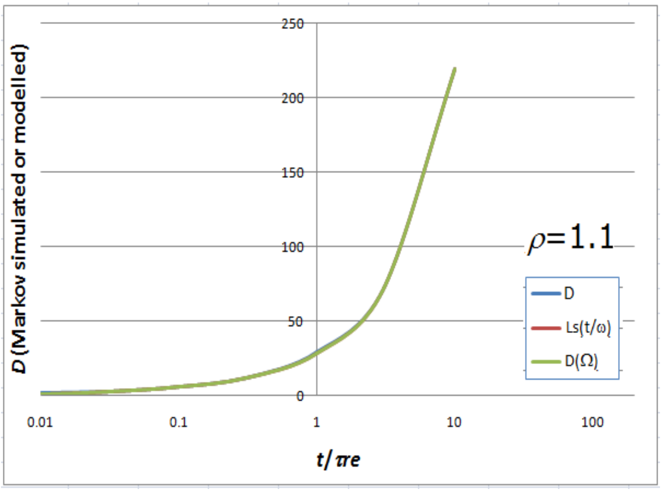


Figure 4.3.9 Markov and ω/Ω-modelled *D* functions for over-saturated growth example

4.3.6 Interpolative correction with general initial queue

Ω_∞ appears to be well defined in all cases, but it remains to evaluate Ω_0 in the general case when the initial queue is non-zero, and this does not appear to be possible using the method applied to Ω_∞. Examples with ρ=0.8 and L_0=2 have been tested. Figure 4.3.10 shows the dependence of ω on *t* in this case[42]. The final values are as predicted by equation (4.3.12). The initial value for equilibrated L_0 is as expected, but in the case of exact L_0 values cannot be calculated for small *t* using ω because *D* is not monotonic – it falls before it rises, whereas L_s can only be monotonic. However, Ω does not suffer from this limitation.

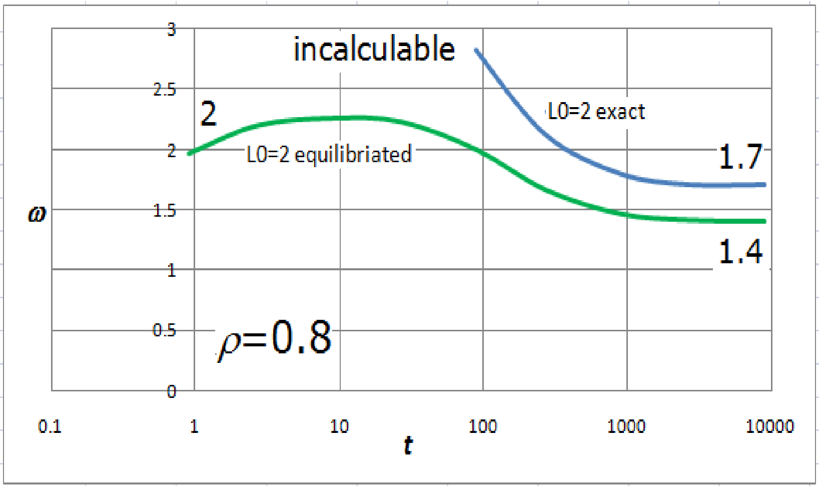


Figure 4.3.10 ω function for non-zero initial queue

Figure 4.3.11 (left) shows that $L_s(t/2)$ is a good approximation to *D*, so while one might ignore the hump in the ω function, Figure 4.3.12 (right) shows that L_s cannot match *D*.

[42] The presence of ω rather than Ω here is for historical reasons, and the adoption of Ω in (4.3.19b) really ought to follow this analysis, but the close relationship between these functions may be kept in mind.

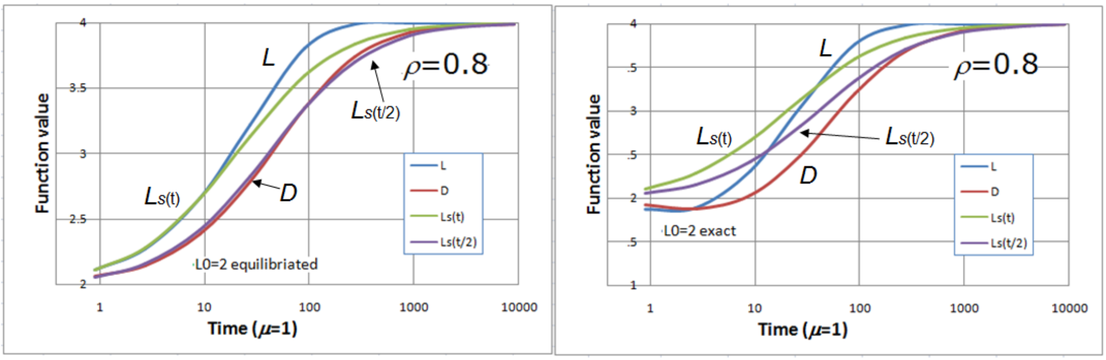


Figure 4.3.11 Queue and delay development for equilibrated or exact initial queue

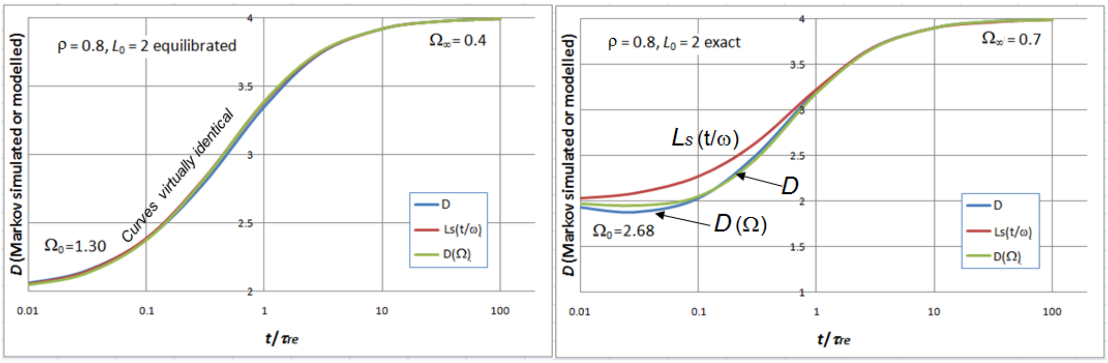


Figure 4.3.12 Markov and ω/Ω-estimated D functions for equilibrated or exact L_0

For the equilibrated initial queue, given that the rising part of the 'hump' in ω covers a relatively short period, if the initial value of ω can be set to the height of the hump, instead of 1.8 or 2, then quite a good fit can be achieved, as Figure 4.3.12 (left) shows, but a satisfactory fit is also obtained using Ω interpolation (4.3.17), and this achieves a much better fit when the initial queue is exact, Figure 4.3.12 (right).

The ω method has the fundamental limitation that it cannot reproduce a dip in D caused by high initial utilisation (e.g. associated with low initial variance). Only the Ω method can do this. The problem is now to determine Ω_0. The formula (4.3.17a) implies a 'limiting value' of Ω as $t \to 0$. This can be estimated by approximating D by $L_0+½(\rho-u_0)\mu t$ where u_0 is set to the value it ought to have at t=0, depending on the initial state assumed. At the same time, an empirical value of Ω_0 is found so as to minimise RMS error between Markov simulated values of D and those estimated by equations (4.3.16c) and (4.3.17b).

Bearing in mind that RMSE is biased by the logarithmic sequence of time points estimated, Table 4.3.1 gives for very small t (0.000001 used to avoid divide error) and three types of initial state (in groups of three cells). The limit Ω values can be rather sensitive and the correlation with the estimated values is weak. Figure 4.3.13 plots the optimum values.

Table 4.3.1 Experimental/estimated values of Ω_0 and calculated values of Ω_∞. In each group: limit Ω_0 values (left), 'optimum' values (middle), and theoretical values if known (right)[43]

ρ	L_0	L_0=0				L_0>0 pure state				L_0>0 equilibrated			
	if>0	Limit	'Opt.'	True	Ω_∞	Limit	'Opt.'	True	Ω_∞	Limit	'Opt.'	True	Ω_∞
0.25	**-**	1.00	0.83	1	0								
0.5	**1**	1.00	0.85	1	0								
0.7	**1**	1.00	0.89	1	0	3.50	2.35	-	0.525	3.49	1.22	-	0.3
0.8	**2**	1.00	0.85	1	0	3.36	2.68	-	0.7	0.97	1.30	-	0.4
0.9	**5**	1.00	0.83	1	0	2.46	2.97	-	0.875	1.00	1.40	-	0.5
0.95	**5**	1.00	0.83	1	0	2.46	1.74	-	0.304	1.00	1.29	-	0.25
0.95	**9**					2.51	2.51	-	0.675	1.41	1.41	-	0.45

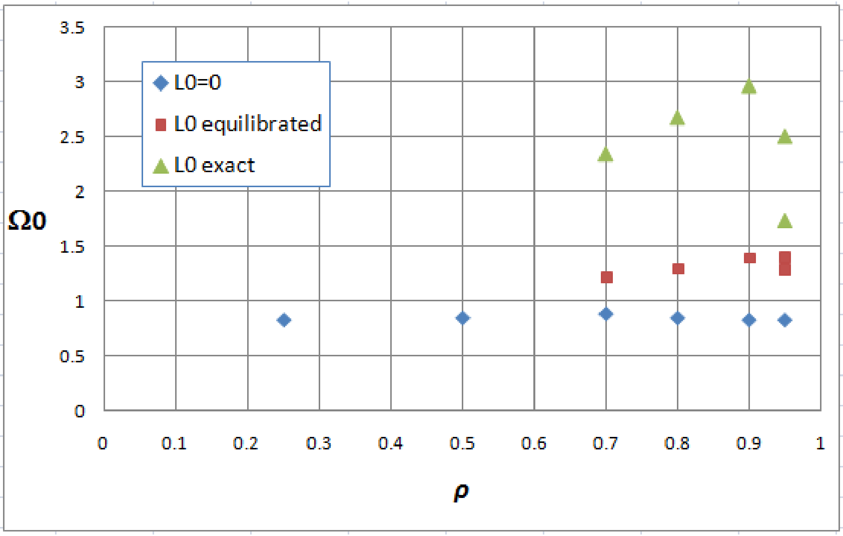


Figure 4.3.13 Optimal Ω_0 for various ρ and initial states

There is a practical limitation on the values of ρ that can be tested in this way. Values near 1 take a long time to simulate, and the Markov simulation program Qsim accepts only exact integer values of the initial queue (in addition of course to equilibrated initial states), meaning only ρ>0.5 can be tested usefully in that mode. However, lower values of ρ have been tested with L_0=0 to see whether $\Omega_{0(opt)}$ tends to 1, which it appears not to, and ρ=0.95 has been tested with two initial queue values, resulting in significant differences. The conclusion is that simple functions of ρ to predict Ω cannot be set up.

[43] Originally it was planned to label initial and asymptotic time correction parameters α and ω, but ω_0 and ω_∞ were considered more informative, hence leading to Ω_0 and Ω_∞. The parameter α was studied in some detail and methods found for matching highly non-linear simulated initial behaviour from low variance initial states. However, this approach was considered to be too complicated and empirical to be practical at this stage.

4.3.7 Interpolative correction for mean queue size

Assuming that Ω_0 can be estimated, this sub-section looks at performance in estimating the mean queue size. The mean queue function can be calculated by applying (4.2.14) and (4.2.16) to (4.3.17). The calculation is best done in stages because the full explicit expression is messy and hence prone to error. From the definition of *D* (2.3.28), the final stage of the evaluation is:

$$L(t) = D(t) + \dot{D}t \qquad \text{where}$$

$$\dot{D} = (1-\Omega)L_s^{/}(t) + \frac{\Omega L_s^{/}(t/2)}{2} - \frac{(\Omega_\infty - \Omega_0)(L_s(t) - L_s(t/2))(\tau_{re} - 1/\mu)}{(\tau_{re} + t - 2/\mu)^2} \qquad (4.3.20)$$

This makes use of the derivative $L_s^{/}$ of the L_s *function*, as given by (4.2.16), and the time derivative of Ω from (4.3.16c) and (4.3.18):

$$\dot{\Omega} = \frac{(\Omega_\infty - \Omega_0)(\tau_{re} - 1/\mu)}{(\tau_{re} + t - 2/\mu)^2} \qquad (4.3.21)$$

Figures 4.3.14 show that for ρ=0.8 the estimates of *L* and *V* based on (4.3.20) and (2.3.27) are quite accurate, except for some over- or undershoot in *V*. Similar results have been got for ρ=0.7 and 0.9[44]. Sheared mean queue and 'naïve' variance approximations are also plotted (dashed graphs), the latter using $L_s(t/2)$ to represent delay, illustrating the former's inaccuracy and the latter's inability to satisfy the steady-state asymptotic constraint. It is interesting that in this case the variance appears to 'lose its memory' of its initial state more quickly than the mean queue. It is noticeable, though not surprising, that the accuracy of *V* is less than that *D* and *L*, even neglecting overshoot. Therefore there may be a case for adjusting Ω to minimise some practical combination of the errors in the variables.

4.3.8 Calculation of initial Ω_0 for undersaturated growth

Since *L* appears to be estimated accurately by (4.3.20), differentiation is expected to lead to a formula for Ω_0. On the face of it, this appears complex, but differentiating (4.3.20) relates the rate of change of mean queue to Ω, and its derivative which is simplified by (4.3.21):

$$\dot{L} = 2\dot{D}t + \ddot{D}t \qquad (4.3.22)$$

[44] There does not appear to be a tendency for errors to escalate as $\rho \rightarrow 1$.

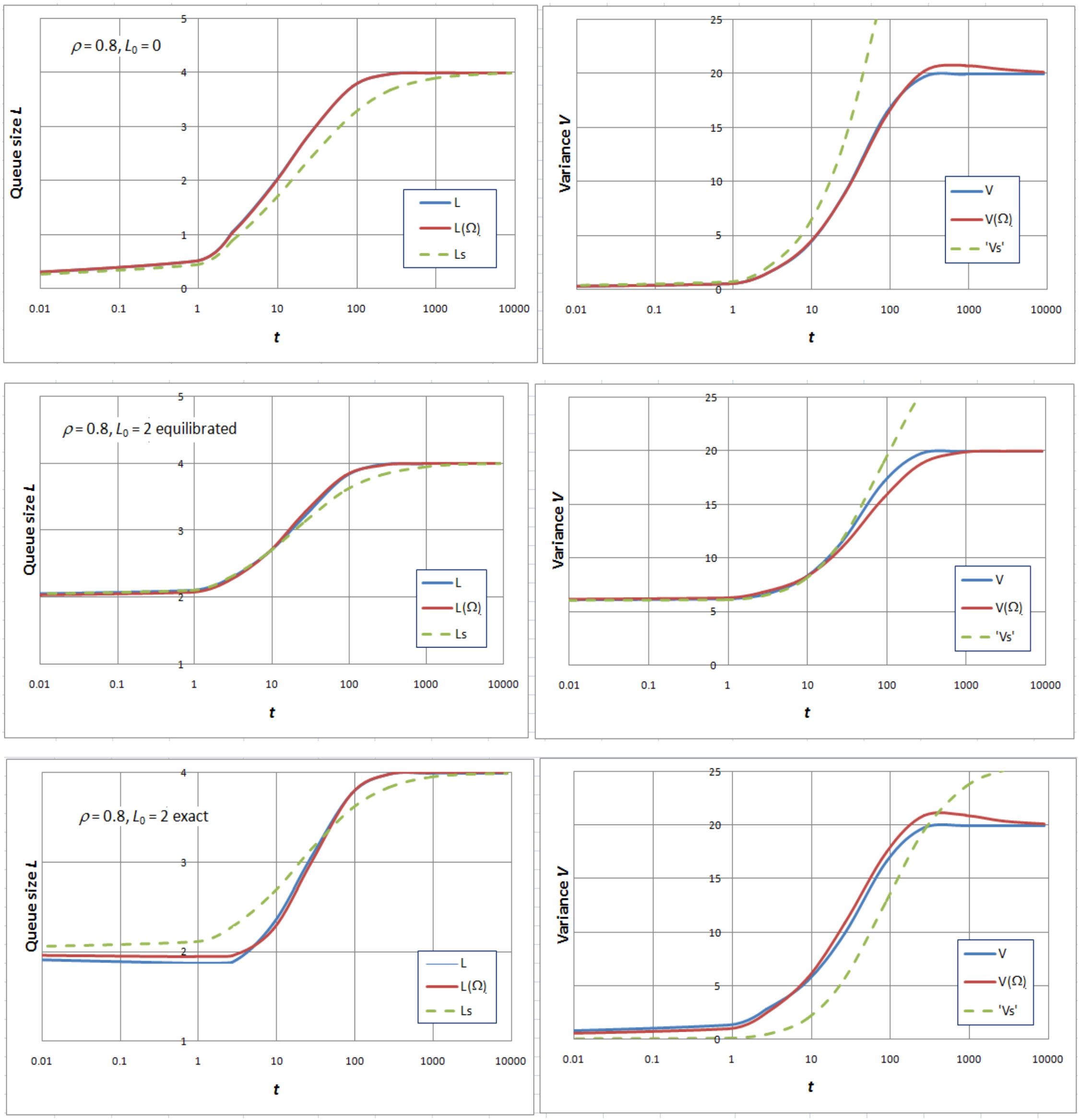


Figure 4.3.14 Fit of Markov and ω/Ω-estimated L and V for various initial states

Expressions that multiply t and remain finite as $t\to0$ will vanish. Since this is true of $L_s^{/}(t)$ and $\Omega^{/}$, it is necessary only be sure that $L_s^{//}(t)$ remains finite to eliminate the last term in (4.3.22). Tedious evaluation of (4.2.19) shows that $L_s^{//}$ is linear in $x_s^{/}$ and $x_s^{//}$, and that $x_s^{//}$ is linear in $x_s^{/}$. The only risk factor lies in the denominator that can become zero when x_s=1. As there is no reason to suppose the expressions 'explode' at the last moment when moving towards an exact initial state, it is assumed that derivatives remain finite. Now from the initial rate of change of the queue size it follows from equation (4.3.20) that:

$$(\rho - u_0)\mu = 2\lim \dot{D}_{t\to0} \qquad (4.3.23)$$

The final term in $\dot{D}$ declines to zero, leaving only the first two terms, and the derivative functions (4.2.18) simplify greatly and smoothly when $t\to0$, hence:

$$\Omega_0 = \frac{(\rho - u_0)\mu - 2L_s^{/}(t)\big|_{t\to 0}}{L_s^{/}(t/2)\big|_{t\to 0} - 2L_s^{/}(t)\big|_{t\to 0}} = \frac{\rho + u_0 - 2x_s}{\rho - x_s} \qquad (4.3.24)$$

while, for completeness, from (4.3.12) and (4.3.19a):

$$\Omega_\infty = \frac{(1-\rho)(W_e - W_0)}{2(L_e - L_0)(I^* + \rho(2-\rho)(C-I))} - 1 \qquad (4.3.25)$$

In (4.3.24), x_s is obtained by inverting the P-K mean function. In the M/M/1 case, Ω_0 can be expressed simply in terms of L_0:

$$\Omega_{0(M/M/1)} = \frac{(\rho + u_0)(L_0 + 1) - 2L_0}{\rho(L_0 + 1) - L_0} \qquad (4.3.26)$$

This necessarily gives the value 1 for all equilibrated initial states tested that are not ‘singular’ (in the sense of ρ=0.5, L_0=1), including L_0=0, since in those cases x_s=u_0. For exact initial states, Ω_0 can take on a range of values >1, since u_0=1 necessarily exceeds equilibrium utilisation. Qsim is able to generate examples with intermediate values of initial variance (using a Normal distribution), and these also produce Ω_0>1 for the same reason. In principle, extreme ‘long tail’ distributions could occur where V_0 exceeds the equilibrium value associated with L_0, but these are likely only when a queue is decaying and would be expected to result in Ω_0<1.

It is interesting, and significant, that Ω_0 involves the initial utilisation and mean queue, but not initial variance, while Ω_∞ involves initial and equilibrium mean queue and variance, but not initial utilisation. Taken together, these quantities send the strong message that all *three* moments - allowing utilisation, or its complement p_0, the style of ‘moment’ – are necessary to characterise the state of a queue. Furthermore, these results not only answer the requirement to make the approximation to queue development consistent with the variance equation, but also bring in the explicit dependence on the initial probability distribution which has been identified as lacking in previous methods where only the mean queue was considered.

Because of the amount of work involved, detailed testing of the effect of different types of initial state has been restricted to ρ values 0.8, 0.9 and 0.95. As Qsim can generate equilibrated and exact initial states only for integral values of L_0, and for $\rho \leq 0.7$ equilibrium is reached very rapidly, ρ=0.8 (L_e = 4) is the lowest value of ρ for which a practically interesting range of cases can be generated.

Examples have been generated for the following initial states:

- exact initial queue size - zero variance
- ‘small’ initial variance - where the variance is broadly similar to the mean
- ‘medium’ initial variance - where the variance is around half the equilibrated value
- initial variance corresponding to equilibrated initial queue

Figure 4.3.15 shows how Ω_0 (calculated), Ω_0 (hand-optimised) and Ω_∞ (calculated) vary against L_0. There appears to be a common pattern covering all cases, with a tendency for the calculated value to exceed the optimum, except when equilibrated (right). Figure 4.3.16 plots Ω_0 optimised against calculated, showing deviations from linearity, while Figure 4.3.17 shows trends against the ratio V_0/L_0, showing differences in trend behaviour. The differences matter because the results can be sensitive to Ω_0.

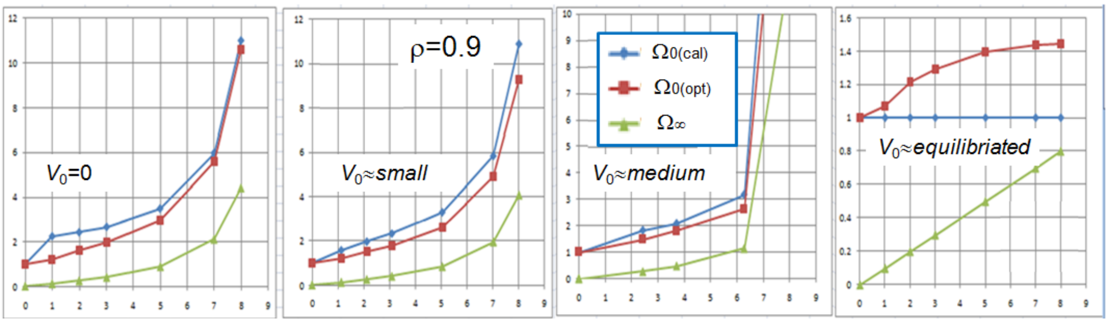


Figure 4.3.15 Ω values for various initial states (defined on previous page)

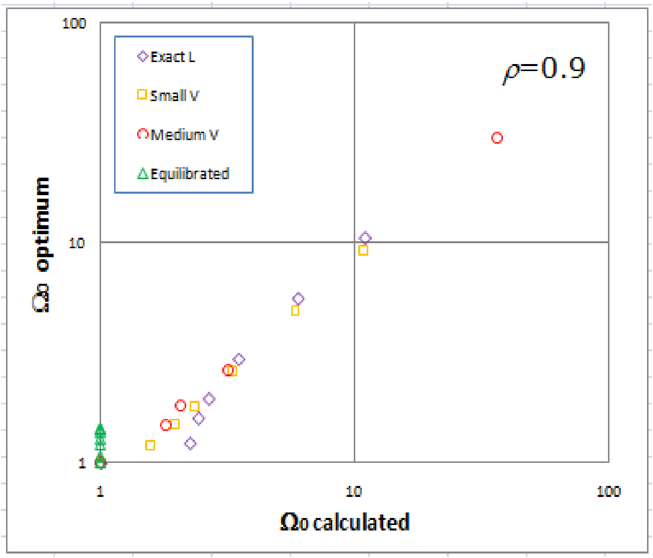


Figure 4.3.16 Fit of calculated Ω_0 for various initial states

When a single factor in the range 0.81-0.84 replaces individual hand-optimised factors, error is increased by 2 to 3 times. Detailed regression of the relationship between Ω_{opt} and Ω_{cal} does not appear worthwhile, but obvious features that could affect accuracy ought to be accounted for. The difference between the four initial cases lies in the shapes of their queue size probability distributions, which are Delta, Normal-like, Poisson-like, and geometric respectively.

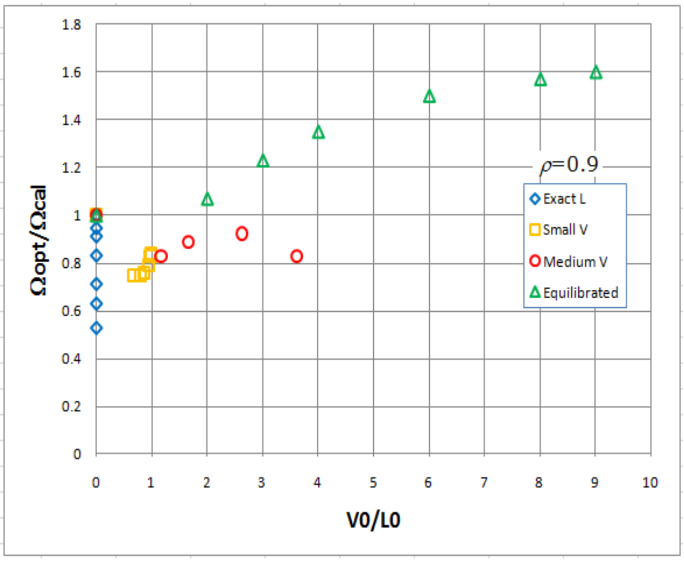


Figure 4.3.17 Ω_0 versus variance ratio for various initial states

The geometric distribution differs fundamentally from the other probability distributions in that it has mode=0, while all the others have mode>0. For this reason, one way to separate cases might be to estimate parameters of the Gamma distribution (2.2.6-7), even if it is not the most appropriate form. Alternatively, equilibriated or near-equilibriated cases can be identified by their relatively large initial variance. Various functions of L_0, V_0 and ρ have been tested to separate cases and to linearise the dependence of the Ω-ratio, and the following appear to work, as evidenced by Figure 4.3.18:

$$V_0 \geq L_0^2: \quad \frac{\Omega_{opt}}{\Omega_{cal}} = 1 + \frac{2}{3}\rho^* \quad \text{(at or near equilibrium)}$$

$$0 < V_0 < L_0^2: \quad \frac{\Omega_{opt}}{\Omega_{cal}} = \frac{3}{4} + \frac{1}{8}\rho^* \quad \text{(typical cases)} \qquad (4.3.27)$$

$$V_0 = 0: \quad \frac{\Omega_{opt}}{\Omega_{cal}} = \frac{3+4\rho^*}{7} \quad \text{(exact initial state)}$$

where a normalised demand intensity parameter is defined by:

$$\rho^* = \frac{2L_0/L_e}{L_0/L_e + 1} = \frac{2L_0}{L_0 + L_e} \qquad (4.3.28)$$

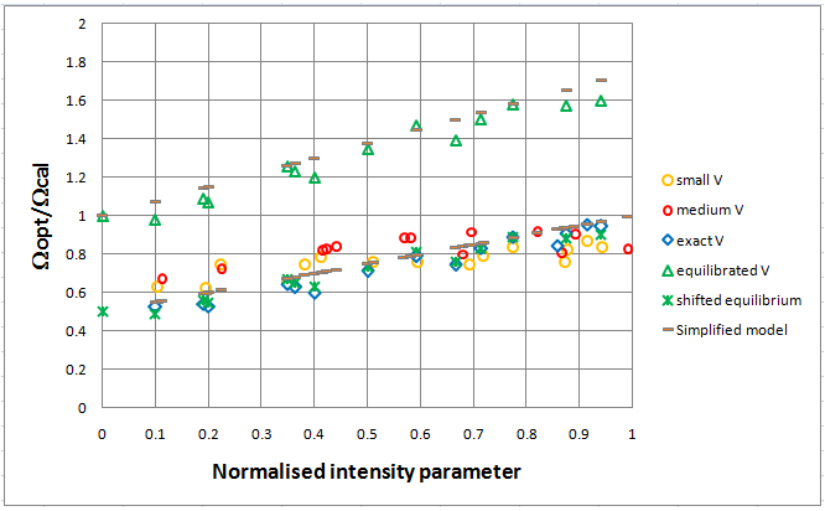


Figure 4.3.18 Linearised trend model for Ω adjustment factor when ρ<1

The impact of the corrections can be judged from Figure 4.3.19, which shows the fit of variance values using optimised, calculated and adjusted Ω_0 respectively. While the corrections cannot achieve the quality of fit of individually optimised Ω_0 values (left), the corrected parameters (right) are clearly better than the uncorrected ones (middle). While it is disappointing that the theoretical formulae alone do not achieve accuracy, and the corrections are not perfect, this could not reasonably be expected given the complexity of queue processes.

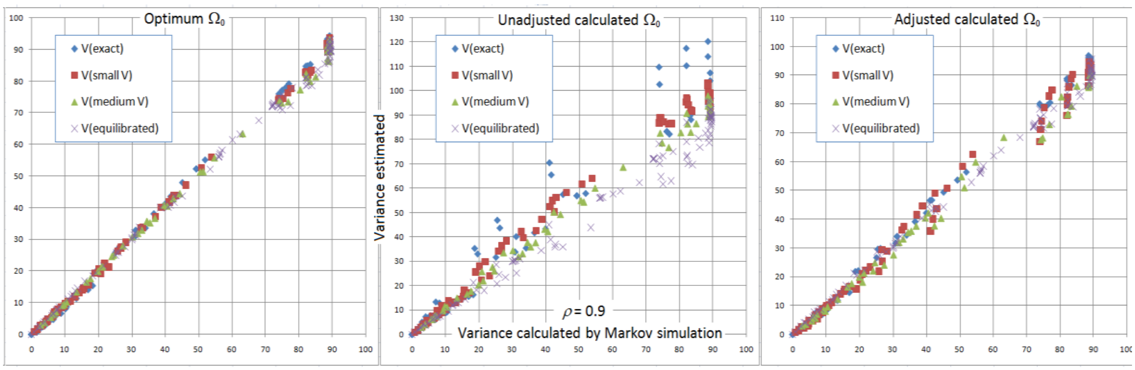


Figure 4.3.19 Effect of Ω_0 adjustment on variance estimates for ρ=0.9 cases

With origin-time shift, the Ω-correction is still possible using the same adjustment formulae, but errors in tests with ρ=0.9 are somewhat greater, as shown in Table 4.3.2.

Table 4.3.2 RMS errors with Ω-corrected models, ρ=0.9

Method	**Queue *L***	**Delay *D***	**Variance *V***
Sheared with L_0	0.064	0.049	0.266
Sheared with t_0	0.096	0.087	0.409

4.3.9 Definition of queue growth methods and labelling of combinations

In later Sections alternative methods of calculating queue development, and specifically queue growth, are coded as shown in the following Table 4.3.3:

Table 4.3.3 Definition of approximation methods for queue growth

Symbol	**Method**
s	Basic sheared queue and delay
t	Origin-shifted sheared queue
d	Sheared delay, derived sheared queue
z	Origin-shifted derived sheared queue
c	Corrected derived sheared
k	Corrected origin-shifted derived sheared

These may be combined in various ways to calculate undersaturated growth, oversaturated growth, and decay, plus the two transitions between these regimes. So a method combination is labeled ‘abcde’ where ‘-‘ can replace a transitional method (b or d) to represent defaulting to the method in an adjacent main regime. Section 4.5 later illustrates and discusses these regimes further, once methods for oversaturated growth and decay have been determined.

4.4. OVERSATURATED GROWTH REGIME

4.4.1 Behaviour relative to a deterministic asymptote

The asymptotic behaviour of the queue when ρ>1 is different from ρ<1 because utilisation can never exceed 1. By Taylor expansion of the inverse-time sheared queue function, as in the previous Section, the asymptotic value of *x* is found to be:

$$x\big|_{z\to 0} = 1 + \frac{I^* + (C - I)\min(\rho,1)}{(1-\rho)\mu t} \qquad (4.4.1)$$

Like (4.3.9b), this is trivially consistent with the deterministic queue formula except for the contribution of the RH term, so that:

$$L_s\big|_{z\to 0} = L_0 + (\rho - 1)\mu t - \frac{I^* + (C - I)\min(\rho,1)}{(1-\rho)} \qquad (4.4.2)$$

The *min*() term has been retained, but since ρ>1 its value is simply 1 and the statistical expression is evaluated using the definition of I^* and the relationship between I_a and c_a in Section 3.4. This gives the result in a revealing form:

$$L_s\big|_{z\to 0} = L_0 + (\rho - 1)\mu t + \frac{C^*}{(\rho - 1)} \text{ where } C^* = \tfrac{1}{2}\left(c_a^2 + c_b^2\right) \qquad (4.4.3)$$

Thus the queue (approximation) never ‘forgets’ its initial value nor its deviation from linearity during initial growth (which the above equations ascribe to randomness but could also be influenced by the initial utilisation). As the time scale is zoomed out the queue graph looks increasingly linear, so $L_s(t/2)$ is expected to be a good approximation to the ‘delay’ associated with mean queue L_s, possibly with a modified third term. Figure 4.4.1 is an example of equations (4.4.3) in action. This shows the devation from the deterministic asymptotes of queue and delay calculated by Markov simulation (‘sL’, ‘sD’) and sheared method (‘sLs’, ‘sLs/’, where ‘/’ indicates that *t* is replaced by *t*/2). The calculated functions tend to a value not far from the theoretical final value. However, as all the functions start and finish at the same values, any correction to the *difference* from the asymptote function must be non-monotonic, meaning that for example estimation methods based on an exponential function cannot be used.

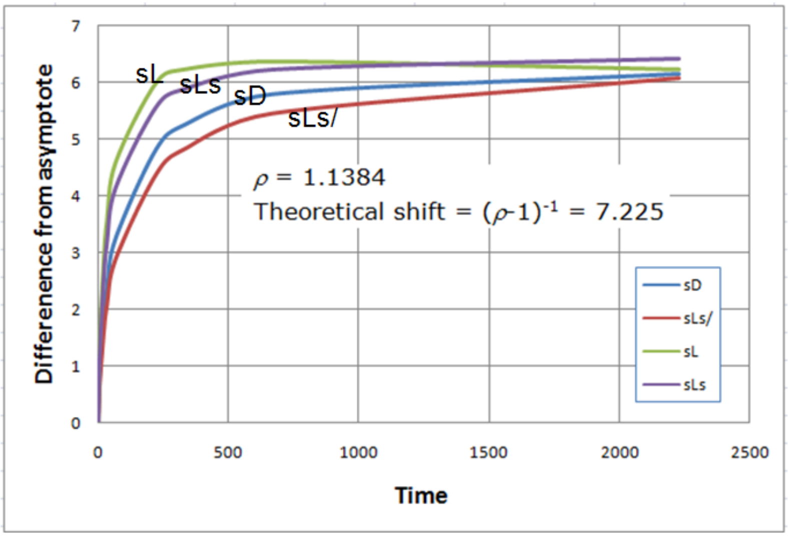


Figure 4.4.1 Difference from deterministic asymptote for various estimation methods

4.4.2 Ω-correction for oversaturated growth

A similar empirical analysis to that of the previous Section has been done for 21 oversaturated growth cases, with ρ in the range to 1.0151-1.3, L_0 ranging from 0 to 100 (mostly at the lower end), and variances ranging from zero to equilibrated. The results are shown in Figure 4.4.2, and equations (4.4.4).

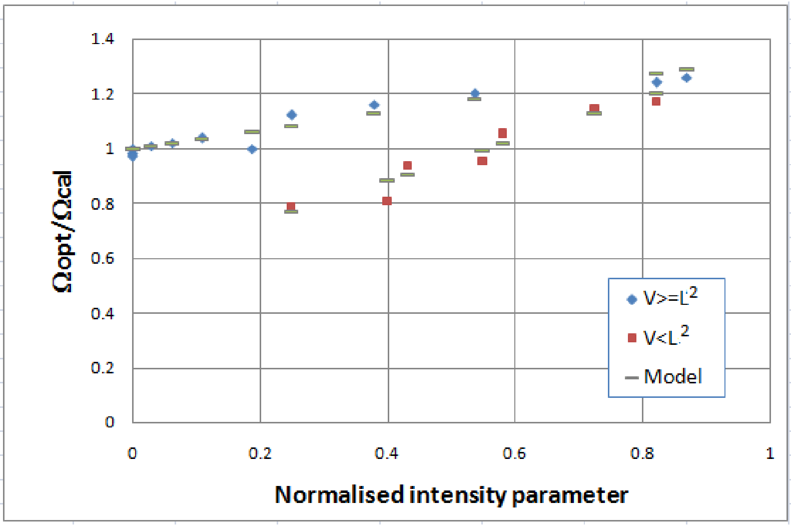


Figure 4.4.2 Linearised trend estimate of Ω ratio for ρ>1

$$V_0 \geq L_0^2: \qquad \frac{\Omega_{opt}}{\Omega_{cal}} = 1 + \frac{1}{3}\rho^*$$

`

$$V_0 < L_0^2: \qquad \frac{\Omega_{opt}}{\Omega_{cal}} = \frac{7+9\rho^*}{12} \qquad (4.4.4)$$

The impact of the correction is evident in Figure 4.4.3. This correction applies to the basic sheared approximation calculated for an extended range of time points from a common origin, not to the derived version, nor to a time-sliced calculation.

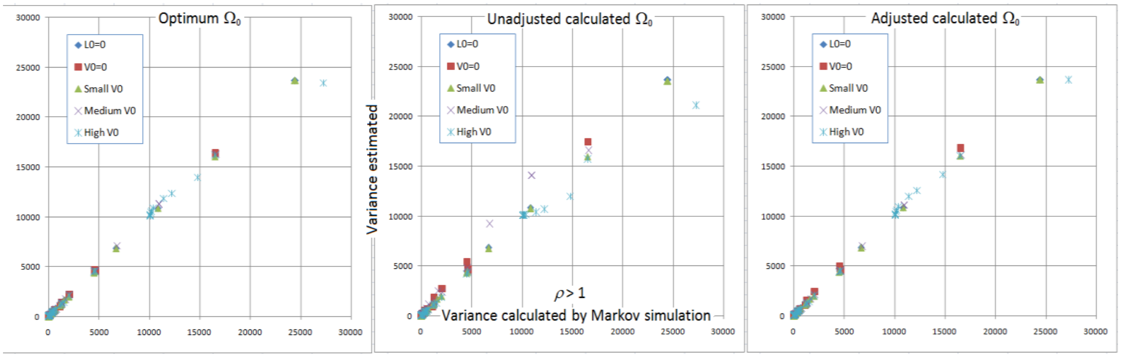


Figure 4.4.3 Effect of Ω_0 on variance estimates for ρ>1 growth test cases

This correction does not work well for the 34 peak test cases. They differ from the 21 pure growth cases in several respects:

- Initial states generated by a complex growth process
- ρ values change with time
- Transitions between ρ<1 and ρ>1
- Relatively short growth times
- Only 2 cases where $V_0>L_0^2$

The 34 peak cases contain a total of 163 time slices where ρ>1, with ρ ranging from 1.0061 to 1.1458. It is found empirically that in each peak case a certain set of optimal factors applied to Ω_0 achieves minimum error (measured for V which is the most sensitive result). These values are mostly near to but not exactly the same as an average values for each peak case. With Ω_∞=1 the effect of the time-dependence (4.3.16c) is small. There is no apparent relationship between the empirical correction factors and V_0/L_0^2. Regressing scatter relationships between the factors and the normalised variable ρ^* defined by equation (4.3.27), as shown in Figure 4.4.4, yields the fairly flat formulae (4.4.5) (where regression coefficients have been rounded). However, the results seem unduly sensitive to the values of the correction factors, and better results appear to come from simply using the origin-shifted sheared ‘t’ method.

$$\rho< 1 \text{ in previous t/s:} \qquad \frac{\Omega_{opt}}{\Omega_{cal}} = 0.93 + 0.1\rho^*$$

$$\rho>1 \text{ in previous t/s:} \qquad \frac{\Omega_{opt}}{\Omega_{cal}} = 0.82 + 0.2\rho^* \qquad (4.4.5)$$

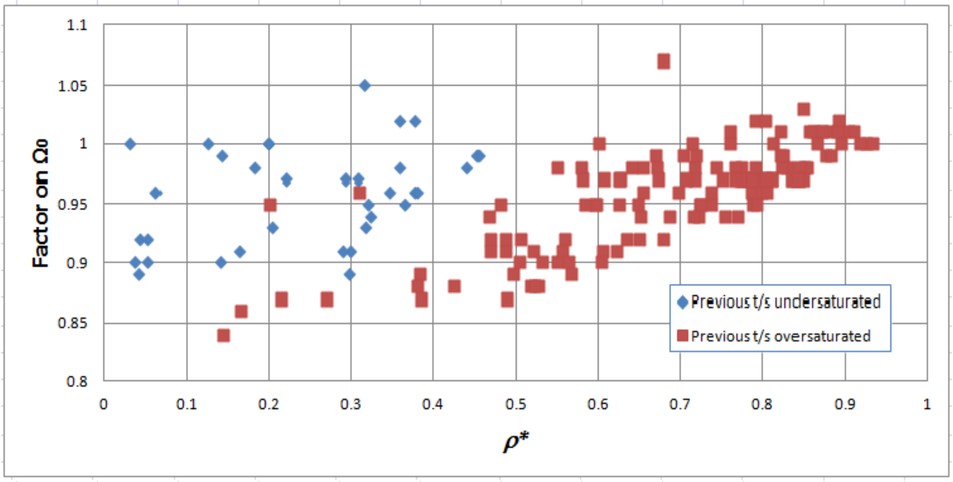


Figure 4.4.4 Correction factors to Ω_0 for $\rho>1$ time slices in peak cases

4.4.3 Transition through saturation and long-term growth

The behaviour of queues is complex, so any simplified or partly empirical approximation must be subject to a limit of performance. At some point it is necessary to stop experimenting and settle on a method that is *sufficiently* good for practical purposes. Even so, experiments with peak cases suggest that performance can be improved by replacing method ‘t’ by ‘k’ in the early oversaturated time-slices, reducing error in the variance. The effect decreases as more time slices are converted, and working from the late end of the oversaturated period is much less effective. To investigate the consistency of this effect a case of pure oversaturated growth has also been tested. In Figure 4.4.5 a queue is growing from an initial equilibrated size of 20 under heavy oversaturation at ρ=1.3. Although plotted on a linear time scale it is calculated in logarithmic time steps so the time slices keep getting longer.

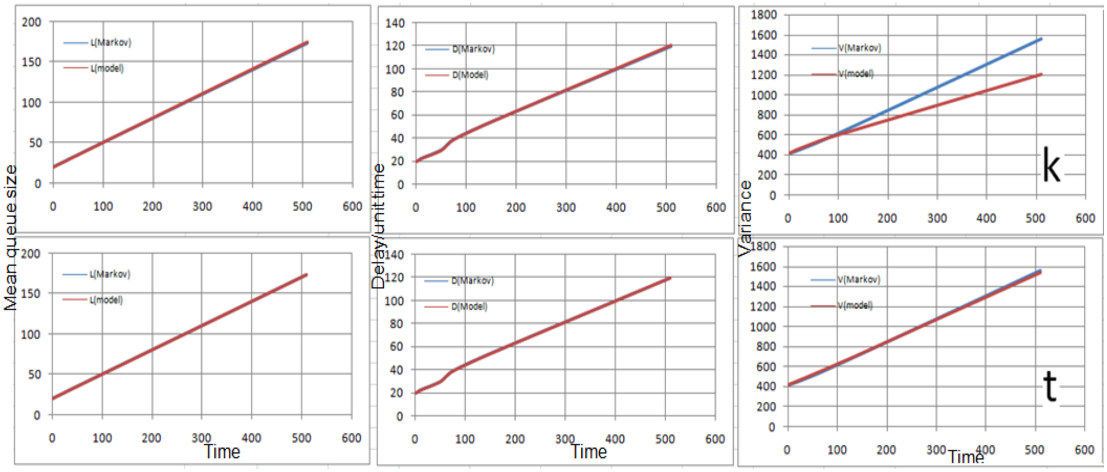


Figure 4.4.5 Oversaturated growth at ρ=1.3, μ=1 from equilibrated initial queue size 20 using two different calculation methods

At first sight it appears that only method ‘t’ produces an accurate result, and method ‘k’ fails to an unacceptable degree (other methods are worse). However, the methods start to diverge only after a time, in this case equivalent to throughput of around 100 (μ=1). That this is not just an appearance can be confirmed by plotting the left end of the graph on an extended time scale. This confirms that the ‘k’ method can be beneficial in the oversaturated transition regime.

If this is a consequence of some ‘relaxation’ process then one would expect it to be related to throughput, not time slices or time. Figure 4.4.6 plots the benefit, in terms of reduction in variance error, of using method ‘k’ in just the first oversaturated time slice in all the peak cases against the throughput in the time slice (the pattern is similar when plotted against the total throughput in oversaturation).

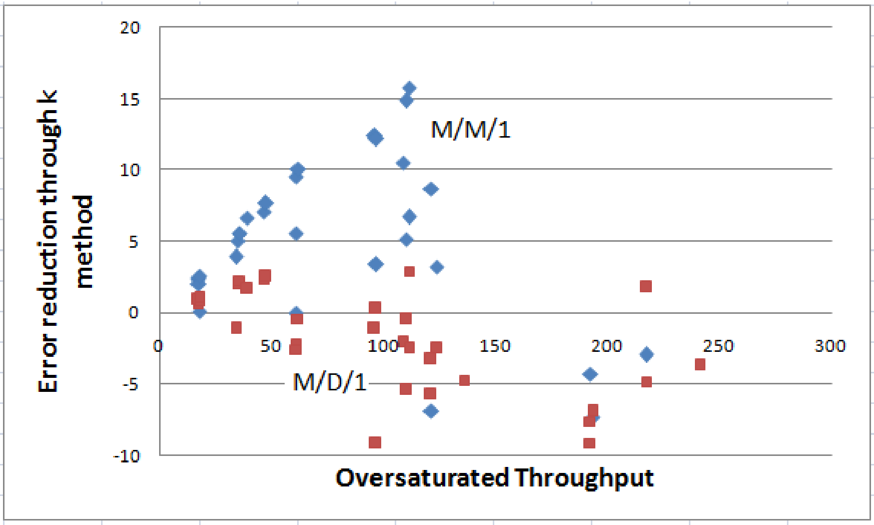


Figure 4.4.6 Benefit of using method ‘k’ in the first oversaturated time slice only

There is no benefit for M/D/1 cases, but there does appear to be a benefit for M/M/1 cases, although it hits a ‘cliff’ at throughput similar to the above. The optimum value to minimise average variance error is 130, but the origin of this number is obscure. In the straight growth case, Figure 4.4.5, error becomes apparent only when method ‘k’ is applied for a total throughput greater than somewhere between 82 and 158, which is consistent with the above value. It is concluded that if over-saturated time slices are estimated by method ‘k’ where the cumulative throughput is less than 130 units, average error is reduced and performance in the lighter peaks visibly improved.

In some cases error can be reduced by applying method ‘k’ in the *last undersaturated* time slice before the oversaturated growth regime. In all such cases the value of ρ is large (0.9463-0.9925), but in some other cases where ρ is in this range accuracy is reduced, so it is unclear what objective criterion could be used to select the best method.

4.5. EXPONENTIAL DECAY METHOD BASED ON GENERAL PROPERTIES

4.5.1 Motivation and approach

The sheared queue method performs least well in representing the decay of queues, especially large queues, as pointed out by Kimber and Hollis (1979) who substituted a combination of a linear approximation and a mirror-image of the growth formula. This might be because of greater departure of the queue size probability distribution from that implied by the quasi-static model. A possible reason for a *qualitative* difference is that decay is dominated by the queue itself rather than a balance between input and output, making it more uni-directional. A queue that can substantially exceed the equilibrium size corresponding to the traffic intensity, can be viewed as applying a proportional 'pressure' driving its own discharge. If previous growth has been undersaturated, it does not follow that the probability distribution will be equilibrium-like, because to produce a queue of any substantial size the traffic intensity must be high, more than 0.9 say, so the growth time could be much less than the stochastic relaxation time. A Normal-like distribution is more likely in practice, so that variance and utilisation have a smaller role than drift of the mean.

Since rate of discharge is limited by capacity, the decay rate of a large queue well above the equilibrium value assocated with the current demand will be virtually constant, agreeing with the linear regime of the Kimber and Hollis (1979) method. However, the rate must eventually fall off as the queue declines. In principle, the sheared formula could then be applied 'as is', but Kimber and Hollis found this to be inaccurate. They do not go deeply into their reasons for using an inverted form of the sheared function, although this appeare to have convenient properties, but point out that the relaxation of a queue is not simple (an unsurprising consequence of an extended probability distribution).

This on the face of it argues against using the next simplest function to linear, an exponential depending on the difference between the current and equilibrium queue sizes which would naturally converge to equilibrium. However, as with the growth approximations described earlier, any function describing queue decay must satisfy at least initial and asymptotic constraints. These will not generally be satisfied by the same function, so a pure exponential function, with a single fixed time constant, could not satisfy the extreme constraints in any case. Arguably, therefore, the next simplest function to try is $e^{f(t)}$, with the exponential providing global relaxation behaviour over time, and $f(t)$ embodying additional time-dependence to match the extremal constraints.

4.5.2 Structural properties of queue development – initial state

The sheared approximation deals with finite time intervals, and real queues involve discrete events, but queue behaviour can analysed theoretically as a continuous or piecewise continuous process. The rate of change of a queue is given by differentiating (2.3.3), giving (4.5.1). Utilisation is necessarily exogenous since it depends on the instantaneous queue size probability distribution which in turn depends on the past history of queue development.

$$\frac{dL}{dt} = (\rho - u)\mu \tag{4.5.1}$$

If the initial state of the queue is in equilibrium then the initial mean utilisation will satisfy the RHS of the P-K mean queue formula (4.2.5) at t=0 (since $x \equiv u$ there) and can be calculated from the initial queue by inverting that equation. This is the assumption made by Kimber and Hollis (1979) for large decaying queues, though they assume this rate remains constant until the queue has fallen to $2L_e$. If the initial state of the queue is an exact size >0 (a 'pure state' or 'shock'), then the initial utilisation is exactly 1. If the initial queue is large then utilisation will be very close to 1. A large queue is therefore initially in 'free fall' and stochastic relaxation behaviour as described by equation (2.3.38) or (2.3.39) has not properly begun. It should be possible to describe such as queue by drift with some diffusion (Newell 1982) as long as p_0 remains sufficiently small so the 'zero barrier' plays no part. The decay rate must eventually decline as the utilisation moves away from 1. The next simplest evolution is exponential, which actually applies to the very simple M/M/1/1 process (see earlier in Section 2.3). If, rather than assuming constant drift, the queue function is assumed to evolve exponentially, at least over the time period of interest, the following can satisfy both initial and equilibrium constraints:

$$L_{(1)}(t) = L_e + (L_0 - L_e)e^{-t/\tau} \tag{4.5.2}$$

Differentiating (4.5.2) and using (4.5.1) the initial value of 'characteristic time' τ in (4.5.2) that satisfies the initial condition is:

$$\tau_i = \frac{L_e - L_0}{(\rho - u_0)\mu} \tag{4.5.3}$$

Equation (4.5.3) can be interpreted as the time scale of initial nearly linear decay. Equation (4.5.2) is readily integrated using the definition of delay D (2.3.28), provided τ can be assumed to be so 'slowly varying' that its variation may be ignored:

$$D_{(1)}(t) \approx L_e + \frac{\tau}{t}(L_0 - L_e)\left(1 - e^{-t/\tau}\right) = L_e + \frac{\tau}{t}(L_0 - L) \tag{4.5.4}$$

This needs slightly careful handling because as it stands the second term becomes singular at *t*=0, but by expanding the exponential for small *t* it is confirmed to have the initial value L_0, as well as asymptotic value L_e. While it is not obvious that equation (4.5.4) is approximately equal to the average of L_0 and L in the neighbourhood of *t*=0, or is generally close to *L*(*t*/2), both these can be shown to be true.

If equation (4.5.4) is taken at face value as the *definition* of *D*, as there is no *a priori* reason to assume that (4.5.2) will turn out to be the most accurate formula, an alternative queue can be obtained as the derivative, equation (4.5.5), which differs from (4.5.2) by the presence of the term involving the derivative of τ. Because this vanishes at *t*=0, the initial value of τ given by equation (4.5.3) remains valid.

$$L_{(1.5)}(t) = L_e + (L_0 - L_e)e^{-t/\tau} + (L_0 - L_e)\left(1 - e^{-t/\tau}\right)\dot{\tau} \tag{4.5.5}$$

It is also possible to start by defining *D* using a formula analogous to (4.5.2), with *L* expressed in terms of its derivative like the ‘derived queue’ in Section 4.2, with or without a term involving the derivative of τ. Then (4.5.2,4) are replaced by (4.5.6,7) repectively:

$$L_{(2)}(t) = L_e + \left(1 - \frac{t}{2\tau}\right)(L_0 - L_e)e^{-t/2\tau} + \frac{t^2}{2\tau^2}(L_0 - L_e)\left(1 - e^{-t/2\tau}\right)\dot{\tau} \tag{4.5.6}$$

$$D_{(2)}(t) = L_e + (L_0 - L_e)e^{-t/2\tau} \tag{4.5.7}$$

Finally, if the complexity of (4.5.6) is considered too much, the option exists to adopt $L_{(1)}$ and $D_{(2)}$ together, and work around any resulting inconsistency. Having said that, it is accepted that achieving maximum accuracy of and consistency between *L* and *D* and, by implication variance, without excessive complexity, is an issue that may benefit from further research.

4.5.3 Structural properties of queue development – asymptotic state

Asymptotically, the queue's evolution must satisfy the variance equation (2.3.27). Substituting (4.5.4) into (2.3.29) and setting t=∞ yields the asymptotic characteristic time:

$$\tau_a = \frac{W_e - W_0}{2(1-\rho)(L_e - L_0)\mu} \qquad (4.5.8)$$

4.5.4 Unsuitability of exponential method for growth

Although derived with decaying queues in mind it is natural to ask whether the whole range of growth and decay could be covered by an exponential method. The method is essentially *deterministic*, with the type of queuing process entering only through steady-state invariants, avoiding the conceptual issues of the quasi-static assumption in shearing. If $L_{(1.5)}$ or $L_{(2)}$ is adopted then, since L_0<L_e, the $\dot{\tau}$ term is negative, which could result in undershoot. However, this does not affect $L_{(1)}$ so its derivative, and hence u, should still behave sensibly. In practice this is not pursued for three reasons: (1) the demonstrated performance of the corrected sheared approximation; (2) the qualitative difference between queue growth and decay; (3) experiemental results showing that the exponential method performs poorly for growth.

4.5.5 Relaxation behaviour of the characteristic time parameter

The next task is to find a time-dependent function $\tau(t)$ to cover the whole range of t, that satisfies the extremal conditions, preserves structural integrity, and is good enough for practical purposes. The function $\tau(t)$ should interpolate τ_i and τ_a smoothly and ideally monotonically. What is the evidence from actual queues?

Initially, queue decay evolution has been calculated for four values of ρ: 0.1, 0.5, 0.8, 0.9, with μ=1, and one of two initial states:

- Equilibrated queue generated by Markov simulation with ρ=0.95, leading to an actual mean queue size of 18.982 corresponding to slightly reduced ρ=0.949955, and corresponding variance.

- Exact initial queue size of 19 with zero variance.

Figure 4.5.1 shows that the queues evolve in a qualitatively similar way against a time scale based on the stochastic relaxation time for each ρ value (equation 2.3.38).

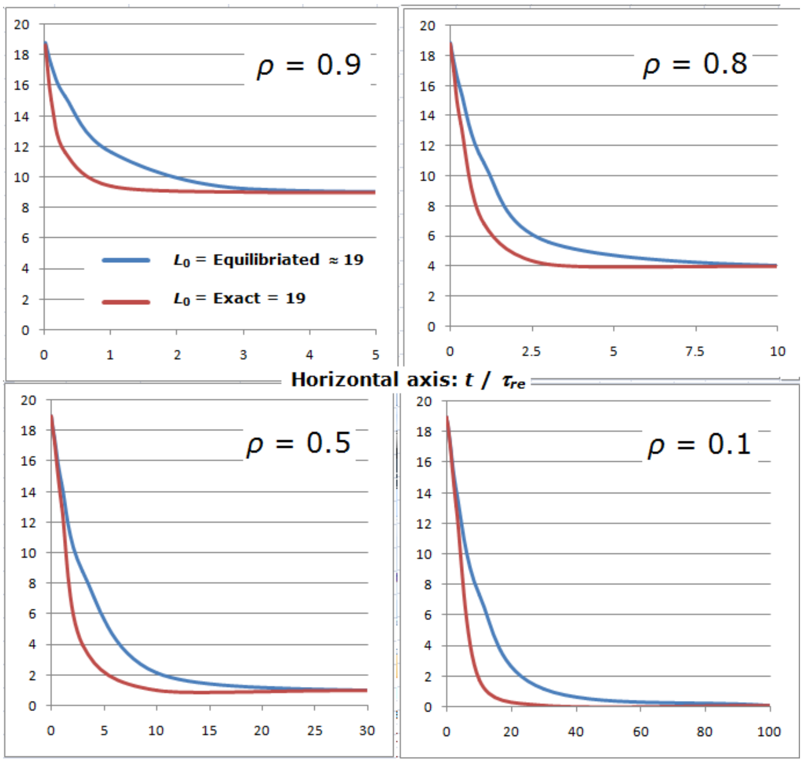


Figure 4.5.1 Evolution of queue size for different traffic intensities and initial states Rates of decay depend on whether the initial queue is exact (steeper) or equilibrated (gentler).

Relaxation and characteristic times are given in Table 4.5.1. The times at which the graphs cease to be practically distinguishable are not obviously related to stochastic relaxation times τ_{re}. However, it appears from the relationships of τ_i to the corresponding τ_{re} that if time were instead scaled according to τ_i the horizontal axes would all be around the same length (~10), supporting the view that τ_i is the critical time constant in the initial relaxation process.

Table 4.5.1 Calculated relaxation and characteristic times (L_0≈ 19, μ=1)

Process		**Equilibrated initial queue**		**Exact initial queue**	
ρ	τ_{re}	τ_i	τ_a	τ_i	τ_a
0.9	379.74	199.82	289.82	100.00	100.00
0.8	86.72	99.91	119.91	75.00	56.67
0.5	11.66	39.96	41.96	36.00	20.89
0.1	2.14	22.20	22.33	20.99	11.17

On inverting (4.5.2), equation (4.5.4) and the variance formula allow alternative τ values to be back-calculated or reconstituted from simulated queue values:

$$\tau_{\_L} = \frac{t}{\ln\left(\frac{L_0 - L_e}{L - L_e}\right)} \tag{4.5.9}$$

$$\tau_{\_D} = \left(\frac{L_e - D}{L - L_0}\right)t \tag{4.5.10}$$

$$\tau_{\_V} = \frac{W - W_0}{2(1-\rho)(L - L_0)\mu} \tag{4.5.11}$$

It is necessarily the case that $\tau_{\_V}$=$\tau_{\_D}$ whatever the method of calculation, but it is not necessary that $\tau_{\_D}$=$\tau_{\_L}$ except in the case of the exponential queue model, since a different queue function would in general lead to a different detailed relationship between *L* and *D*.

Equation (4.5.9) can run into difficulty if the benchmark fails to converge exactly to the theoretical equilibrium queue, so in preference to the theoretical L_e value the benchmark mean queue at the largest *t* value tested (=100τ_{re}) is used. This in turn can result in the logarithm being undefined, so limiting the range of values of *t* for which $\tau_{\_L}$ is estimated.

Figure 4.5.2 shows that when only valid ranges are plotted, the values of $\tau_{\_L}$ are only broadly consistent with the theoretical values given in Table 4.5.1. For the equilibriated cases, dependence on time is quite weak, and two of the cases with exact initial queue appear not to be monotonic. No firm conclusion can be drawn from these rather variable results.

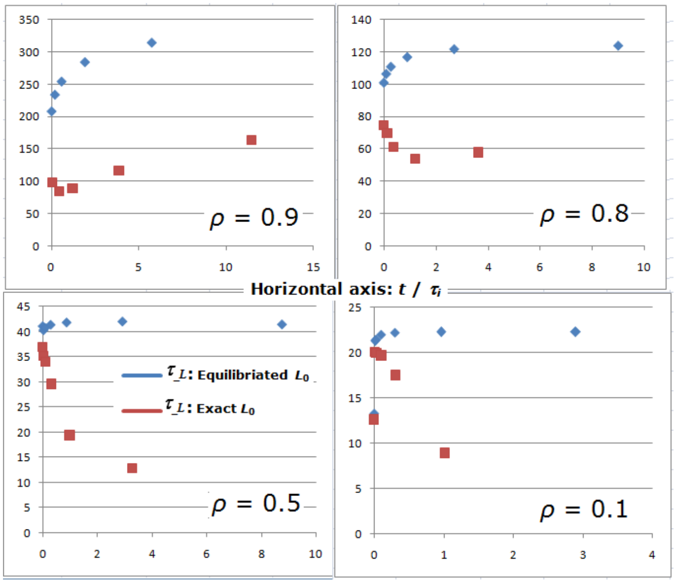


Figure 4.5.2 Characteristic time values estimated from benchmark queue sequences

4.5.6 A possible refinement to allow for non-monotonicity

Abate and Whitt (1987) point out that a queue starting from a range of initial values, including some greater than its equilibrium value, can fall below it before climbing up again, so the function of time is non-monotonic, something the exponential approximation cannot accommodate. They suggest that a queue's moments can be modelled using two components as in equation (4.5.12) both of which *are* monotonic, so if the second term in particular can be approximated, this problem might be avoided.

$$L(L_0, t) \equiv L(0,t) + \left(L(L_0, t) - L(0,t)\right) \qquad (4.5.12)$$

However, the effect is most pronounced when a queue starts from an initial pure state, i.e. is associated with low initial variance. Although this is one of the tests above, it is relatively uncommon in traffic modelling and especially unlikely in the case of a decaying queue, that thanks to randomness during growth is likely to begin from an extended probability distribution. For this reason such refinement is left for potential future research.

4.5.7 Testing various exponential interpolations

Continuing the empirical approach for the sake of simplicity, several alternative interpolation functions based on the characteristic times (4.5.3, 4.5.8) and the generic exponential queue approximation (4.5.2) have been tested:

$$L_i(t) = L_e + (L_0 - L_e)e^{-t/\tau i} \qquad (4.5.13)$$

$$L_a(t) = L_e + (L_0 - L_e)e^{-t/\tau a} \qquad (4.5.14)$$

$$L_x(t) = L_a + (L_i - L_a)e^{-t/\tau i} \qquad (4.5.15)$$

$$L_y(t) = L_a + (L_i - L_a)e^{-t/\tau a} \qquad (4.5.16)$$

$$L_m(t) = L_e + (L_0 - L_e)e^{-t/\tau m} \qquad (4.5.17)$$

$$L_n(t) = L_e + (L_0 - L_e)e^{-t/\tau n} \qquad (4.5.18)$$

The additional compound time parameters are defined by equations (4.5.19-20), and (4.5.15) can be rewritten in the form of (4.5.2) if τ is replaced by τ_x as defined by (4.5.21):

$$\tau_m(t) = \tau_a + (\tau_i - \tau_a)e^{-t/\tau i} \qquad (4.5.19)$$

$$\tau_n(t) = \tau_a + (\tau_i - \tau_a)e^{-t/\tau a} \tag{4.5.20}$$

$$e^{-t/\tau x} = e^{-2t/\tau i} + e^{-t/\tau a}\left(1 - e^{-t/\tau i}\right) \tag{4.5.21}$$

As Figure 4.5.3 shows for the J2P4 peak case (defined earlier in Section 2.5), several interpolations produce wild excursions, including negative values in Ts 4-7[45] where ρ>1. Figure 4.5.4 shows that greater excursions occur in variance estimates (calculated from 2.3.27). L_n is stable only in the decay regime. L_m, calculated using a 'mixed characteristic time' τ_m, appears to produce a smooth stable graph and a close fit to the true mean queue and variance profiles, although its value in Ts 3 is low, making it second best to L_x in the decay regime. In the peak regime Ts 4-7, when ρ>1, L_m achieves the lowest RMS error, while other interpolations are unstable. Both of these methods have in common that they use τ_i to interpolate the characteristic times.

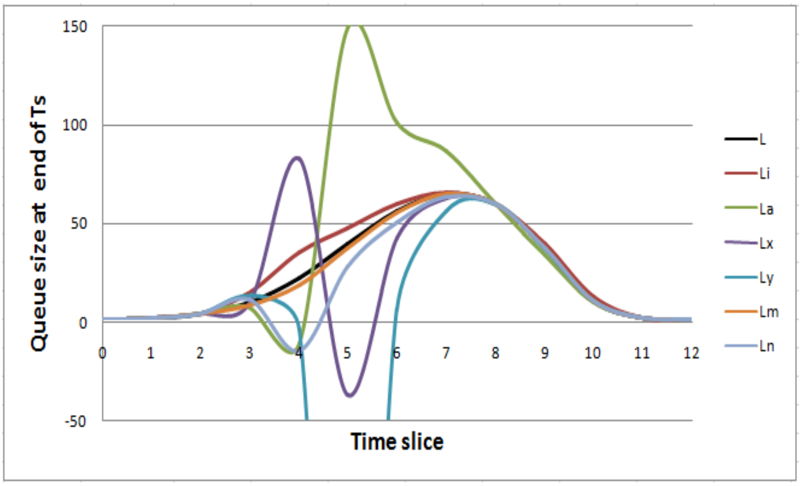


Figure 4.5.3 J2P4 peak case: queue size profiles estimated by various models

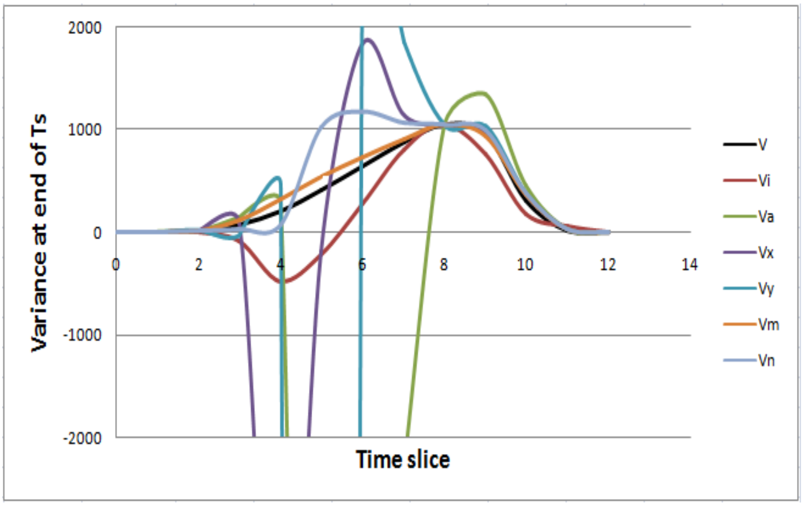


Figure 4.5.4 J2P4 peak case: variance size profiles estimated by various models

[45] Time slice 0 is used to set up the initial equilibrium state.

Table 4.5.2 Best interpolation methods for J2P4 – second best in brackets where error <2x

Best model in regime	**Growth**	**Peak**	**Decay**	**Growth+Decay**
	Ts 1-3, ρ<1	Ts 4-7, ρ>1	Ts 8-12, ρ<1	Ts: ρ<1
L	x	m	n(x)	x
D	x	m	m(x)	x
V	x	m	m(x)	x

To get a sense of the error involved in basing this on the simple $L_{(1)}$ approach (4.5.2), the derivative of τ_m can be calculated from the definition (4.5.17):

$$\dot{\tau}_m = \frac{\tau_a - \tau_m}{\tau_i} \qquad (4.5.22)$$

Assuming τ_m to be monotonic, this could make a significant contribution to (4.5.5) through the $\dot{\tau}$ term at middle values if *t*, but at least it is simple to evaluate. However, as Figure 4.5.5 (next page) shows[46] (in this peak case at least), it leads to serious distortion of the queue profile 'L(alt)', compared to the simple method labelled 'L(model)', suggesting that (4.5.4) might not be a good approximation to delay if τ has the degree of variability indicated by (4.5.22) and the $\dot{\tau}$ term in (4.5.5) leads to unacceptable error.

4.5.8 Practical application of simple exponential interpolations of characteristic times

The quandary can be avoided by adopting $L_{(1)}$ with $D_{(1)}$. Figures 4.5.3-4 show that, surprisingly, the method appears to be accurate in Ts 4-7, even though it is not expected to work when ρ>1. Furthermore τ_m can go negative, though this may be acceptable where relaxation to equilibrium is not involved, and it may also help that the time slices are relatively short (9 minutes) compared to the magnitude of τ_m (at least 22 minutes).

Two particular kinds of issue can arise in practical application to the decay regime:

- $\tau_i > \tau_a$, where intuitively $\tau_i \leq \tau_a$ is expected as τ_a relates to an asymptotic condition
- One or both of τ_a, $\tau_i < 0$

While $\tau_i \leq \tau_a$ is true for lighter peaks like J1P1, it turns out that in heavy peak cases like J3P9, $\tau_i > \tau_a$ is more usual. In intermediate cases like J2P4, $\tau_a \approx \tau_i$.

[46] To achieve the close fit at around 54 minutes it is necessary to take the initial state values from the Markov simulation up to that point, rather than from iterative calculations in which the result in each time-slice provides the input for the next. After 54 minutes the model using the Markov initial states makes little difference. As predicted at the outset, the exponential model performs poorly in the early growth part of the peak.

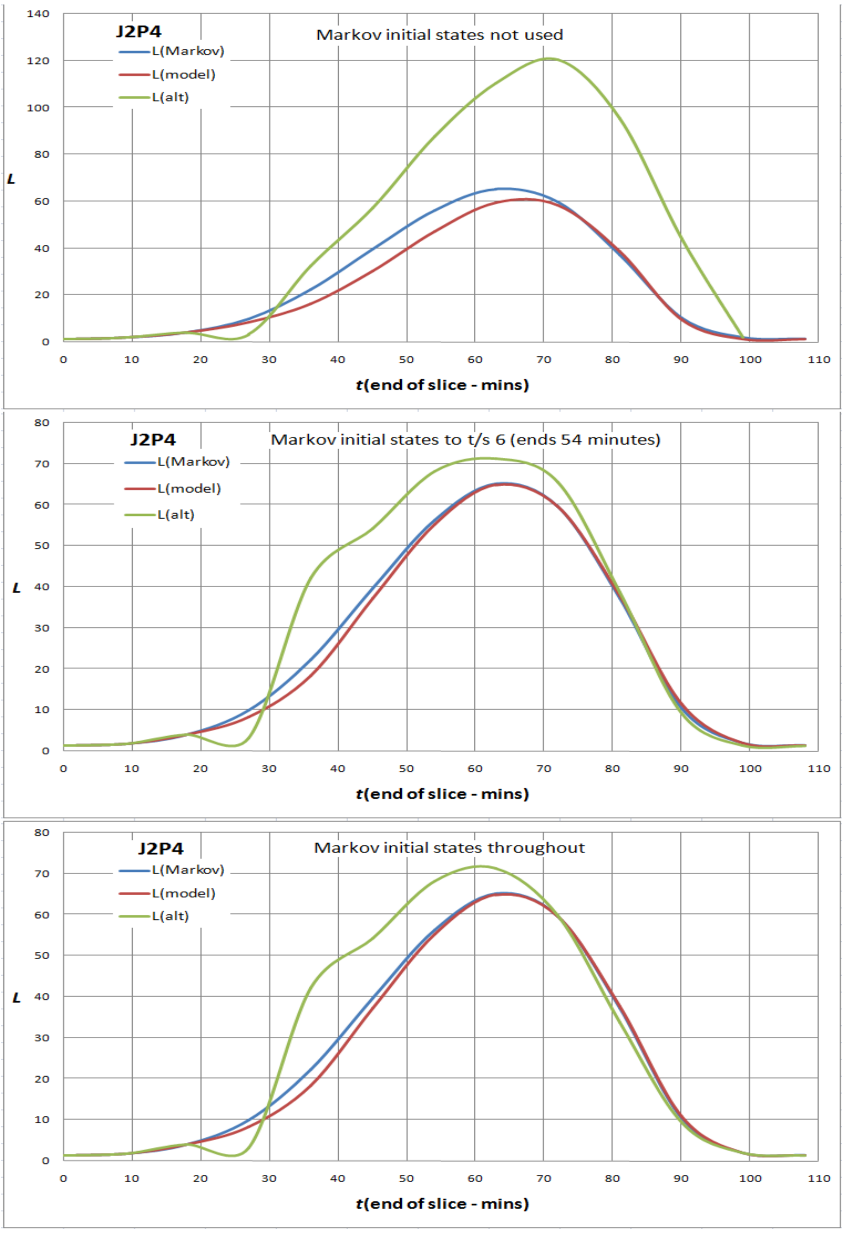


Figure 4.5.5 Effect on queue of initial states and (alt) added term in equation (4.5.5)
('alt' graphs are different from the other two similar graphs)

So there is a pattern that appears rational. Initially, a time parameter calculated as the average of τ_i and *min*(τ_i,τ_a) was tried on the basis that it *ought* to be closer to τ_i than to τ_a. However, this was found unsatisfactory, and would actually be illogical in cases where neither τ_i nor τ_a is consistently greater. On the principle that a working parameter should represent whichever relaxation process 'gets there first', (4.5.18) is modified to:

$$\tau_m(t)=\tau_a+(\tau_i-\tau_a)e^{-t/\min(\tau i,\tau a)} \qquad (4.5.22)$$

Negative values of τ_a, τ_i do not seem a problem for estimation of *L*, since the mean could increase as well as decrease, and *D* from the integral of *L* is meaningful in either case.

However, stability of V calculated from these quantities is not guaranteed, and this is more of an issue where $L_0 \approx L_e$ results in τ_a becoming highly sensitive. To avoid this problem, a simple exponential interpolation of V using the relaxation time τ_{re} has been attempted when either of τ_a, τ_i is negative, but this also proved unsatisfactory. However, cases with this problem are rare. Consequently certain basic measures may be adopted to protect the results:

- If $\tau_a < 0$, set $\tau_m = \tau_i$.
- Absolute value of $t/min(\tau_i,\tau_a)$ is limited to avoid overflow in the exponential function
- V_m is subject to a lower limit of $min(V_0,V_e)$

4.5.9 Comparative performance of decay estimates for run-out to equilibrium

Several methods of estimating queue delay have been compared against Markov simulation using the tails of two peak cases, starting in each case from two different initial states drawn from the full Markov simulations, and extended to run out to equilibrium at constant traffic intensity, as defined in Table 4.5.3.

Table 4.5.3 Definition of decay run-out cases

Case name	Starting point	Run-out ρ	Duration	Time-slices
J3P9x1	Markov values at start of decay	$\geq$0.8084	140 mins	10 mins
J3P9x2	Markov values at start of run-out	0.8084	120 mins	10 mins
J4P10x1	Markov values at start of run-out	0.9206	166.5 mins	4.5 mins
J4P10x2	Markov values at start of run-out	0.7480	157.5 mins	4.5mins

The average RMS errors of the four tests are given in Table 4.5.4, using the methods defined in Table 4.3.3 plus ‘m’ exponential.

Table 4.5.4 Average errors in decay run-out tests

Method	Error in *L*	Error in *D*	Error in *V*
s: Basic sheared queue and delay	3.75	4.16	1052.63
t: Origin-shifted sheared queue, $L_T = 1.5\ L_e$	7.30	6.78	1604.81
t: Origin-shifted sheared queue, $L_T = 2\ L_e$	6.17	5.82	1464.95
t: Origin-shifted sheared queue, $L_T = 2.5\ L_e$	5.25	5.11	1234.29
d: Sheared delay, derived sheared queue	4.90	4.62	1172.55
z: Origin-shifted derived sheared queue	23.06	25.93	7701.67
c: (Growth) Corrected derived sheared	123.80	113.18	76685.67
k: Corrected origin-shifted derived sheared	122.03	111.85	77830.03
m: Mixed exponential	0.91	0.78	132.63

The results for methods ‘c’ and ‘k’ are so poor that these methods are clearly inappropriate for decay – they are designed to correct growth. Of the pre-existing methods, surprisingly the original sheared method ‘s’ gives the smallest errors, but those of the exponential method ‘m’ are an order of magnitude smaller. This can be ascribed to its inherent conformance to the correct asymptotes, as well as possibly being structurally superior. Figure 4.5.6 shows differences in performance of the methods on three run out tests, compared to Markov simulated run out profiles, showing method ‘s’ is unsuitable for decay of a heavy queue (top right). Figure 4.5.7 plots mean, delay and variance of four run-out tests (the additional second case being a variation of the first with later resetting to a Markov initial state).

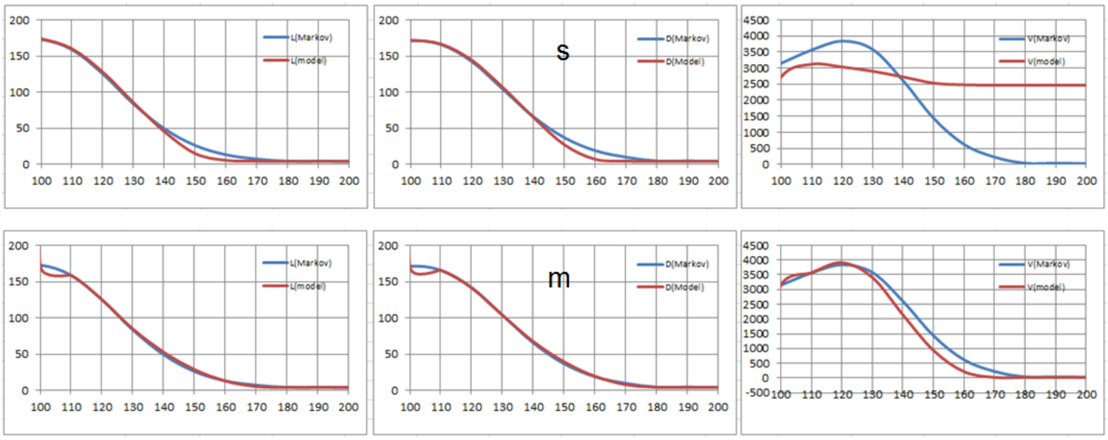


Figure 4.5.6 Results of run-out tests using methods ‘s’ (above) and ‘m’ (below)

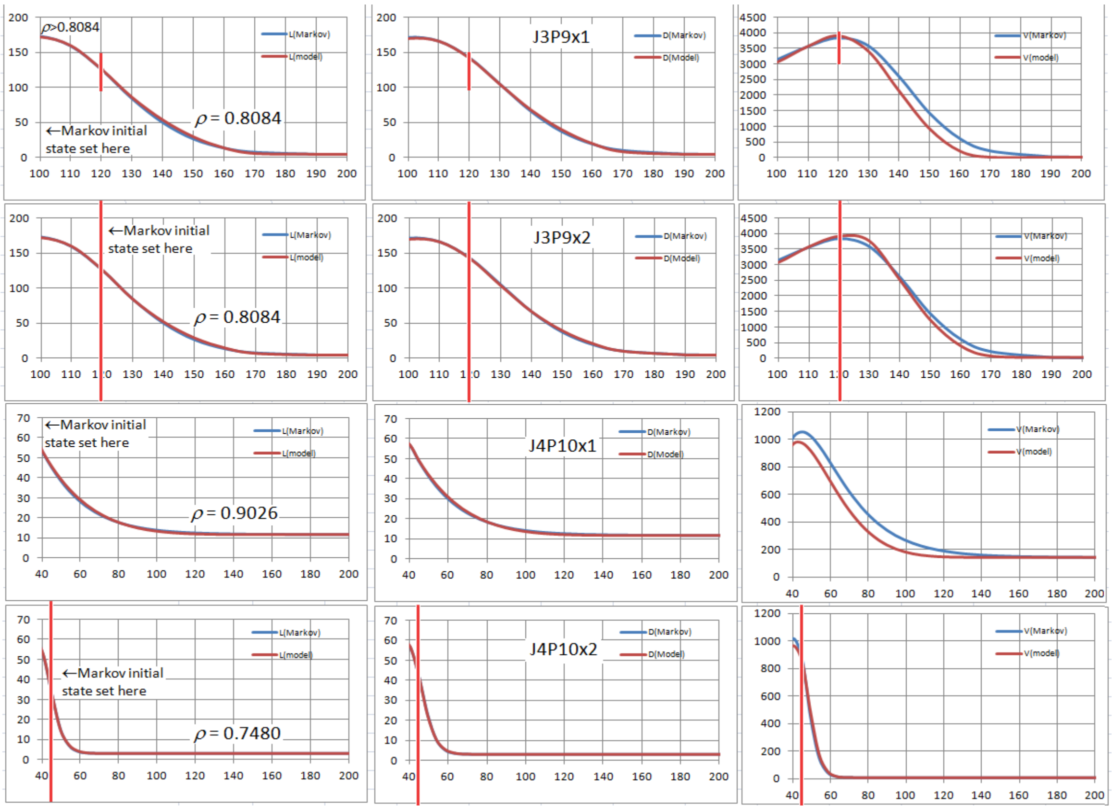


Figure 4.5.7 Results of the four run-out tests using ‘ckt-m’ combination

The ‘kinks’ in the first time slice of the ‘m’ examples above arise because of the need to force the input of the last oversaturated time slice to the Markov value. This is not required in subsequent tests, where the optimum method combination is used, together with the ‘Markov restarts’ given in Table 4.5.3. Estimated and simulated means and delays are almost indistinguishable, though the variances show that they are not perfect. This gives confidence that the exponential method is at least usable for the entire relaxation to equilibrium.

4.5.10 Accuracy of the time parameters and sources of uncertainty

On substituting the Markov simulated values for all M/M/1 peaks into equations (4.5.9-11) for the three alternative back-calculated parameters, Figure 4.5.8 shows values are similar.

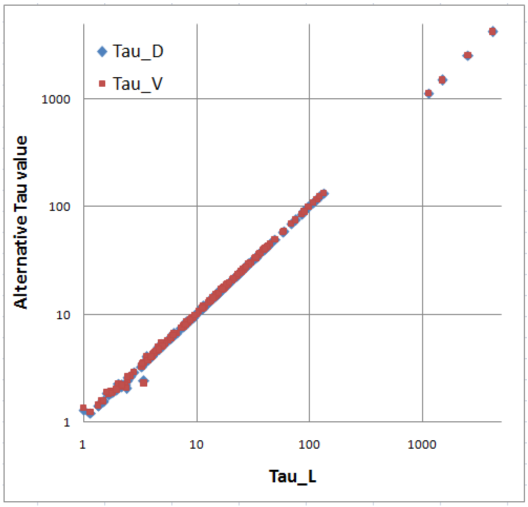


Figure 4.5.8 Similarity of alternative τ values back-calculated from Markov simulations

Figure 4.5.9 shows a fair match (on logarithmic axes) between estimated time parameters and the average of the nearly identical ‘benchmark’ values, with some scatter.

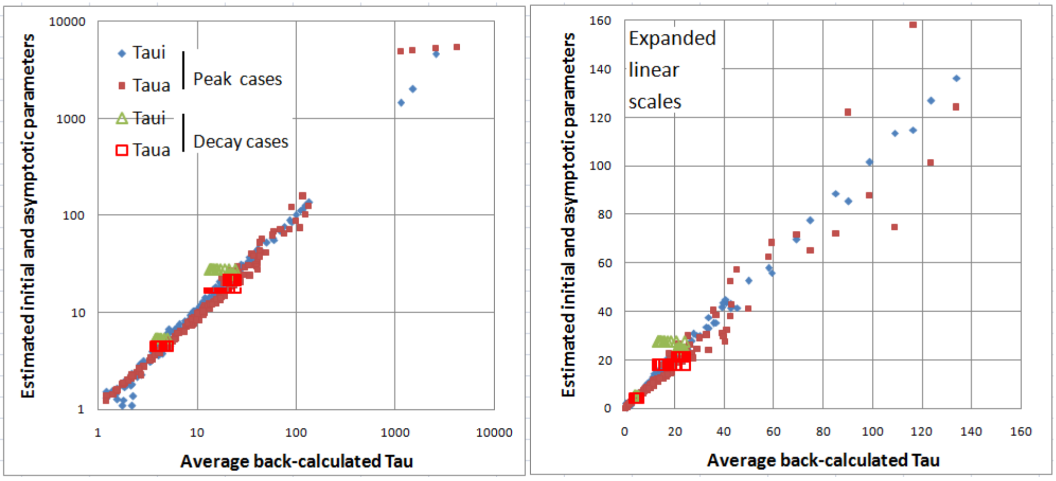


Figure 4.5.9 Estimated τ parameters versus τ back-calculated from Markov simulations

In Figure 4.5.10, supplementary data have been obtained from extended decays to equilibrium at constant ρ values. On expanded linear scales, most scatter is in τ_a since most of the evaluation times are quite short compared with the characteristic times, confirming that $\tau(t)$ should initially be close to τ_i. The equilibrating decays involve longer times so this should not apply to them. The different behaviour of the back-calculated τ values is emphasised for the equilibrating decays, also revealing limitations of the Markov simulation. Red lines mark where the Markov simulated mean queue ceases to converge, without actually reaching the precise equilibrium value. This may be a result of insufficient precision or too large a step size in the calculation, but in any case results are not meaningful beyond that point. As expected, τ_V converges to τ_a, but there appears to be no general tendency for τ_L to converge to a limit.

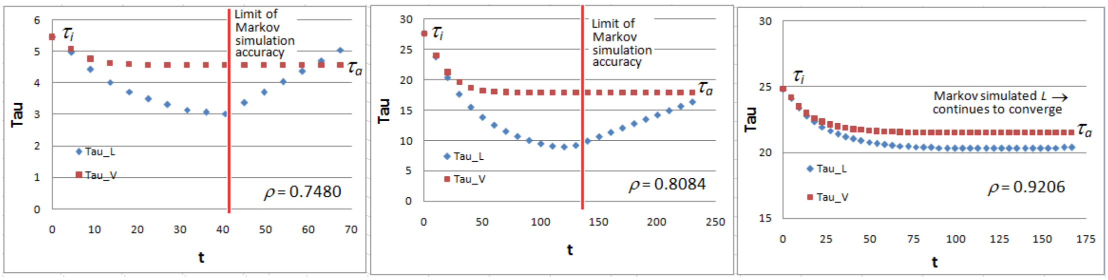


Figure 4.5.10 Behaviour of back-calculated τ for three equilibrating queues

Lower graphs are τ_L which unlike τ_V does not appear to converge to a limit

The most obvious error in the estimation of peaks occurs in the variance during the decay of two heavy peaks, J3P6 and J3P9, where the rate of decay is particularly high, as shown in Figure 4.5.11. Error in the mean queue or delay is hardly evident, although it is sufficient to produce some error in variance through the magnifying effect of the variance formula.

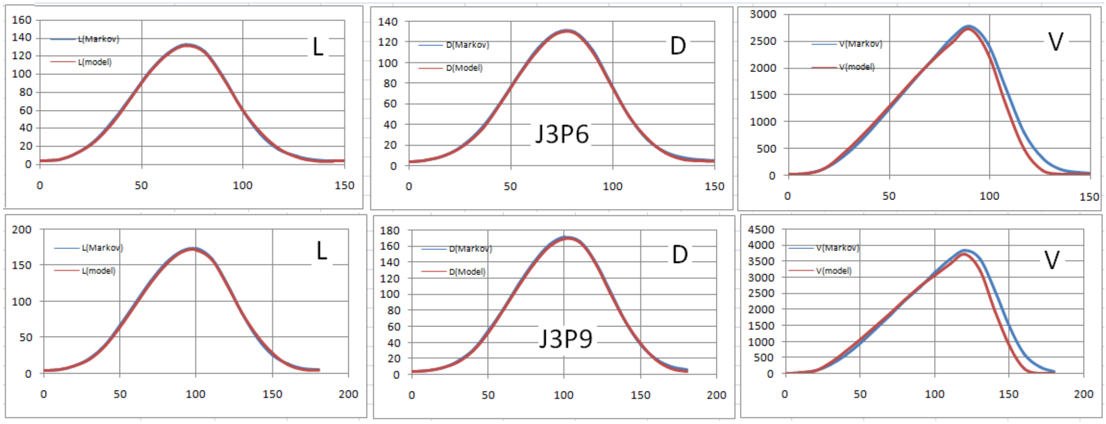


Figure 4.5.11 Predicted and Markov profiles for heavy peaks (ρ_∞=0.8086)

The maximum absolute error in variance in any one time slice in these cases represents underestimation by 24% (Markov prediction 2604, approximation estimate 1989). Greater percentage underestimation occurs but at much smaller absolute values, so is of less practical importance. In terms of standard deviation, the maximum absolute error occurs at a later time and amounts to underestimation by 47%. While these are substantial errors, they occur in a dynamic context where the queue and its variance are falling rapidly, as evident in Figure 4.5.11. Another way of looking at them is that the approximation predicts the end of a 2½ hour peak 10 minutes early, which seems less serious. It is natural to ask how the average τ_m in the exponential decay formulae would have to be adjusted to give the correct variance. Figure 4.5.12, drawn from the extended decay of J3P9, shows no simple pattern, and even with an adjustment results could still be very sensitive to small variations in parameters.

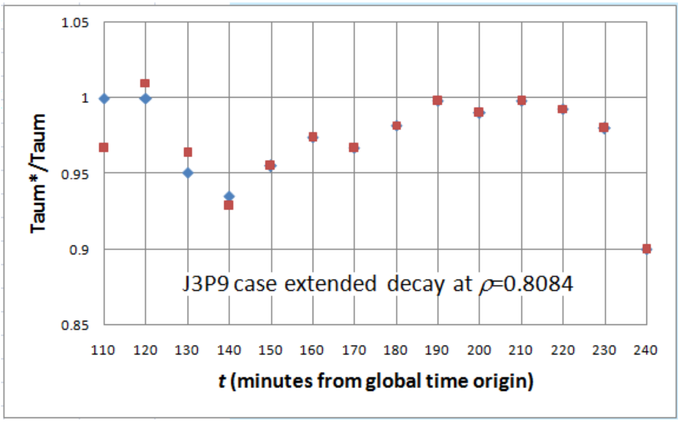


Figure 4.5.12 Factor to apply to τ_m to correct variance during decay in one peak case

Rapidly changing queue size and variance suggests that the queue size distribution is volatile. Moving ahead briefly to the material to be covered later in Chapter 5, Figure 4.5.13 (lower, overleaf) shows that the initial probability distribution for the final run-out period in J3P9 (Ts 12 running from *t*=110 to *t*=120) is complex, reminiscent of a water wave crashing into a sea wall, having developed quite rapidly over the previous 20 minutes from a near-Normal shape (top left) through a mixture of exponential- and Normal-like functions (top right) - note also the difference in the vertical (probability) scales as p_0 rises rapidly though still to nowhere near its equilibrium value.

Dynamic probability distributions such as this make it difficult to achieve accurate predictions with simple approximations, so a more detailed approach may be needed where highest accuracy is required. However, it is not clear that this is a practical necessity given the uncertainties in data that are likely to be involved, and that a probability distribution is not a prediction. For example, it seems unlikely that the precise queue size probabilities could be validated. Instead, the best practical prediction would be that in a certain time period following the peak, queue lengths can be very volatile, with short queues on some days and long ones on other days, for no obvious reason. So enhancement is left for future research.

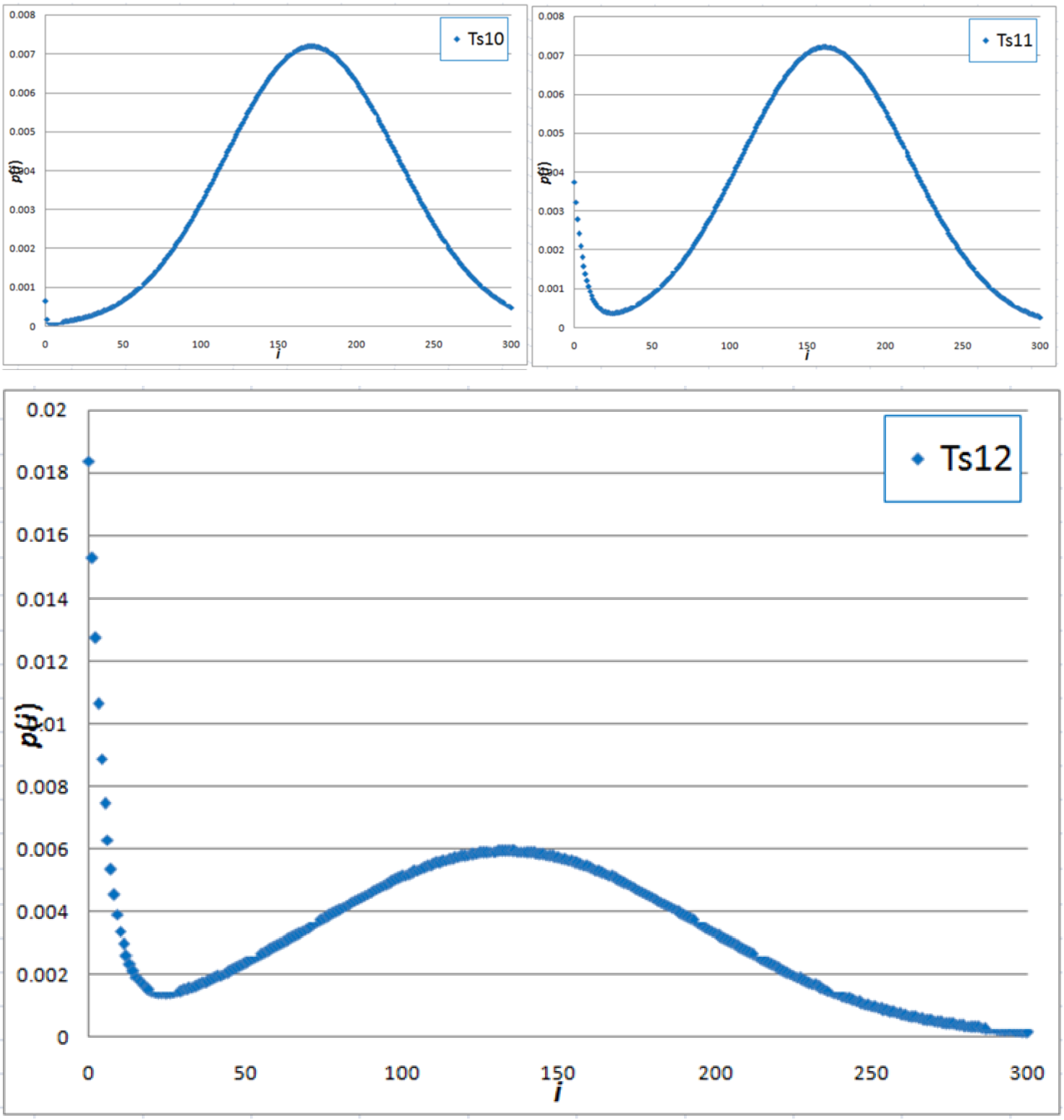


Figure 4.5.13 Probability distributions at start of J3P9 decay and 10 and 20 minutes later

4.5.11 Estimating error in time parameters and potential for correction

Using Markov simulated data as the input isolates consecutive time slices, so constituting a limited and artificial test. If the input for each time slice is taken more realistically from the output of the previous one, the value of τ can still be forced to produce the correct final value of one of the moments, by substituting one of the back-calculation formulae (4.5.9-11). At the same time, free values of τ_i, τ_a and τ_m from (4.5.22) can be calculated for comparison. Comparing τ_m with the 'optimum' τ should then give an idea of the practical magnitude of error and the conditions under which it occurs. Figure 4.5.14 shows that where $\tau_{\_L}$ is taken as optimum then most of the calculated τ_m are similar but a few are greatly overestimated.

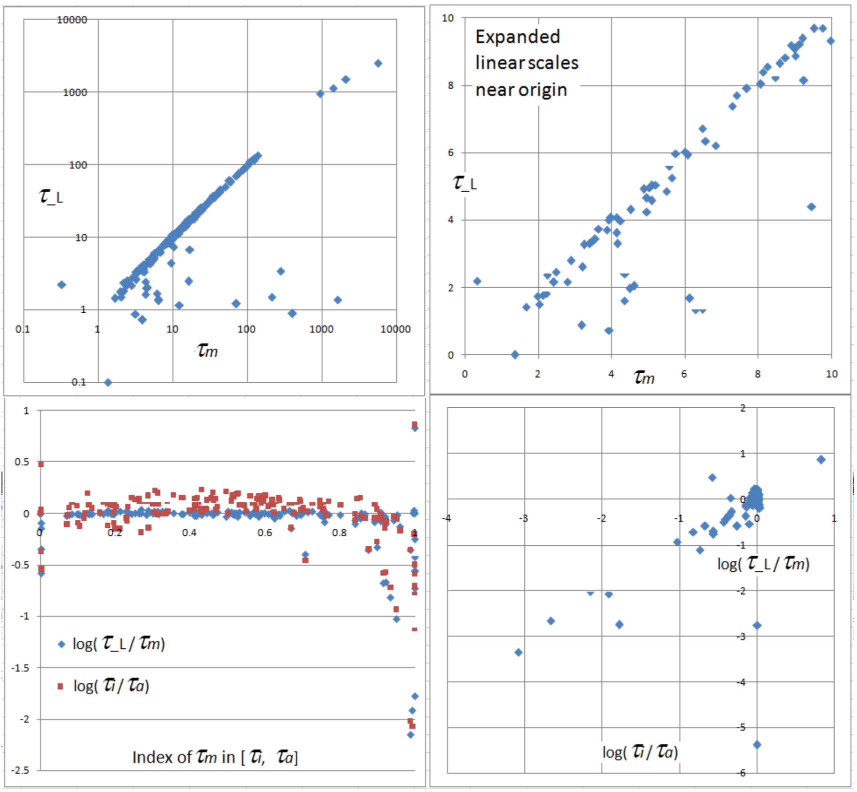


Figure 4.5.14 Comparison of estimated τ_m and ‘optimum’ based on *L*

In Figure 4.5.15, where back-calculated τ_V is taken as optimum, a few points are similarly scattered though not only overestimated (the scattered points are not even predominantly from the heavy cases selected above).

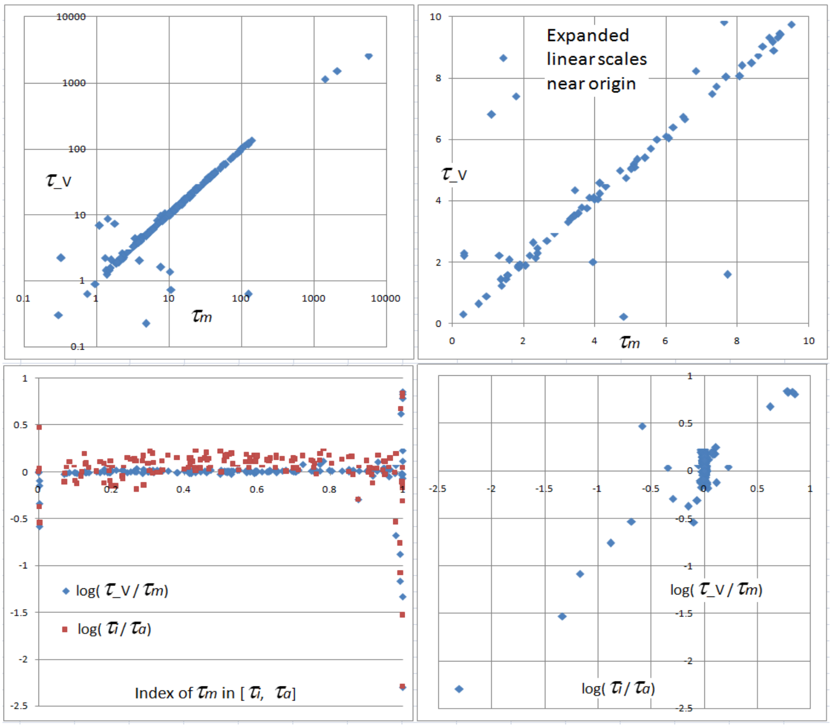


Figure 4.5.15 Comparison of estimated τ_m and ‘optimum’ based on *V*

In many cases the optimum τ lies somewhat outside the range of [τ_i, τ_a] suggesting that the monotonic form of (4.5.22) is an oversimplification, as accepted earlier. The lower graphs show that in *most* cases, where '~' means 'similar to':

$$\frac{\tau_{\_x}}{\tau_m} \sim \frac{\tau_i}{\tau_a} \qquad \text{(where '}x\text{' is } L \text{ or } V\text{)} \qquad (4.5.23)$$

The problem with applying this factor to (4.5.22) is that it gives the wrong extreme values. The cases where τ_m is most poorly predicted mostly occur where τ_i is much less than τ_a, and t is significantly greater than τ_i, so the exponential term in (4.5.22) has little influence, forcing τ_m close to τ_a. However, applying (4.5.23) uncritically can result in serious errors elsewhere, and there seems no reliable way to identify the cases that need correction.

4.5.12 Reality check on queue decay rate

In principle a queue can grow at any rate if there is sufficient demand, but it cannot decay at a greater rate than capacity. This maximum rate is what Kimber and Hollis (1979) assume for large queues in their modified sheared method. The deterministic queue formula applies in this case and places a lower limit on the queue size at any subsequent time. The exponential decay method 'm' ought not to predict a queue smaller than this at the end of the time period, but this could occur where the initial queue is large and the arrival rate is very low. While the mean queue could be corrected by adjusting τ_m, in practice this cannot be relied upon to ensure the consistency of the delay and variance, so in case of the exponential method 'm' lead to an anomalous result in one time slice, the time-origin-shifted method 't' is adopted. Somewhat unsatisfactory as this is, its resolution may depend on further research to improve the method of estimating queue decay. None of the peak cases produces this anomaly, but it can occur briefly in the more extreme random profile cases described next.

4.5.13 Random profile tests

To challenge the method further, three random profiles each with several oversaturated peaks have been tested. In the profiles, shown in Figure 4.5.16, ρ, μ and time-slice lengths have been generated randomly within specified ranges: traffic intensity $0.5 \leq \rho \leq 1.5$; capacity $20 \leq \mu \leq 24$ (veh/min), time slices 8-15 minutes. Traffic intensity is the critical factor, and this is allowed to exceed considerably the highest value in the peak cases.

All three ‘random’ cases are modelled using the ‘ckt-m’ method combination (for explanation see section 4.3.9 earlier) and M/M/1 process. These cases have been embedded and run in the Demonstrator software described later in Chapter 7. In all three there are several points where the demand intensity drops abruptly to low values, something that does not occur in the peak cases, resulting in certain ‘anomalies’ which fail the ‘reality check’. For example in case 1 these occur at around 50 and 160 minutes, while case 3 experiences many anomalies including a cluster in the range 210-300 minutes. With the substitution of methods applied in the time slices affected, the results are as shown in Figure 4.5.17. There is a tendency in all cases for the queue to grow on average over the test period, an expected consequence of the choice of a range of ρ centred on 1.

Any similarity between demand and mean queue profiles disappears as one moves from Case 1 to Case 3, the last being severely overloaded. The match of queue and delay is close in all cases. Variance is reasonably well estimated in the first two cases, but distinctly underestimated in the third. Having said that, the maximum underestimation of *standard deviation* is 22.3%. Case 3 is virtually insensitive to the methods used for growth, suggesting that it suffers from insufficient constraint on the variance, as in other growth tests.

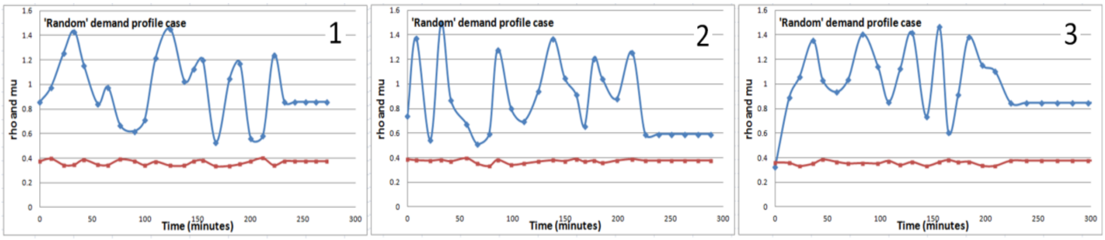


Figure 4.5.16 Three ‘randomly generated’ traffic profiles: ρ above, μ below

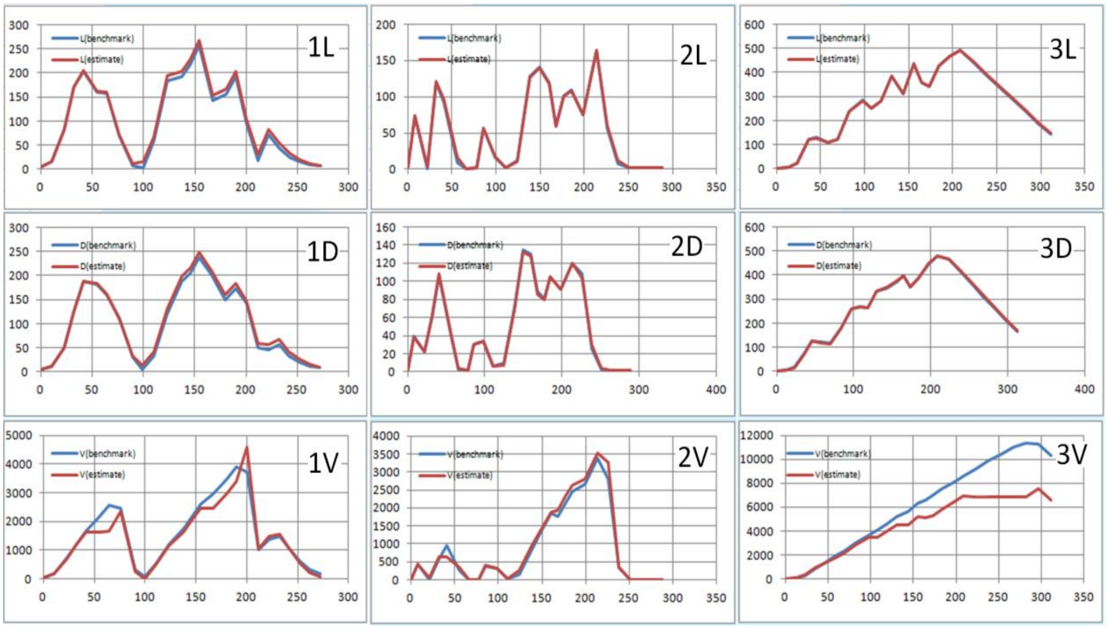


Figure 4.5.17 Modelled v. Markov mean queue, delay and variance for ‘random’ cases

4.5.14 Conclusion on mixed exponential approximation

The foregoing has shown that the simple exponential approximation is effective in most cases of decay including from large queues produced by oversaturation and over long periods of relaxation to equilibrium. The method can run into trouble in some more extreme cases but among the test cases these are relatively rare. Situations that can lead to problems include:

- Heavy queuing is followed by low demand, resulting in rapid queue decay
- Complex shape of initial probability distribution
- Large difference between initial and asymptotic time parameters.

Although some indicators have been identified, no simple modification to the mixed exponential method has been found that can deal with special cases without harming accuracy in the majority of cases. This suggests that a more general method would be more complex, making it more difficult to assure stability. Results with the peak cases for both M/M/1 and M/D/1 processes show that using 'optimum' τ values, rather than the estimated τ_m, would reduce average error by a modest 23-35%. However, the benefit of the best combination of approximation methods over uncorrected methods is orders of magnitude greater. For these reasons, the results are considered sufficient for the present purpose, while an improved approximation to queue decay could be a topic for future research.

4.6. TRANSITIONAL REGIMES AND COMBINED PERFORMANCE ON PEAKS

Figure 4.6.1 summarises the role of transitional regimes around the boundaries between under- and over-saturated regimes in an idealised peak case, giving a five-regime combination.

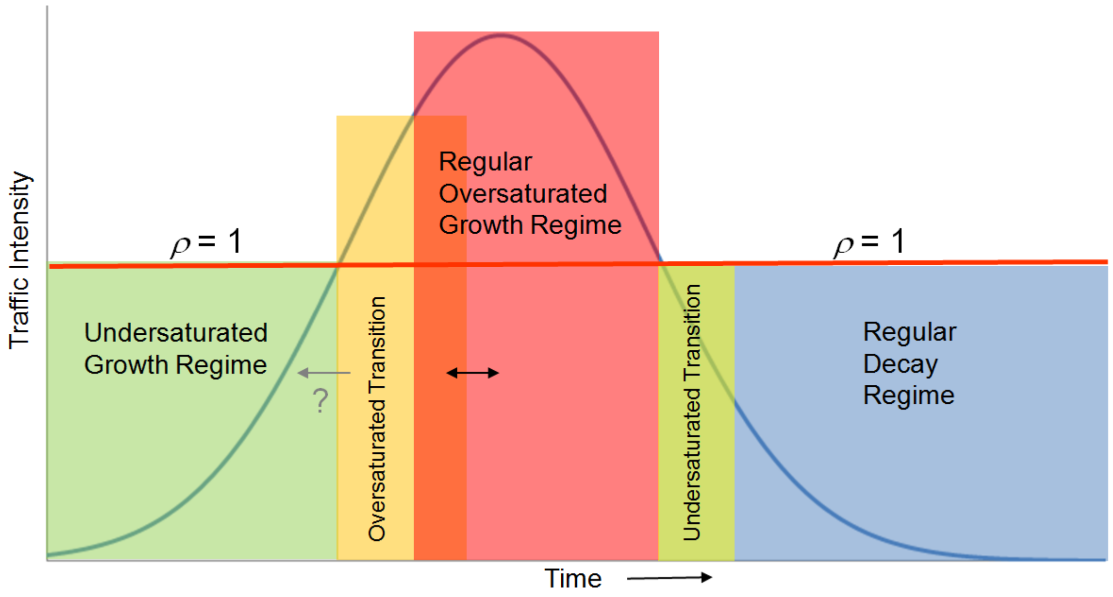

Traffic Intensity
$\rho = 1$
Regular Oversaturated Growth Regime
$\rho = 1$
Undersaturated Growth Regime
?
Oversaturated Transition
Undersaturated Transition
Regular Decay Regime
Time

Figure 4.6.1 Model of queue development regimes

For early oversaturated time slices, the use of transitional method ‘k’ in place of ‘t’ was described earlier in Section 4.4. This represents a merging of the basic undersaturated and oversaturated growth methods: ‘c’=corrected derived sheared + ‘t’=origin-shifted sheared ⇒ ‘k’=corrected origin-shifted derived sheared. This seems rational, but it works only for the first 130 or so units of throughput (see earlier in Section 4.4), and where this number comes from is a not known[47]. In a few cases it appears beneficial for the modification to extend into the undersaturated regime but there is no obvious criterion for this.

The ‘undersaturated transition’ (post-peak) matters in rare cases where queue decay does not follow a simple pattern, for example the mean queue increases or the variance increases. Modifications have been proposed that have some logical justification, but cannot be guaranteed optimal. A problem can occur where $\tau_a < 0$ in the time slice immediately following the oversaturated regime. This occurs in only one M/M/1 peak case, J2P10, and what works is simply to use the ‘c’ method instead of ‘m’, no other growth method working in this case.

[47] It is not known whether it is an absolute number or related to some common feature of the peak shapes which may not translate to more general cases. If interpreted directly as a relaxation parameter, 130 units throughput corresponds to ρ=0.832 or L_e≈5, or ρ=1.183 if it is accepted that a relaxation parameter is meaningful for ρ>1.

A less rare and extreme situation arises where a high value of L_e in the first decay time slice means that $L_e<L_0<2L_e$. For M/M/1 in these cases it is found that the ‘k’ method can give better results than ‘m’, while the ‘c’ method is less satisfactory, but for M/D/1 this results in catastrophic error in one case, J1P6, so is unproven in general. However, the fallback is the basic method combination that is no different in most cases and not much worse on average.

Questions remain about the best ways of dealing with these transitional regimes and special cases. Further research may reveal patterns that can be exploited but there is also a risk that over-optimisation may reduce the robustness of approximation.

In full tests of all 34 peak cases, where the input of each time slice is cascaded from the output of the previous time slice, results for both M/M/1 and M/D/1 are generally good. The optimum method combinations and their average RMS errors can be as low as shown in Table 4.6.1, where the growth methods are as defined in Table 4.3.3. More detail including non-optimal combinations is given in the next two main Sections.

Table 4.6.1 Summary performance of best method combinations in all peak cases

Method used in regime					**Average RMS error across all cases**		
Growth $\rho<1$	/	**Growth** $\rho\geq1$	\	**Decay** $\rho<1$	**L**	**D**	**V**
All 34 peak cases using M/M/1 process							
c	**k**	**t**	**-**	**m**	**0.40**	**0.59**	**30.41**
c	k	t	m	m	0.40	0.59	30.41
c	k	t	k	m	0.39	0.59	30.06
‘cktkm’ with Markov initial states					**0.19**	**0.32**	**19.37**
All 34 peak cases using M/D/1 process							
c	-	t	-	m	0.47	0.47	16.40
c	-	t	m	m	0.47	0.47	16.40
c	**k**	**t**	**-**	**m**	**0.35**	**0.37**	**16.27**
‘c-t-m’ with Markov initial states					**0.32**	**0.31**	**19.02**

The most reliable combination for both queue processes appears to be ‘ckt-m’, giving errors comparable with those of ‘optimum input and combination’ tests where time slices are calculated independently using the actual Markov estimated values as the inputs.

Figure 4.6.2 shows that for all peak cases together the estimation of variance is good. However, Figure 4.6.3 shows a range of performance for variance estimation, with some underestimation in the decay regime of J3P9, which is also found in other heavy peak cases – the slight overestimation in the lighter peak J1P1 originates in the growth regime.

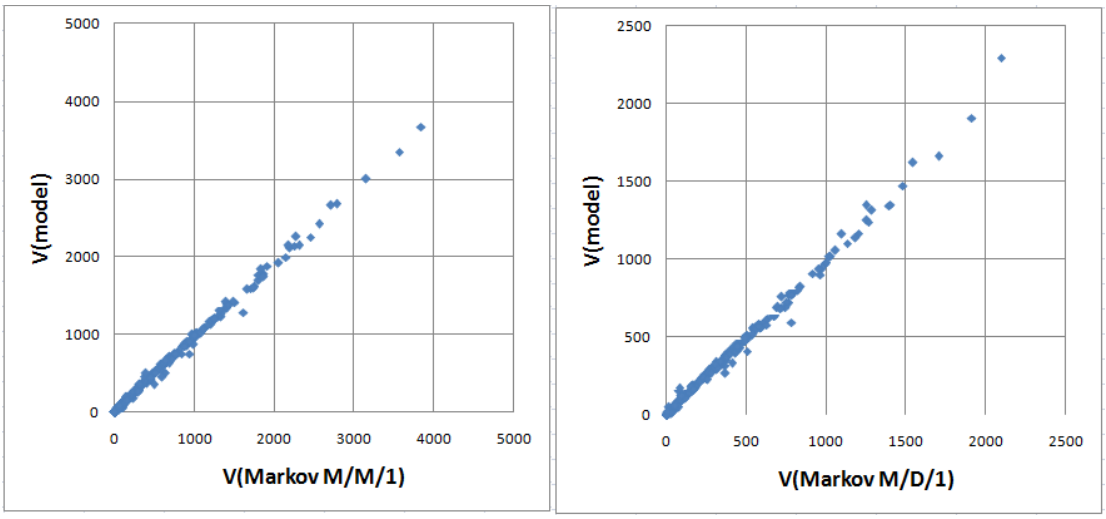


Figure 4.6.2 Match of variance estimates in all peak cases, using exponential decay method

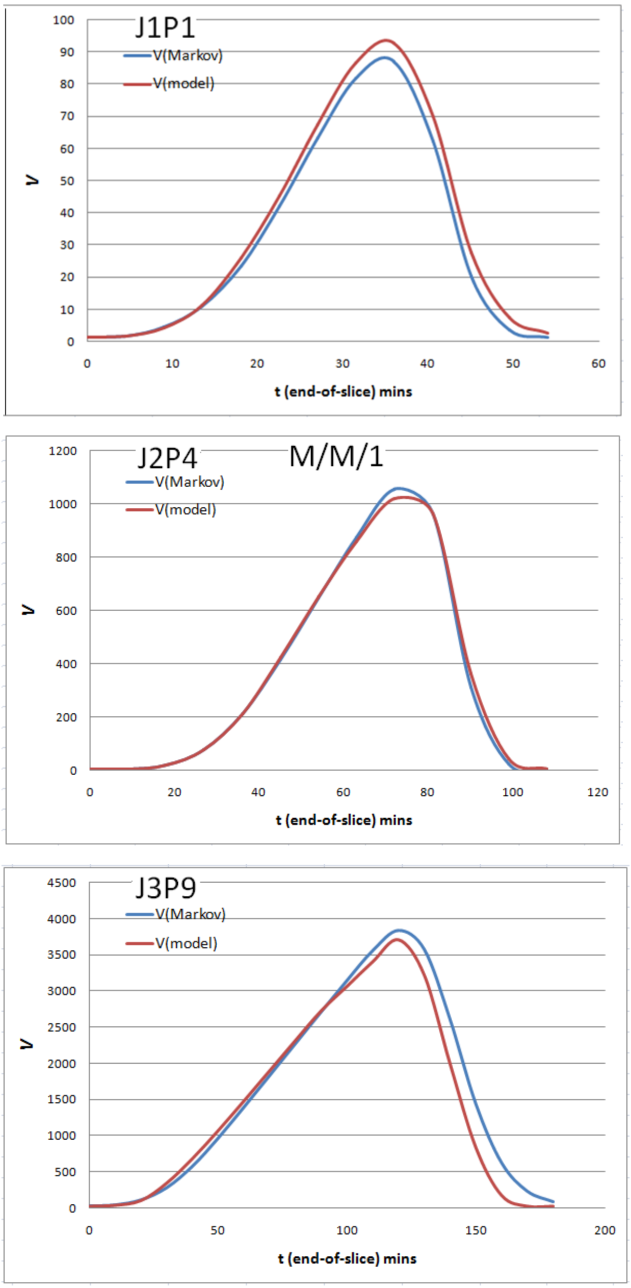


Figure 4.6.3 Estimation of variance in light, moderate and heavy M/M/1 peak cases

4.7. METHODS AND SUMMARY RESULTS FOR M/M/1 PEAK CASES

4.7.1 Methods tested

A test-bed spreadsheet has been set up so that several options can be tested on all 34 peak cases simultaneously, average errors noted from a central point, and comparisons of results with simulation plotted. Cases are limited to 12 time slices (sufficient in most cases to give adequate run-off) except for J1P9 (16), J3P6 (17), J3P9 (18) and J4P7 (14), resulting in a total of 425 calculations.Table 4.7.1 summarises the alternative methods available in various combinations.

Table 4.7.1 Summary of approximation methods

Symbol	Method
s	Basic sheared queue and delay
t	Origin-shifted sheared queue
d	Sheared delay, derived sheared queue
z	Origin-shifted derived sheared queue
c	Omega-corrected derived sheared
k	Omega-corrected origin-shifted derived sheared
m	Exponential

The main choice provided for is the method in each of the three main queue development regimes (see previous Section), but additional choices are provided for the first time slice of oversaturated growth, reflecting the alternative correction factors in equations (4.3.32), and the first time slice after oversaturation where it is found that the equilibrium queue size can exceed the initial queue, resulting in principle in an undersaturated growth situation. In the former case, a choice may be exercised between Ω-correcting or not, and in the latter a version of the exponential method can be applied. These choices are summarised in Table 4.7.2. Since ‘t’ and 'z' do not include correction they cannot guarantee accurate steady-state variance, though this is seldom approached in the peak cases, and is not an issue for oversaturation.

Table 4.7.2 Methods used in different regimes of queue development

Development regime	ρ	Typical methods
Undersaturated growth	<1	c, k, s, t, z
Oversaturated transition	≥1	*k, s, t*
Oversaturated growth	≥1	k, s, t
Undersaturated transition	<1	*c, m, t, z*
Decay	<1	m, t

The convention adopted here and in the test spreadsheet is to default the transitional methods if they are not specified, represented by dashes between the main methods, e.g. ‘c-t-m’. The default for the oversaturated transition is the main oversaturated method, while the default for the undersaturated transition depends on the relationship between L_0 and L_e.

4.7.2 Comparative performance of methods in peak cases with M/M/1 process

Table 4.7.3 lists the RMS errors (across all time slices) in the mean queue, delay and variance respectively. Variances tend to be the most sensitive, but it should be remembered that the error on the standard deviation, at most half that in the variance in percentage terms, may be thought a fairer criterion. The highlighted rows show that the basic sheared method 's-s-s', and the origin-time shifted modification proposed by Kimber and Hollis (1979) 't-t-t', perform much less well than the combinations involving corrections, but the corrected method does not work so well in the oversaturated regime. In a post-oversaturation transition the new exponential method appears safer than one of the methods derived from shearing. This may because the initial variance at that point is severely incompatible with quasi-equilibrium.The best methods are not far worse than 'optimal' calculations where Markov-simulated initial states are input in each time-slice rather than values calculated in the previous time slice.

Table 4.7.3 Errors for method combinations on M/M/1 peak cases

Method used for M/M/1					**Initial data, substitution etc.**	**Average RMSE (all 34 peaks)**		
Growth $\rho<1$	/	**Growth $\rho\geq1$**	\	**Decay $\rho<1$**		**L**	**D**	**V**
t		**t**		**t**		**19.05**	**21.18**	**4924.66**
m		m		m		182.58	173.62	503.77
d		d		d		4.44	4.92	413.39
s		**s**		**s**		**3.58**	**4.16**	**376.99**
d		d		m		4.03	4.55	368.07
c		c		m		3.55	3.96	339.62
c		s		m		3.51	3.95	330.78
s		s		m		3.38	3.79	326.52
z		z		m		1.52	1.58	192.15
c		z		m		1.54	1.57	191.88
k		z		m		1.46	1.55	189.36
z		t		m		1.21	1.73	76.11
k		k		m		0.86	1.44	75.20
k		t		m		1.30	1.81	73.97
c		k		m		0.42	0.59	69.82
t		t		m		1.40	1.93	65.70
c	k	t		m	$\tau_{\_\mathrm{L}}\rightarrow\tau_{m}$	0.35	0.58	44.89
c		t		m		0.77	1.07	38.68
c		t	m	m		0.77	1.07	38.68
c		t	k	m		0.75	1.06	36.89
c	**k**	**t**		**m**	**(Preferred)**	**0.40**	**0.59**	**30.41**
c	k	t	m	m		0.40	0.59	30.41
c	k	t	k	m		0.39	0.59	30.06
c	k	t		m	$\tau_{\_\mathrm{D}}\rightarrow\tau_{m}$	0.37	0.53	20.03
c	**k**	**t**		**m**	$\boldsymbol{\tau_{\_\mathrm{V}}\rightarrow\tau_{m}}$	**0.37**	**0.52**	**19.82**
c	k	t		m	Markov data	0.19	0.32	19.38
c	**k**	**t**	**k**	**m**	**Markov data**	**0.19**	**0.32**	**19.37**

The results labelled ‘Markov data’ take data for each time slice from the Markov simulation, so those results are only a benchmark. All the estimation methods cascade the output of each time slice to the input of the next. Stability issues can arise when this is done over many time slices with no external constraint, and this was a critical hurdle for the development of the correction method. The best performance is by variations of the combination ‘c-t-m’. Whether an appropriate method is chosen automatically for a transitional time slice depends on the exact results emerging from the previous time slice. There are so many possible combinations that only the ones that perform well, or extremely poorly, without exercising the transition options, are further tested with these options.

4.7.3 Summary of performance in peak cases with M/M/1 process

Figure 4.7.1 plots estimated versus Markov-simulated values of main results for all the 34 peak cases. Figure 4.7.2 shows for comparison plots for combinations ‘s-s-s’ and ‘t-t-t’, basic and origin-shifted sheared.

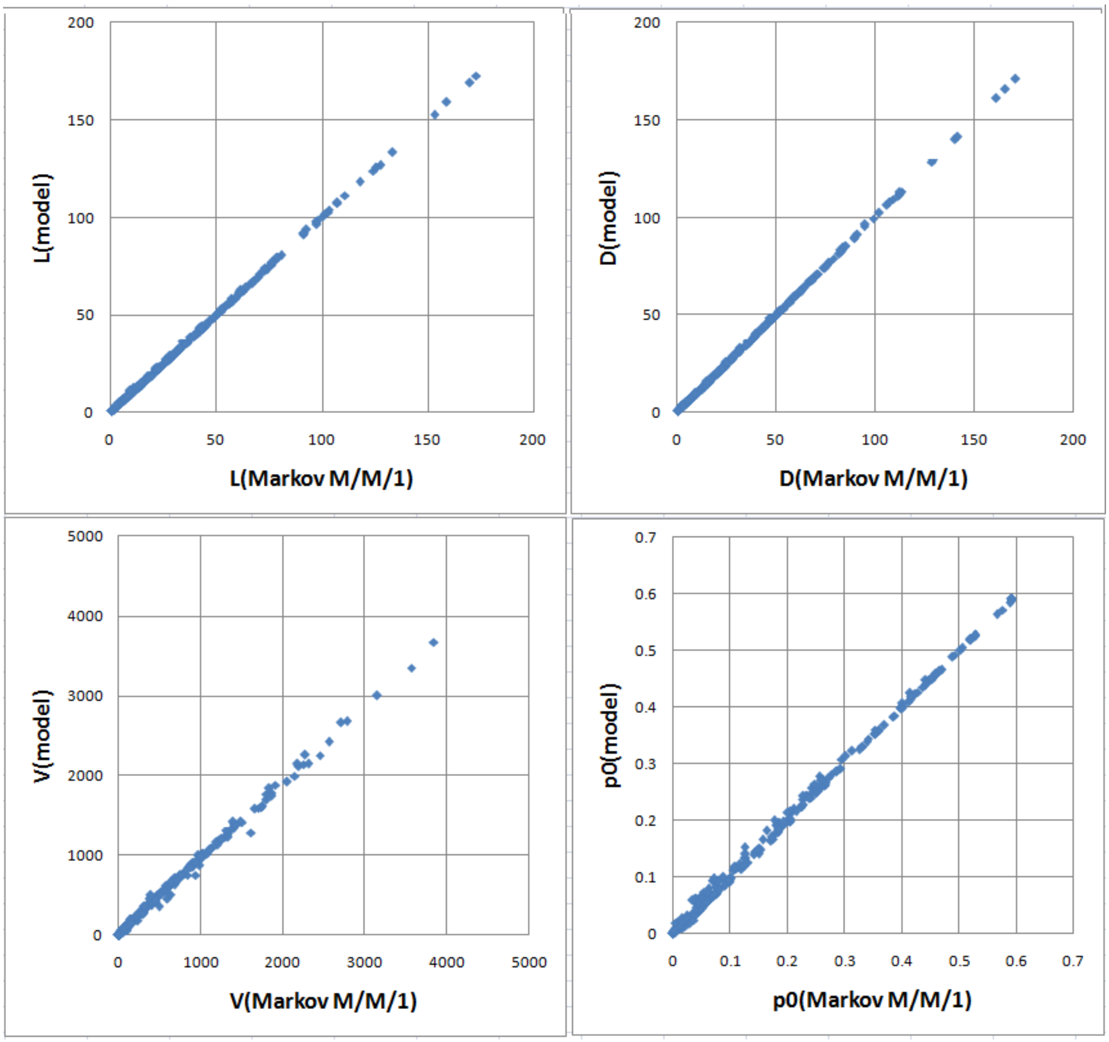


Figure 4.7.1 Performance of optimal ‘ckt-m’ model on 34 M/M/1 peak cases (408 points)

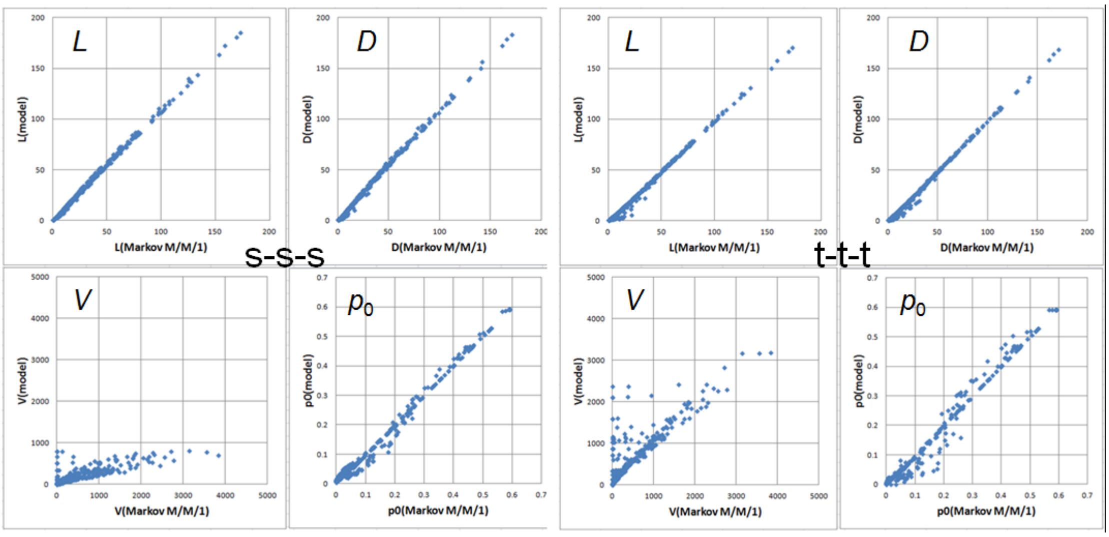


Figure 4.7.2 Performance of basic (left) and origin-shifted sheared (right) M/M/1 methods

Figure 4.7.2 shows that it is possible to get away with uncorrected models for estimating mean queue size and delay, but not for calculating variance. The greater scatter in variance *V* and zero queue probability p_0 compared to Figure 4.7.1 is evident. When individual cases are plotted the difference is particularly obvious. Figure 4.7.3 shows one of the heaviest and longest peaks, J3P6, for several model combinations. The sensitivity of variance is evident given the close fit of both *L* and *D* over most of the time range in all cases, but the performance of the uncorrected methods shows wild errors in variance compared to the corrected method.

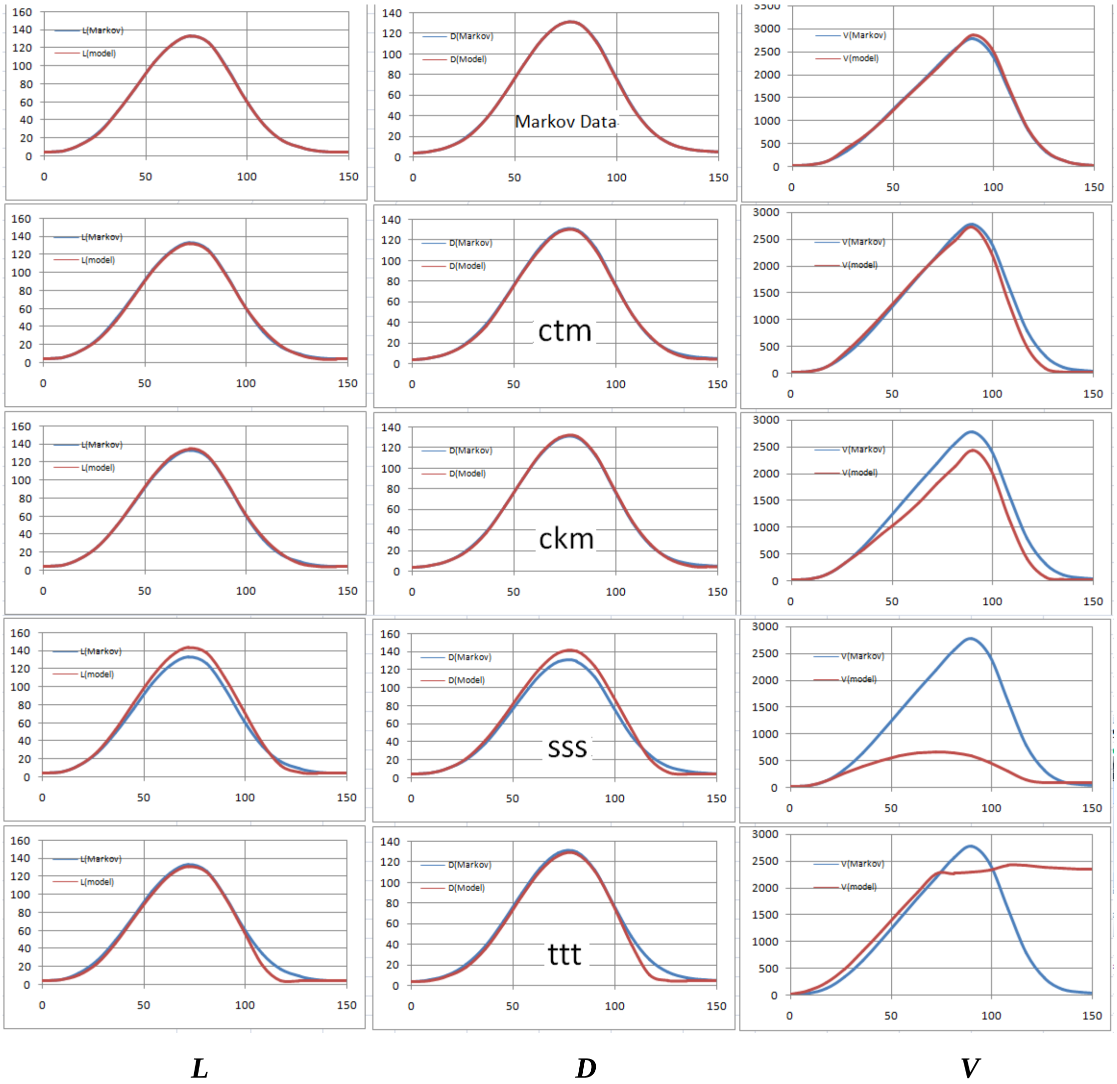


Figure 4.7.3 Performance on heavy peak case J3P6 (17 time slices), using M/M/1 process

(The Markov simulated benchmark profiles are common to all methods)

4.8. MODIFICATION AND SUMMARY RESULTS FOR M/D/1 PEAK CASES

4.8.1 Modification of calculation for M/D/1

Since almost all of the expressions derived in the previous Sections either include the statistical parameters of the P-K formula, or are independent of them and hence essentially deterministic, the method should in principle be applicable to any queue process including M/D/1.Before proceeding to test this, an issue must be resolved that requires revisiting an earlier point about the distinction between the average probability of zero queue and the value that appears in the theoretical probability distribution. The calculation methods work exclusively in terms of mean utilisation. If a value of p_0 is taken from a simulated probability distribution, in general it must first be translated into the corresponding average, the complement of utilisation. This does not affect actual calculations but only comparisons with Markov results. No changes should be needed to Ω or any calculations provided that the appropriately parameterised P-K formula or its inverse have been included in all calculations, and this is confirmed by results.

Whether it is simpler to calculate the Markov average utilisation from the Markov distribution p_0, or the estimated distribution p_0 from the estimated utilisation depends on the relationship. In the case of M/D/1 the latter is simpler, because the distribution p_0 is given at equilibrium by:

$$p_{0e} = e^{\rho}(1-\rho) \qquad \text{(in probability distribution)} \qquad (4.7.1)$$

Based on a quasi-static *assumption*, p_0 at any point can be estimated from the utilisation by:

$$p_{0[distribution]} = ue^{1-u} \qquad (4.7.2)$$

The alternative, starting from $p_{0[Markov]}$, is to solve (4.7.2) for a value of u to compare with the estimated utilisation. While such a solution must be numerical the solution method need not be efficient as it is needed only for verification and testing.

4.8.2 Comparative performance of methods in peak cases with M/D/1 process

Using an identical spreadsheet test-bed to M/M/1, with the appropriate statistical parameters but without modifying the case definitions, the 34 peak models have been recalculated for M/D/1 and compared with the results generated by the Markov simulation. The calculated value of p_0 is compared with the average p_0 as calculated above, rather than the simulation value.

Table 4.8.1 gives errors for M/D/1 peak cases using the same combinations of methods as for M/M/1. The smaller absolute errors are consistent with the smaller values of the moments compared to M/M/1. Combination ‘ckt-m’ is superior to ‘c-t-m’ and ‘c-k-m’, but all can be considered acceptable. The best combination for both queue processes is ‘ckt-m’.

Table 4.8.1 Errors for method combinations on M/D/1 peak cases.

Method used for M/D/1					**Initial data, substitution etc.**	**Average RMSE (all 34 peaks)**		
Growth $\rho<1$	/	**Growth $\rho\geq1$**	\	**Decay $\rho<1$**		**L**	**D**	**V**
d		d		d		3.26	2.87	554.14
s		**s**		**s**		2.43	2.42	539.61
t		**t**		**t**		8.45	8.53	486.51
c	k	t	k	m		0.76	0.73	325.89
m		m		m		279.20	262.13	297.94
d		d		m		2.74	2.55	188.01
c		s		m		2.36	2.19	167.34
s		s		m		2.27	2.10	166.00
c		c		m		1.84	1.67	121.60
c		z		m		1.07	0.87	100.05
k		z		m		0.94	0.77	92.68
z		z		m		0.98	0.80	93.45
c	k	t		m	$\tau_L \to \tau_m$	0.24	0.44	50.98
k		k		m		0.77	0.92	30.05
z		k		m		0.67	0.83	28.39
t		t		m		0.74	0.76	22.18
c		k		m		0.49	0.61	22.28
c	k	t		m	Markov data	0.54	0.47	21.02
k		t		m		0.61	0.64	20.97
z		t		m		0.55	0.58	20.24
c		**t**		**m**	**Markov data**	**0.32**	**0.31**	**19.02**
c		t		m		0.47	0.47	16.40
c		t	m	m		0.47	0.47	16.40
c	**k**	**t**		**m**	**(Preferred)**	**0.35**	**0.37**	**16.27**
c	k	t		m	$\tau_D \to \tau_m$	0.29	0.34	13.70
c	**k**	**t**		**m**	$\boldsymbol{\tau_V \to \tau_m}$	**0.30**	**0.34**	**12.68**

4.8.3 Summary of performance peak cases with M/D/1 process

Figure 4.8.1 plots estimated versus Markov-simulated values of main results for all the 34 peak cases. Figure 4.8.2 shows for comparison plots for combinations ‘s-s-s’ and ‘t-t-t’, traditional and origin-shifted sheared. The appearance of these plots, and comparative performance generally, is similar to that for the M/M/1 cases. Again, the greater scatter in variance V and zero queue probability p_0 compared to Figure 4.8.1 is evident. Figure 4.8.3 shows one of the heaviest and longest peaks, J3P6, for several model combinations, the results being somewhat similar to those for M/M/1 (Figure 4.7.3).

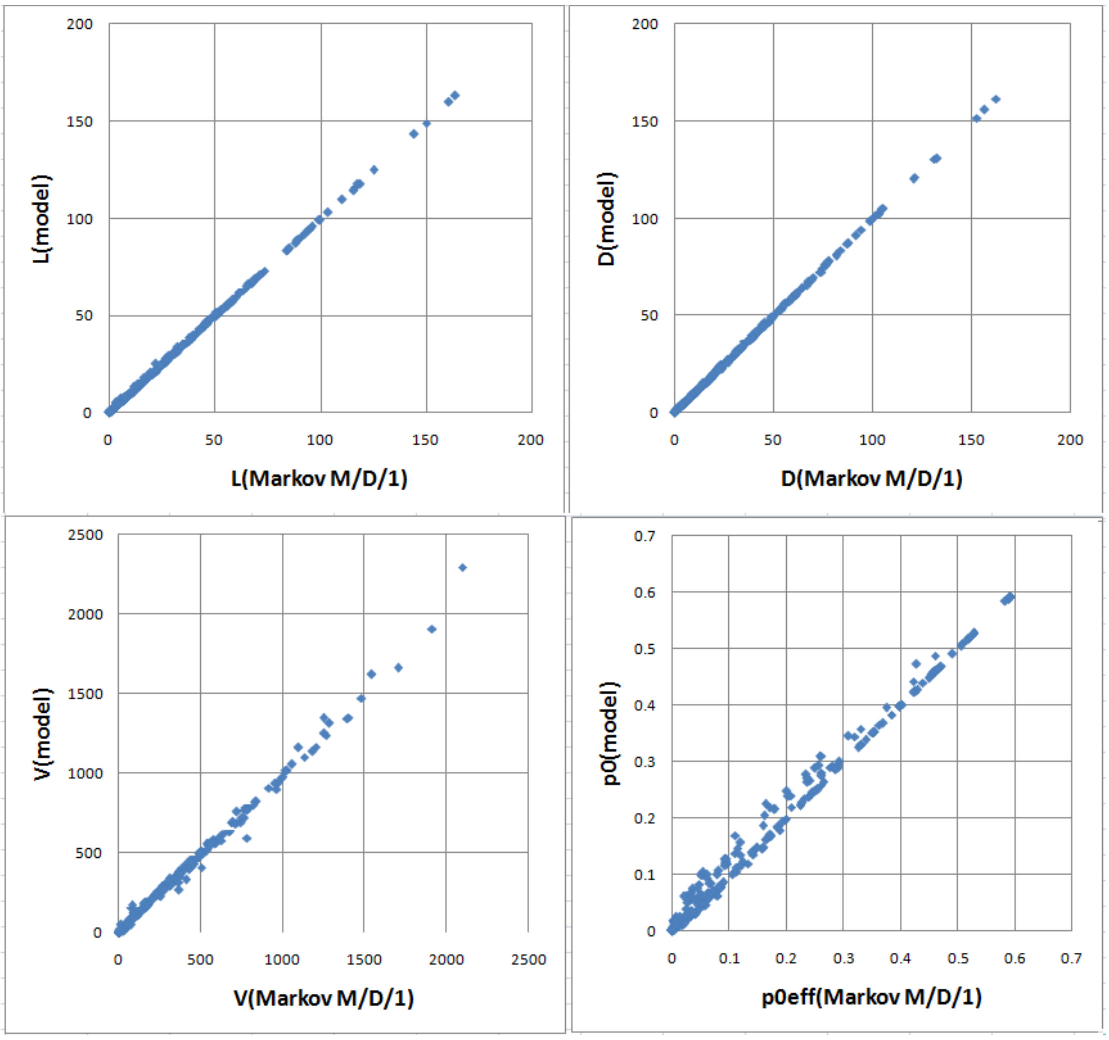


Figure 4.8.1 Performance of optimal ‘c-t-m’ model on 34 M/D/1 peak cases (408 points)

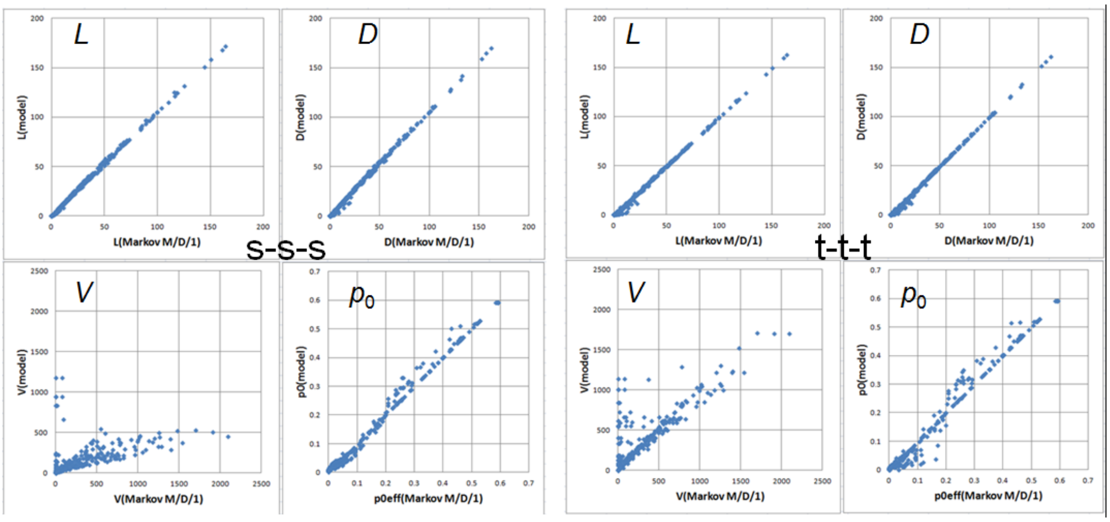


Figure 4.8.2 Performance of basic (left) and origin-shifted sheared (right) M/D/1 methods

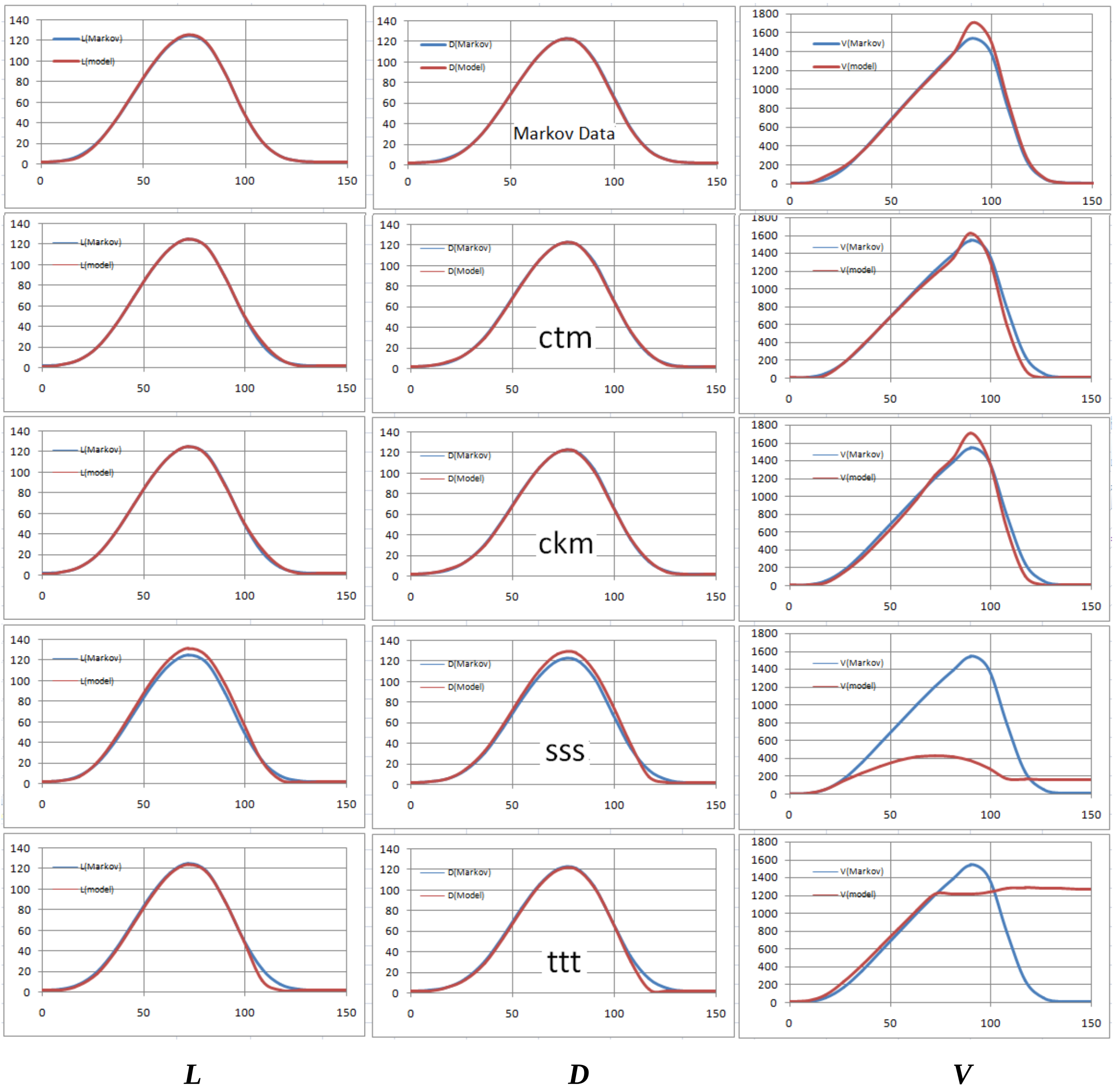


Figure 4.8.3 Performance on heavy peak case J3P6 (17 time slices), using M/D/1 process

(The Markov simulated benchmark profiles are common to all methods)

4.9. CONCLUSIONS ON TIME-DEPENDENT APPROXIMATIONS

In this Chapter 4 an improved and extended time-dependent approximation for queue development has been described which:

- Satisfies the time-dependent variance formula.
- Calculates essential queue properties: p_0 (~utilisation), mean and variance.
- Handles overcapacity seamlessly through its use of an enhanced sheared method applied to growth and a mixed exponential method applied to decay.
- Can accommodate different queue processes through the statistical parameters of the Pollaczek-Khinchin equilibrium mean queue and new equilibrium variance formula.
- Can accommodate some non-monotonic queue growth, e.g. from a ‘pure’ initial state.
- Is computationally efficient through the use of closed-form formulae.
- Is stable in the sense that queues can be handed on iteratively from one time slice to the next over many cycles to develop the prediction for an entire peak profile.
- Subject to qualification concerning variance estimates in a few cases, offers good accuracy compared to benchmark simulations of over-saturated peak test cases, modelled as a sequence of time slices with constant parameters within each time slice.

The methods have several possible or inherent weaknesses:

- The correction of the sheared method for queue growth assumes that stochastic relaxation time can be used to normalise the time scale of queue development, though this is supported by general arguments and empirical evidence.
- Fine tuning of the correction of queue growth regimes employs adjustment factors involving ‘free parameters’ (although the main correction does not).
- Sensitivity around the edges of oversaturation has been addressed by *ad hoc* rules.
- The mixed exponential approximation for queue decay is based on general principles applied independently to mean and variance, and uses an heuristic method of interpolation over the time range, which is challenged in a few extreme cases and cannot accommodate non-monotonic queue development.

A legitimate question in any engineering discipline is “when is good good enough?”, or “when is bad not bad enough to worry about?” This is essentially the precision v. accuracy dichotomy again. Arguments have been put forward that the results are sufficiently good for the practical purposes objectives of this research, which can be considered to have been achieved, and further improvement in the methods can be left for future research.

CHAPTER 5: ESTIMATING PROBABILITY DISTRIBUTIONS

5.1. INTRODUCTION

Computationally convenient ways of approximating the queue size probability distribution at any time are explored, given the three queue properties available from the foregoing time-dependent approximation methods: p_0, mean and variance. Standard continuous diffusion approximations are first reviewed, but are considered to be both more complex and less flexible than is required, bearing in mind that a fully time-dependent solution is not required since this is already embodied in the queue moments obtained by the estimation methods described earlier. Alternative approaches are developed based on continuous functions that are easier to work with than discrete forms and can exploit standard solution methods.

5.2. POTENTIAL OF CONTINUOUS APPROXIMATIONS TO DISTRIBUTIONS

Figure 5.2.1 superimposes normalised simulated distributions for the 12 time-slices of the J2P4 peak case (see Section 2.5 earlier). In addition to being normalised to have equal maxima, the distributions have been shifted so that all their means lie at x=0.

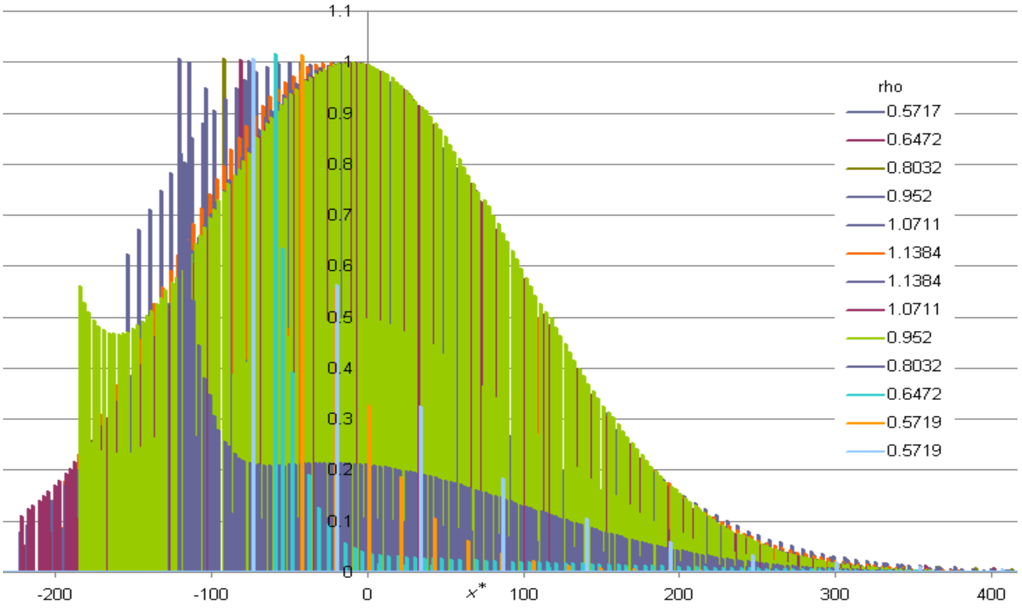


Figure 5.2.1 Superimposed normalised probability distributions from J2P4 peak case

Graphical examples in Kobayashi (1974b) show that a diffusion approximation can reproduce many of the distribution shapes between Normal and exponential/geometric-like, including bi-modal with the 'duck-tail' at the left, though not as pronounced as in the simulation. However, they appear unable to reproduce the more extreme distribution shapes in some later time slices.

Any queue distribution can be viewed as a linear superposition, like the solution given by Morse (1958). States below the equilibrium mean will behave like growing queues, while those above it behave like decaying queues. This hints at the possibility of simply combining exponential and Normal functions, but it remains to be shown that this can be done uniquely.

5.3. TIME-DEPENDENT DIFFUSION SOLUTIONS FOR M/M/1

5.3.1 The diffusion equation

The Kolmogorov Forward Equation or Fokker-Planck Equation (FPE), as quoted e.g. by Newell (1968a), is considered the prototype for a continuous analogue of recurrence relations. Its coefficients represent drift at a rate determined by the difference between capacity and arrivals, and diffusion at a rate determined by the variance. The diffusion term includes the indices of dispersion of arrival and service, I_a and I_μ:

$$\frac{\partial p(x,t)}{\partial t}=\mu\left[\frac{(I_a\rho+I_\mu)}{2}\frac{\partial^2 p(x,t)}{\partial x^2}+(1-\rho)\frac{\partial p(x,t)}{\partial x}\right] \qquad (5.3.1)$$

Newell (1968a) says, for a Poisson (i.e. M/M/1) process, only that both I_a and I_μ are ‘suitable coefficients’ expected to be ‘comparable to 1 and essentially independent of [ρ] or μ’, which is certainly consistent with c_b^2=1, but raises a question about what is meant by I_a here (one suspects it corresponds to c_a^2). The term is the exact variance only in the deterministic case, suggesting that a solution will apply to ‘heavy traffic’, i.e. where ρ≈1. It is not clear to what extent this can cover the full range of the P-K formula.

A diffusion equation also arises in an analysis of platoon bunching, derived from M/M/1 time-dependent recurrence relations (Kühne and Lüdke 2013). After translating variables, this amounts to:

$$\frac{\partial p(x,t)}{\partial t}=\mu\left(\frac{\partial^2 p(x,t)}{\partial x^2}+(1-\rho)\frac{\partial p(x,t)}{\partial x}\right) \qquad (5.3.2)$$

Equation (5.3.1) is consistent with this only if I_a is interpreted as c_a^2, rather than $c_a^2\rho$ in accordance with equation (3.4.5). This is inconsistent with the assertion of Newell (1968a) that I_a in equation (5.3.1) should be around 1, while $I_\mu = c_b^2 = 1$ remains uncontroversial. This ambiguity will be commented on further shortly, but the use of the diffusion approximation in what follows will be pragmatic.

5.3.2 Diffusion solutions combining exponential and Normal functions

The exponential function is a solution of the static FPE (LHS of (5.3.1/2) = 0), and the Normal function is a solution of FPE remote from *x*=0, making for an analogy with the equilibrium and deterministic queue formulae. An initial pure state or ‘shock’ will initially drift and diffuse into a Normal distribution whose mean evolves naturally according to the difference between arrival and departure rates and whose standard deviation increases approximately as the square root of time. If there were no barrier at *x*=0, the distribution would spread out until it becomes uniform. According to Rose (1995), the zero queue state is not a true reflecting barrier but a ‘reflecting sticky barrier’, meaning that on reaching zero the state remains there until it receives a positive impulse. Ultimately the distribution should relax to a stable exponential form.

Kobayashi (1974a,b), Newell (1968a-c, 1982), Kleinrock (1976), Gross *et al* (2008) and others give solutions for the time-dependent M/M/1 queue size probability distribution evolving from an initial pure state or ‘shock’ towards the equilibrium distribution under constant arrival and service rates. Gross *et al* give a solution, expressed here in slightly modified notation[48]:

$$p(x,t|x_0)=\frac{1}{\sqrt{2\pi(1+\rho)\mu t}}\left[e^{-\frac{(x-x_0+(1-\rho)\mu t)^2}{2(1+\rho)\mu t}}+e^{-\frac{2(1-\rho)x}{(1+\rho)}}\left(e^{-\frac{(x+x_0+(1-\rho)\mu t)^2}{2(1+\rho)\mu t}}+\frac{2(1-\rho)}{1+\rho}E(x,t|x_0)\right)\right]$$

(5.3.3)

where

$$E(x,t|x_0)=\int_x^{\infty}e^{-\frac{(y+x_0+(1-\rho)\mu t)^2}{2(1+\rho)\mu t}}dy=\sqrt{\frac{\pi(1+\rho)\mu t}{2}}erfc\left(\frac{x+x_0+(1-\rho)\mu t}{\sqrt{2(1+\rho)\mu t}}\right)$$

(5.3.4)

The formula can be rewritten in a form more convenient for evaluation:

$$p(x,t|x_0)=\left[\frac{e^{-\frac{(x-x_0+(1-\rho)\mu t)^2}{2(1+\rho)\mu t}}+e^{-\frac{2(1-\rho)x}{(1+\rho)}}e^{-\frac{(x+x_0+(1-\rho)\mu t)^2}{2(1+\rho)\mu t}}}{\sqrt{2\pi(1+\rho)\mu t}}\right]+\frac{(1-\rho)}{1+\rho}e^{-\frac{2x(1-\rho)}{(1+\rho)}}erfc\left(\frac{x+x_0+(1-\rho)\mu t}{\sqrt{2(1+\rho)\mu t}}\right)$$

(5.3.5)

Kobayashi (1974a) points out that the barrier at *x*=0 makes this a Wiener process, which can be characterised by drift (ρ-1)μ*t* and variance (ρ+1)μ*t*.

[48] Some authors use λ and μ representing arrival and service rates. Others use expressions involving μ_x in the sense of interval times. Newell and Kobayashi also transform or scale variables and extract common terms to simplify formulae. This can aid calculation substantially, but does not always assist physical interpretation, so we stick with variables used elsewhere in the dissertation.

Figure 5.3.1 shows that this equation generates realistic distributions, and suggests that the left and right parts of (5.3.5) represent Normal-like and exponential-like components respectively. A feature of the model is that $p(x,t|x_0)$ integrates to 1 over the range $(-\infty,+\infty)$ rather than $[0,\infty)$, as a result of which the integral over $[0,\infty)$ declines with time, and the mean, variance and p_0 have to be normalised by dividing by the integral of p, p_0 being assumed to be the normalised integral over [0,1). Accuracy begins to deteriorate at large values of t.

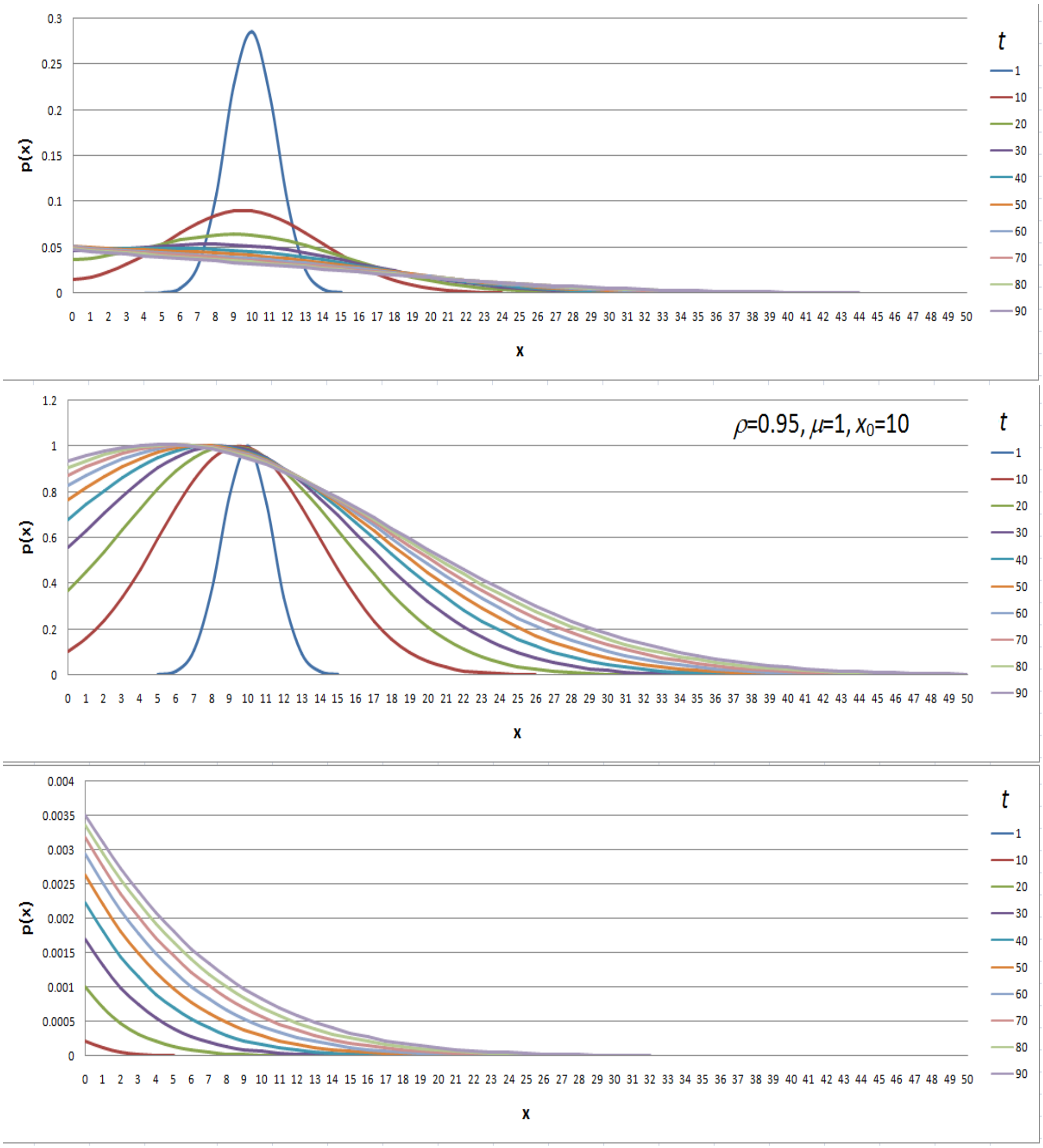


Figure 5.3.1 Distributions generated by Gross *et al* diffusion approximation (5.3.5), respectively the full distribution, two LH terms and RH term (p values unnormalised). Development is shown over a short time period after an initial 'impulse', during which the upper (Normal-like) component diffuses, while the lower (equilibrium-like) component grows

In principle, predicting the queue size distribution from an arbitrary initial distribution, as may be required when calculating the result of an arbitrary demand/capacity profile, requires convolving (5.3.5) with an arbitrary initial distribution. One way to avoid this would be to calculate for some x_0 and $t+t_0$ where t_0 is the time at which the distribution most closely resembles the required starting distribution, analogous to origin-shifting the sheared queue. Calculating the mean and variance requires integrating terms in *erfc*, which should be relatively straightforward since terms in $x^n erfc(x)$ integrate to expressions involving $x^{n+1} erfc(x)$.

The distribution function must move from a Normal-like shape to an exponential-like shape, and these components are virtually independent, analogous to the deterministic and P-K components of the sheared formula. However, the former consists of not one but two Normal-like components, because of the mirror term. Functionally, this component arises as a link between the two main functions, rather as the utilisation variable *x* in the sheared solution.

Gross *et al*'s expression appears to suffer from confusion about the sign of the (1-ρ) term (though this could lie with the present author). Kobayashi (1974b) quotes a differential form of the diffusion approximation (apparently misprinted in Kobayashi (1974a)), which he transforms into two 'coordinate-free' solutions applying to ρ<1 and ρ>1 respectively. The expression involves a function Φ similar to *E* in (5.3.4) except that the integral is over (-∞,*x*), and (1-ρ) appears as its absolute value. As far as can be ascertained, when evaluated for ρ~1, the probabilities sum to near 1 for all *t*, and moments approach their equilibrium values accurately. The different forms for ρ<1 and ρ>1 can be absorbed into a single expression, with change of sign of (1-ρ) being accommodated by use of the absolute in some terms. Reverting to the untransformed variables, a replacement for (5.3.5) which embraces the full range of ρ is:

$$p(x,t|x_0) = \left[\frac{ e^{-\frac{(x-x_0+(1-\rho)\mu t)^2}{2(1+\rho)\mu t}} + e^{-\frac{2x|1-\rho|}{(1+\rho)}} e^{-\frac{(x+x_0-(1-\rho)\mu t)^2}{2(1+\rho)\mu t}} }{ \sqrt{2\pi(1+\rho)\mu t} } \right] + \frac{|1-\rho|}{1+\rho} e^{-\frac{2x|1-\rho|}{(1+\rho)}} \overset{\leftrightarrow}{E}(x,t|x_0)$$

(5.3.6)

where:

$$\overset{\leftrightarrow}{E}(x,t|x_0) = erfc\left(sign(1-\rho)\left(\frac{x+x_0-(1-\rho)\mu t}{\sqrt{2(1+\rho)\mu t}} \right)\right), \quad erfc(-|z|) \equiv 2 - erfc(|z|)$$

(5.3.7)

Figure 5.3.2 shows the difference between the Gross and Kobayashi formulae results. The calculations are for integral values of *x*, and the sums of probabilities of the Kobayashi formula for ρ=0.95 are up to 3.7% high, but the moments are quite accurate when adjusted for Σp_i, with

p_0 converging to about 0.0487. For smaller values of ρ the error in Σp_i is greater, for example 37% at ρ=0.5, but again the moments are quite accurate, with p_0 approaching 0.476.

While it is possible that Gross *et al*'s function has been misunderstood, as it stands it produces strange results (left), in particular the sum of probabilities decays with time while the mean rises well above the expected equilibrium value. Kobayashi's version as interpreted (right) gives sensible results where the sum of probabilities remains around 1, and the mean converges to the equilibrium value, so it is adopted as fulfilling the requirement.

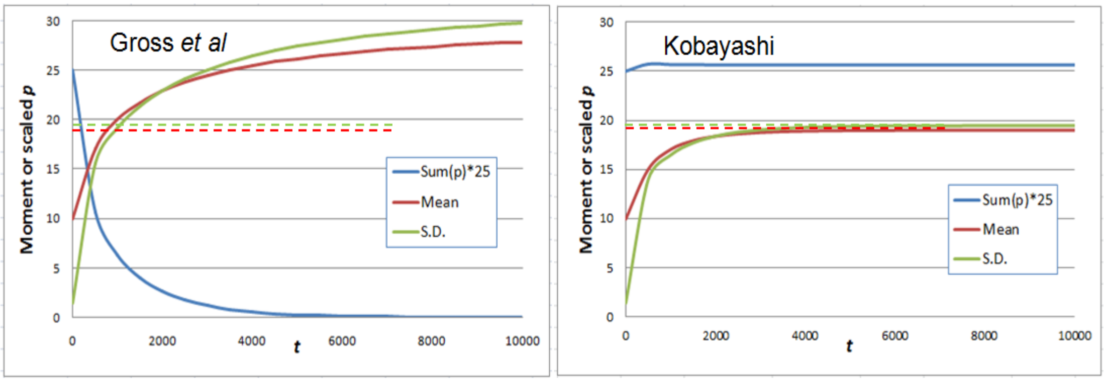


Figure 5.3.2 Time-development of moments of approximations for M/M/1, ρ=0.95, L_0=10

Figure 5.3.3 shows the full distributions and their components produced by (5.3.6-7). Comparing with Figure 5.3.1, the main visible difference is in the combined distribution. At first it appears that the two Normal components have taken over the whole distribution, but actually the exponential component is still growing with time. The changing balance between them over a longer time period is graphed by Figure 5.3.4, showing that the exponential component does eventually become dominant.

Drawbacks of the diffusion approximation have already been touched upon. They include:

- Starts from an initial 'impulse', so in principle needs to be convoluted with an initial probability distribution, or retro-fitted to one by shifting the time origin.
- Less accurate for smaller values of ρ.
- Modification for general queue process statistics is questionable.
- Analytical calculation of moments requires complicated integrations.

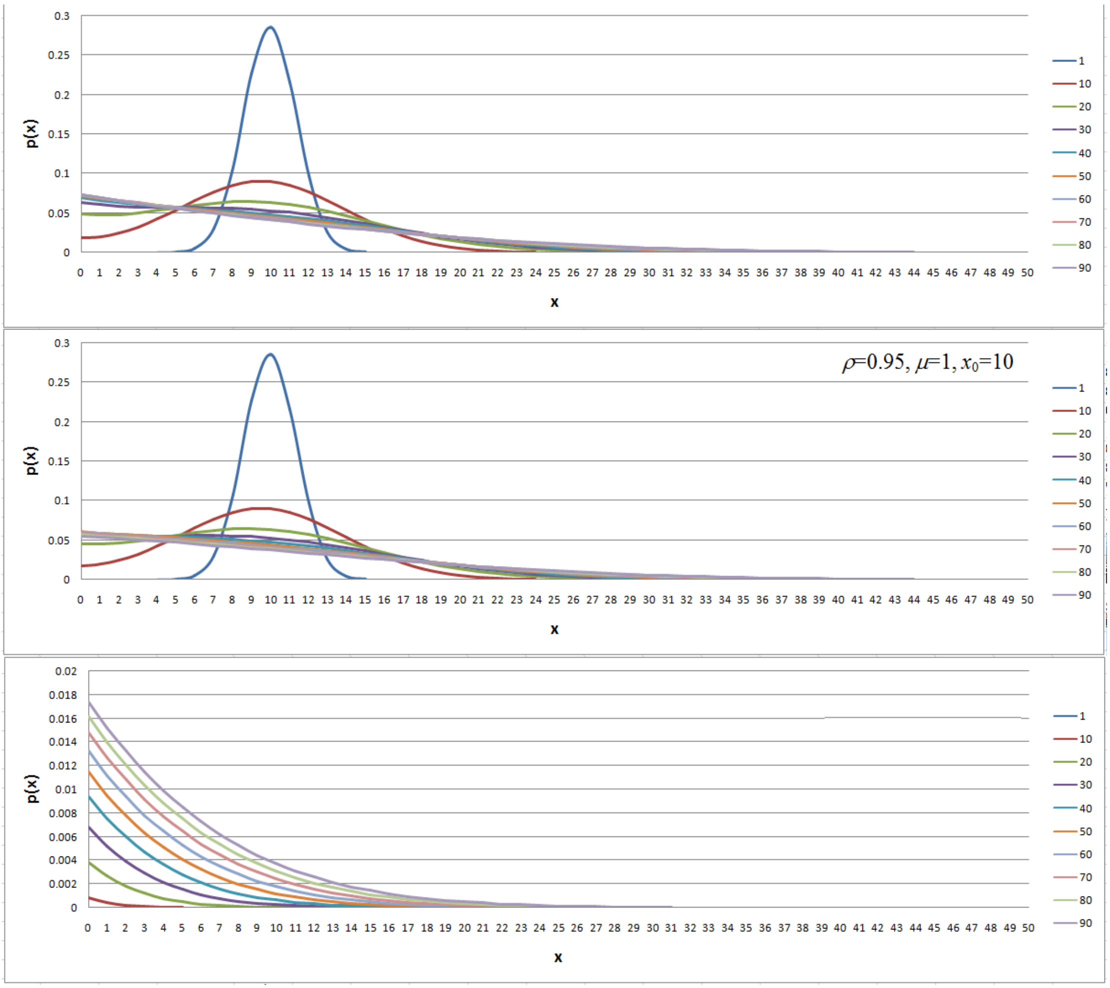


Figure 5.3.3 Distributions generated by Kobayashi diffusion approximation (5.3.6) Development is shown over a short time period after an initial 'impulse', during which the upper (Normal-like) component diffuses, while the lower (equilibrium-like) component grows

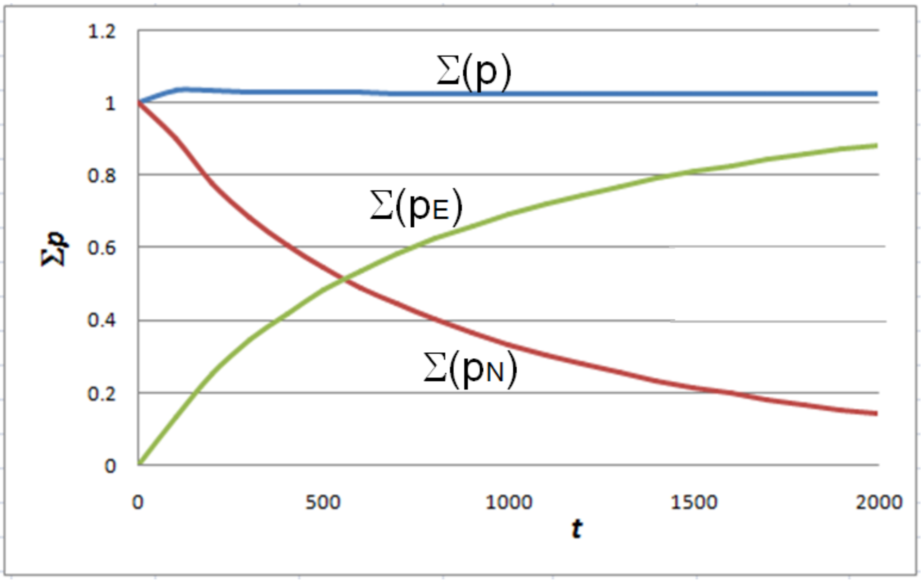


Figure 5.3.4 Time-development of sums of probabilities in diffusion approximation

5.3.3 Equilibrium distributions with more general statistics

Kobayashi (1974b) proposes a generalisation of the process-dependent parameters that enter into his derivation of the equivalent of (5.3.6-7). In effect, wherever (1+ρ) appears it is replaced as follows[49]:

$$(1+\rho) \rightarrow \left(c_a^2\rho + c_b^2\right) \tag{5.3.8}$$

Referring back to the c_a versus I_a discussion earlier in Chapter 3, if it is assumed that $I_a=c_a^2\rho$ then the RHS of (5.3.8) loses its explicit dependence on ρ and can no longer be properly matched with the LHS. However, comparing (5.3.8) with (5.3.1) suggests the identification $I_a=c_a^2$ and $I_b=c_b^2$. Taking (5.3.8) at face value, if c_b is set to zero to reflect M/D/1 then the time developments follow a similar pattern with reduced moments as shown by Figure 5.3.5, where the initial queue has been appropriately reduced, and once again Kobayashi's version performs correctly (although the discrete distribution as calculated is not perfectly normalised).

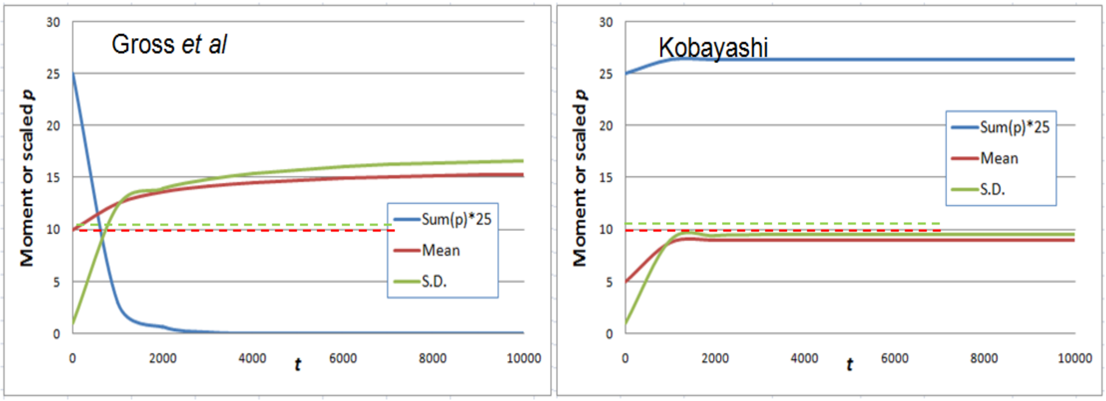


Figure 5.3.5 Time-development of approximations to M/D/1, ρ=0.95, L_0=5, c_b=0

Although the moments approach the expected values in general terms the results are not precise enough to distinguish between with and without unit-in-service. The moments are too low and therefore nearer the latter, but p_0 tends towards 0.095 for ρ=0.95 and 0.829 for ρ=0.5, which values are too high. The approximation certainly works best for M/M/1 and for ρ~1.

Kobayashi (1974a) finds the equilibrium solution of (5.3.1) with (5.3.8):

$$p(x) = \frac{2(1-\rho)}{\left(c_a^2\rho + c_b^2\right)} \exp\left\{-\frac{2(1-\rho)x}{\left(c_a^2\rho + c_b^2\right)}\right\} \tag{5.3.9}$$

[49] In Kobayashi's paper his c_x are the *squares* of the coefficients of variation.

Integration over unit intervals just produces the geometric series:

$$\int_i^{i+1} p(y)dy = \hat{\rho}^i - \hat{\rho}^{i+1} = (1-\hat{\rho})\hat{\rho}^i \qquad \text{where} \qquad (5.3.10)$$

$$\hat{\rho} = \exp\left\{-\frac{2(1-\rho)}{\left(c_a^2\rho + c_b^2\right)}\right\} \qquad (5.3.11)$$

which makes for the obvious identification:

$$p_i = (1-\hat{\rho})\hat{\rho}^i \qquad (5.3.12)$$

However, in the M/M/1 analogy, $\hat{\rho}$ is not equal to ρ, so an improved discrete interpretation is proposed, the singly-nested distribution (3.6.3-4) in Chapter 3, but even that is not perfect.

Figure 5.3.6 shows how this distribution compares with the continuous distribution, where it can be seen that the match between p_0 and the integral over the interval [0,1] becomes less accurate at smaller values of ρ.

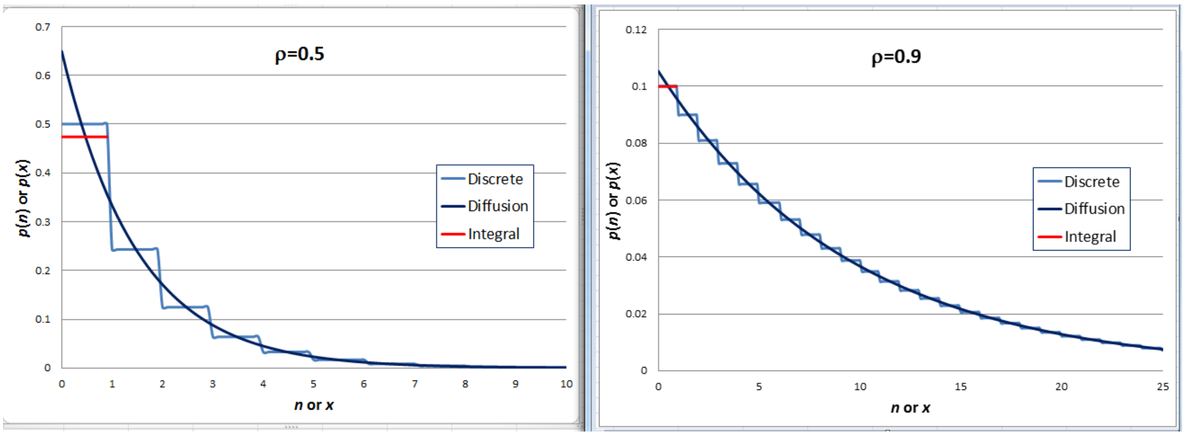


Figure 5.3.6 Comparison of nested discrete and continuous equilibrium distributions

Recalling Whitt (1982), there are many alternative versions of a waiting time formula, and hence the closely related mean queue formula, leading to different generalisations of (1+ρ), none of which appears completely satisfactory, so for this practical application it may be sufficient to accept whatever works well enough.

5.3.4 Consequences for variance of assuming Kobayashi's formula

Recalling equation (2.3.51) for the deterministic limit of the variance, if L_e is substituted from the P-K formula, the simple M/M/1 form is replaced by the rather awkward:

$$V = \left(2\left(\left(I^* - \tfrac{1}{2}\right) + (C - I)\rho\right)\lambda + \mu\right)t = \left(\left(2(I-1) + I_a + \left(c_b^2 + 1 - 2I\right)\rho\right)\lambda + \mu\right)t \qquad (5.3.13)$$

When I=1, I_a=1 and c_b^2=1, (5.3.13) collapses to the simple deterministic result. If it is then assumed that I_a=$c_a^2\rho$:

$$V = \left(\left(2(I-1) - (2I-1)\rho + \left(c_a^2 + c_b^2\right)\rho\right)\lambda + \mu\right)t \qquad (5.3.14)$$

Apart from the seductive alignment of the coefficients of variation without the ρ factor on c_a^2, it is hard to see what this is saying in practical terms. A particular oddity is that c_b^2 appears in association with λ rather than μ. This casts doubt on the generality of (5.3.8).

5.4. CONTINUUM APPROACHES FOR EQUILIBRIUM QUEUES

5.4.1 Exponential solution of the Fokker-Planck diffusion equation

The Fokker-Planck Equation (FPE) (5.3.1) with constant drift and diffusion can be written more generally as (using the coefficients α and β commonly found in some standard works):

$$\frac{\partial p(x,t)}{\partial t} = \frac{\alpha}{2}\frac{\partial^2 p(x,t)}{\partial x^2} + \beta\frac{\partial p(x,t)}{\partial x} \qquad (5.4.1)$$

The continuous exponential distribution is already known to be a solution of the static form of (5.4.1), and can be related to a discrete geometric distribution as in equations (5.3.8-11). However, the corresponding 'traffic intensity' $\hat{\rho}$ is *not* equal to ρ. An alternative equilibrium solution in which it *is* equal to ρ is the exponential distribution:

$$p_E(x) = -\rho^x \ln\rho = \nu e^{-\nu x} \qquad \text{where} \quad \nu = \frac{2\beta}{\alpha} = -\ln\rho = p_E(0) \qquad (5.4.2)$$

When discrete probabilities are identified with integrals over unit intervals, as before, this satisfies all the usual relationships making it analogous to the discrete geometric distribution:

$$\int_0^\infty p_E(x)dx = 1, \quad p_0 = \int_0^1 p_E(y)dy = 1-\rho, \quad p_i = \int_i^{i+1} p_E(y)dy = (1-\rho)\rho^i \quad (5.4.3)$$

However, the moments of the geometric distribution do not exactly match the exponential:

$$L_e = \int_0^\infty x p_E(x)dx = \frac{1}{\nu} \qquad V_e = \int_0^\infty x^2 p_E(x)dx = \frac{1}{\nu^2} \qquad (5.4.4)$$

The *relationship* between the variance and the mean is the same as for a *discrete* exponential distribution (as e.g. used to describe headways), but differs from that for the geometric distribution. The mean itself has a different form from that of M/M/1, suggesting that (5.4.2) is in some sense a limiting case of M/M/1 rather than a direct analogy. Figure 5.4.1 graphs the discrete M/M/1 geometric and continuous exponential functions, confirming that they match well over the range. The red lines represent the integral of the continuous function over [0,1) which is indistinguishable from the discrete p_0 in these cases.

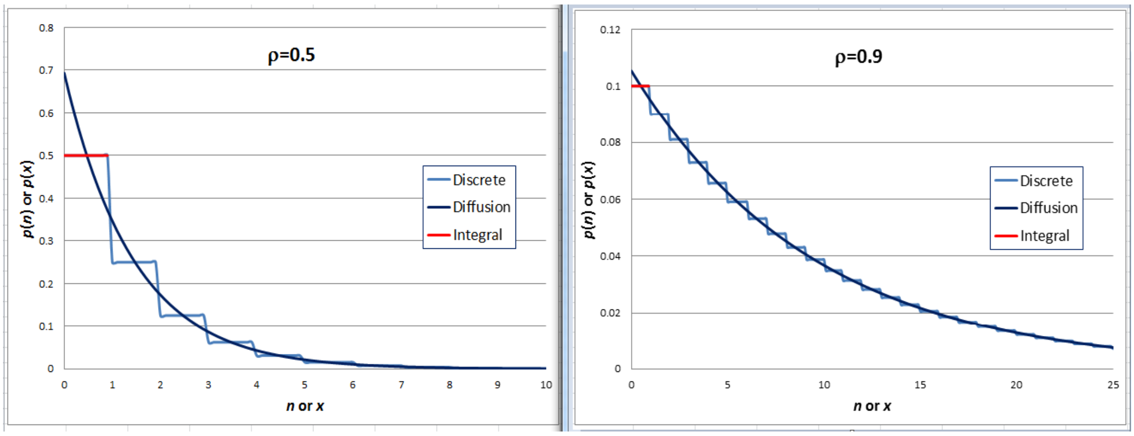


Figure 5.4.1 Comparison of discrete M/M/1 and continuous equilibrium queue distributions

However, this begs the question what is the meaning of $p_E(x)$ at non-integral values of x, and in particular the meaning of $p_E(0)$. A way to explore this is to move from a discrete to a continuous interpretation by gradually reducing the step size, but (5.4.1) is difficult to simulate numerically since, when x is discretised in steps of δ (say), it reduces to a Markov process similar to (2.3.22-23), to which the following apply but offer no new insight:

$$\sum_{i=0}^{\infty}\left(\rho^{\delta}\right)^{i} = \frac{1}{1-\rho^{\delta}} \qquad \frac{\Delta p_i}{\Delta x} = \frac{p_{i+1}-p_i}{\delta}\ (i>0) \qquad \frac{\Delta p_0}{\Delta x} = -p_0$$

$$\frac{\Delta^2 p_i}{\Delta x^2} = \frac{p_{i+1}-2p_i+p_{i-1}}{\delta^2}\ (i>0) \qquad \frac{\Delta^2 p_i}{\Delta x_2} = \frac{p_0-p_1}{\delta} \qquad (5.4.5)$$

Integrating (5.4.1) with respect to x yields the rate of change of the 'zeroth moment', which is of course zero, the RHS rearranging to:

$$\left.\frac{dp}{dx}\right|_{x=0} = -\nu p(0) \qquad (5.4.6)$$

Moments of (5.4.1) can be calculated by using integration by parts with respect to x over the range $[0,\infty)$, and the general results $p(\infty,t)=0$, $\int p=1$, $\int xp=L$ etc. The first moment yields the time-dependent deterministic relationship (5.4.7), consistent with static interpretation of ν and $p(0)$, leads to an identification of capacity μ, and is analogous to the rate of change of the discrete mean queue (2.3.24):

$$\frac{dL}{dt} = \int_0^{\infty} x\frac{\partial p}{\partial t}dx = \frac{\alpha}{2}\int_0^{\infty} xp^{//}dx + \beta\int_0^{\infty} xp^{/}dx = \frac{\alpha}{2}\left[xp^{/}\right]_0^{\infty} + \int_0^{\infty}\left(\beta x - \frac{\alpha}{2}\right)p^{/}dx$$

$$= 0 + \left[\left(\beta x - \frac{\alpha}{2}\right)p\right]_0^{\infty} - \beta\int_0^{\infty} p dx = \frac{\alpha}{2}p(0) - \beta \qquad (5.4.7)$$

Hence:

$$\frac{dL}{dt} = \left(p(0) - \nu\right)\mu \qquad \text{where} \quad \nu = \frac{2\beta}{\alpha}\,\text{(as before)} \qquad \text{and} \;\; \mu = \frac{\alpha}{2} \qquad (5.4.8)$$

Note, however, that while the second of equations (5.4.2) is satisfied, meaning that *dL*/*dt* is zero at equilibrium, the variables do not have the usual relationship to ρ. Since the coefficients on the RHS are constants, integrating (5.4.8) over time gives the deterministic queue formula:

$$L(t) = L_0 + \left((1-\nu) - x(t)\right)\mu t \qquad \text{where} \quad x(t) = \frac{1}{t}\int \left(1 - p(0)\right)dt \quad (5.4.9)$$

The second moment of (5.4.1), evaluated similarly, is:

$$\frac{d\left(V + L^2\right)}{dt} = \int_0^\infty x^2 \frac{\partial p}{\partial t}dx = \frac{\alpha}{2}\int_0^\infty x^2 p^{//}dx + \beta\int_0^\infty x^2 p^{/}dx = \frac{\alpha}{2}\left[x^2 p^{/}\right]_0^\infty + \int_0^\infty (\beta x - \alpha)xp^{/}dx$$

$$= 0 + \left[(\beta x - \alpha)xp\right]_0^\infty - \int_0^\infty (2\beta x - \alpha)p\,dx = 0 + 0 - 2\beta L + \alpha \qquad (5.4.10)$$

Substituting the coefficient values from (5.4.8), and then integrating over time:

$$\frac{d\left(V + L^2\right)}{dt} = 2(1 - \nu L)\mu \qquad (5.4.11)$$

$$\left(V + L^2\right) = 2\nu\left(\frac{1}{\nu} - D\right)\mu t + \left(V_0 + L_0^2\right) \qquad \text{where} \quad D = \frac{1}{t}\int L\,dt \qquad (5.4.12)$$

Remembering the form of L_e in (5.4.4), equation (5.4.12) has the form of the discrete variance formula (2.3.27) apart from missing terms L and L_0. In (2.3.27) these come from the p_0 term in (2.3.26), which in turn arises from the difference between the recurrence relations (2.3.22) and (2.3.23) caused by the boundary at zero. As step size δ in the discrete distribution is reduced from 1 to 0 the influence of the boundary at zero decreases, and the contribution of this term decreases linearly, so the definition of W (equation 2.3.29) can be generalised to:

$$W \equiv V + L(L + \delta) \qquad (5.4.13)$$

Thus not only the deterministic queue formula, as would be expected through conservation, but also the deterministic variance formula, arise naturally from the FPE diffusion process.

5.4.2 Adding arrival and service statistics through Gamma approximation

The essence of the modification by Kobayashi (1974a) lies in the relationship between the $\hat{\rho}$ that fits the exponential model (5.4.2) and the real ρ, as per (5.3.8-11). However, the distribution itself is not changed. In reality, the distribution always changes. One possibility would be to make $\hat{\rho}$ a non-linear function of ρ. The problem with this is evident from the M/D/1[G] results, that p_0, which nominally should be the integral of the function over [0,1], is increasingly out of scale with the rest of the distribution as G increases. Although it is straightforward to fit a discrete doubly-nested geometric distribution, it would be very difficult to fit a continuous function.

In addition to a shift of origin, a useful distribution needs two parameters, eliminating Poisson. The Gamma distribution has several advantages: its mode can be 0 or >0; one of its parameterisations yields the exponential distribution and a limiting parameterisation leads to the Normal distribution. Apart from this, as will be shown, simulation also gives empirical support for the Gamma distribution as an approximation to the extended distribution $\{p^*_{-G} \ldots p_{i>0}\}$ when its origin is shifted appropriately[50]. Technically, the approximating distribution is Erlang (2.2.3) because it is calculated discretely, but the ability to make non-integral adjustments to parameters in Gamma is an advantage. The basis of fitting Gamma is to set the shape and scale/rate parameters according to known invariants of the target distribution, the mean and variance, and in the case of M/D/1[G] the mode. Properties of the Gamma distribution, using k and ν as the shape and rate parameters, are given by:

$$Gamma(k,\nu,x)=\frac{k\nu(k\nu x)^{k-1}e^{-k\nu x}}{\Gamma(k)} \qquad (5.4.14)$$

$$Mean=\frac{1}{\nu},Variance=\frac{1}{k\nu^2},Mode=\frac{(k-1)}{k\nu},Maximum=\frac{[(k-1)/e]^{k-1}}{k\Gamma(k)\nu} \qquad (5.4.15)$$

Using results from Chapter 3, the shifted queue size distribution satisfies:

$$Mean=\sum_{-G}^{\infty}(i+G)p_i^{(*)}=\sum_{-G}^{0}ip_i^{(*)}+\sum_{0}^{\infty}ip_i+G\sum_{-G}^{\infty}p_i^{(*)}=L_e-L^*+G=L_e+G\rho$$

(5.4.18)

$$Mode\approx G\rho \qquad (5.4.19)$$

[50] Olszewski (1990) implies that a Gamma or possibly Negative Binomial distribution can be used to describe the probability distribution of a *time-averaged* signal overflow queue, but as this is not origin-shifted is not clear that it has any bearing on the distribution of an extended M/D/1[G] distribution including notional queue states.

The extended M/D/1[G] queue size probability distributions can be approximated by:

$$p_{i(est)}^{(*)} \cong Gamma\left(k^*(\iota,\kappa),\kappa\nu,i+G\right) \qquad i\in[-G,\infty) \qquad (5.4.20)$$

where ι and κ are adjustments to the mode and mean respectively such that equations (5.4.15) are satisfied by the parameters k and ν as modified, and the origin is shifted by G. The 'real' value of p_0 is obtained by summing components $-G$ to 0, as described earlier in Chapter 2. Numerical fitting can be achieved in most cases using a standard method such as Excel's Solver (a Newton method), with the target to minimise the simple RMS error between the simulated and calculated probability components, but this may not work where ρ is small and G large. The Anderson-Darling difference measure has also been tested but gives a poor fit.

To estimate k the variance is needed. The results of Chapter 3 enable the variance of the notional probability terms to be calculated from the *Variance* of the extended distribution and the 'real' part[51]. Now k can be estimated from moments of the extended distribution:

- Given L_e estimate *Mean* from (5.4.18)
- Estimate *Variance* by the procedure in Chapter 3 (equations 3.7.23-25).
- Estimate k as *Mean*^{2}/*Variance* or from the formula for *Mode*

To fit the distribution it is now necessary to optimise the adjustments, for example by minimising the error between the estimated and simulated distributions. Several methods exist, differing in the way they weight the queue size probabilities. The Cramér-von Mises Statistic is just unweighted least squares, although expressed formally as an integral, while the Anderson-Darling Statistic places greater emphasis on the ends of the distribution (e.g. Laio 2004). In this case simple least-squares is preferred precisely because it does not emphasise what are likely to be the least reliable regions of an approximate distribution.

Matching the maxima of the distributions helps guide automatic selection of the adjustments. To do this, using similar methods to those of Chapter 3 to approximate M/D/1[G] moments, the maxima $h(\rho,G)$ of the simulated extended probability distributions are approximated by the following, where the second equation defines a link function as discussed previously:

$$h(\rho,G) = 0.2(1-\rho)(1-\ln(H(\rho,G))) \qquad \text{where} \qquad (5.4.21)$$

[51] As in Section 3.7 we use italics to identify the moments of the *unshifted* notional distribution.

$$H(\rho, G) = \frac{(G - (1 - \rho))\rho}{\tau_{re}} \tag{5.4.22}$$

This approximation achieves absolute RMS error of 0.042 in normalised predictions of maxima values in the range [0.2,1.5], giving the trend shown in Figure 5.4.2.

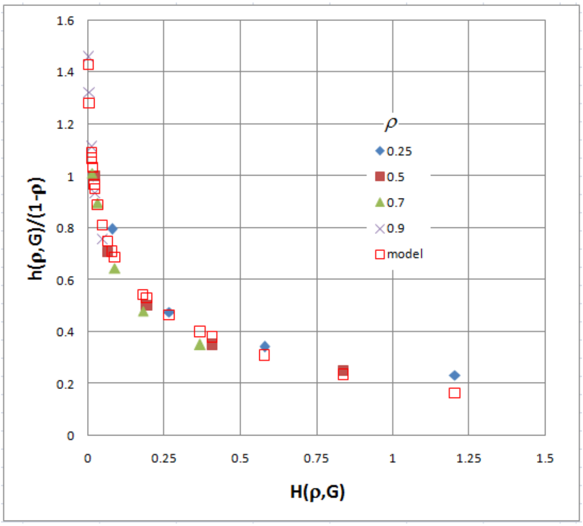


Figure 5.4.2 Linking of transformed maxima of Gamma distributions and estimates

For the purpose of testing, a least-squares fit using manual adjustment of parameters with step size of 0.05 has been used, giving the parameters, fit and errors graphed in Figures 5.4.3-6.

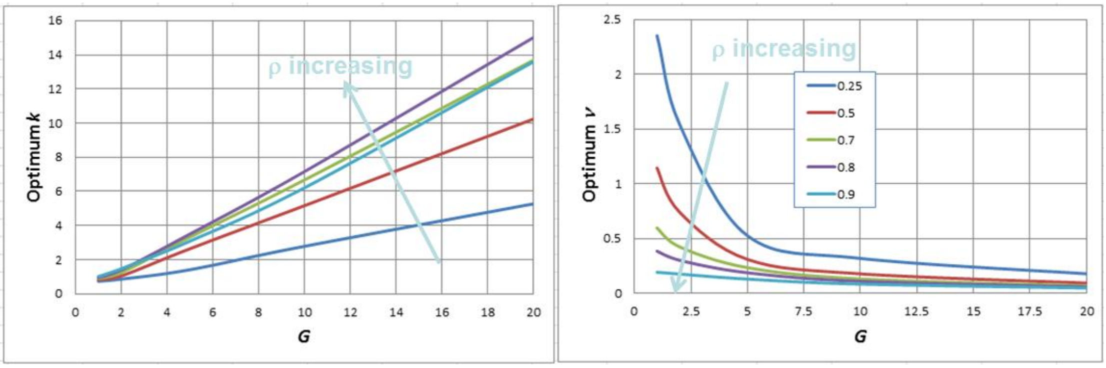


Figure 5.4.3 Optimum *k* and ν parameters for Gamma distributions

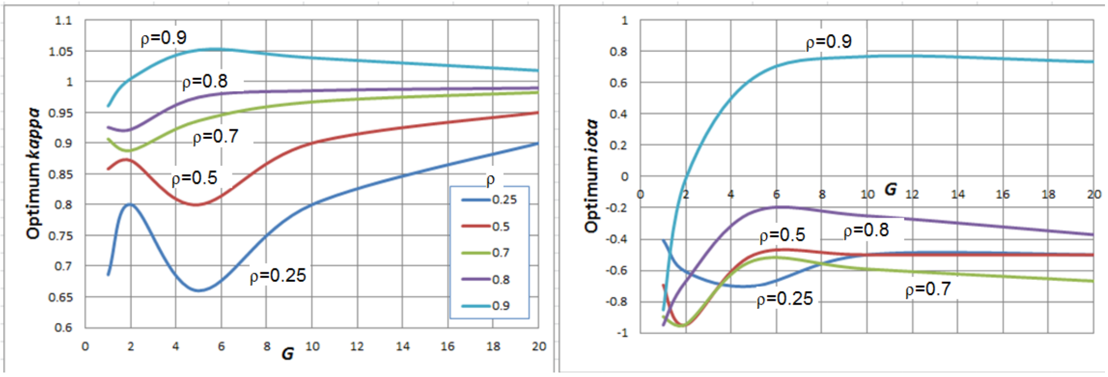


Figure 5.4.4 Optimum κ and ι adjustments for Gamma distributions

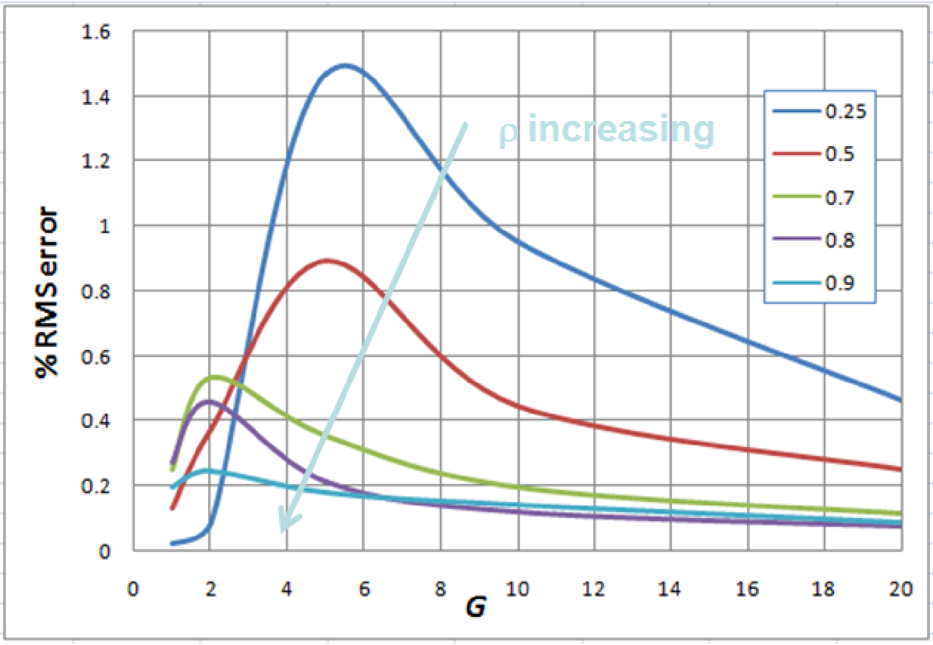


Figure 5.4.5 Percent RMS error of Gamma distribution fits for various values of ρ

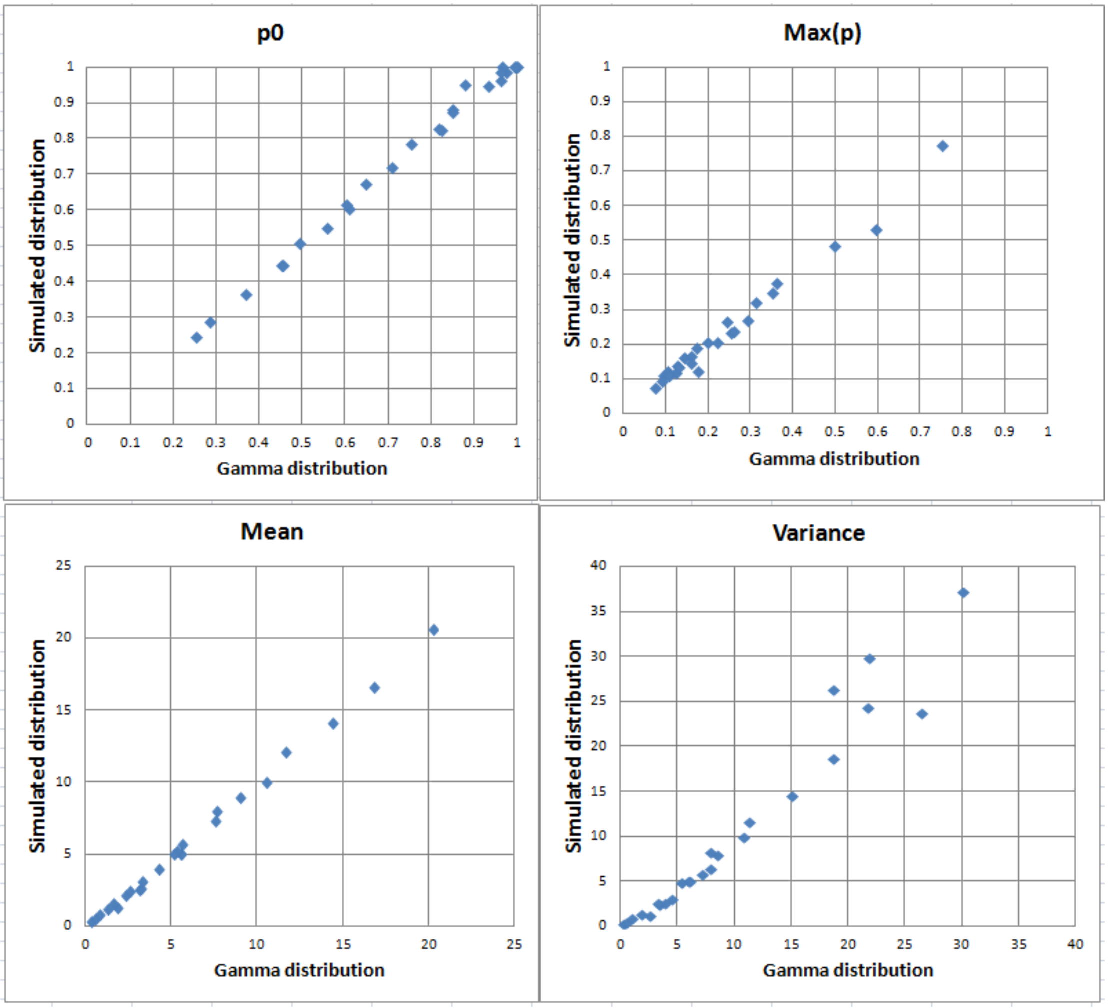


Figure 5.4.6 Fit between simulated M/D/1[G] and Gamma distribution moments

The error is greater the smaller the value of ρ but remains moderate on average, although the match of variance can be poor. Figure 5.4.7 shows the graphical fit between the simulated (solid) and Gamma (broken) distributions for four ρ values.

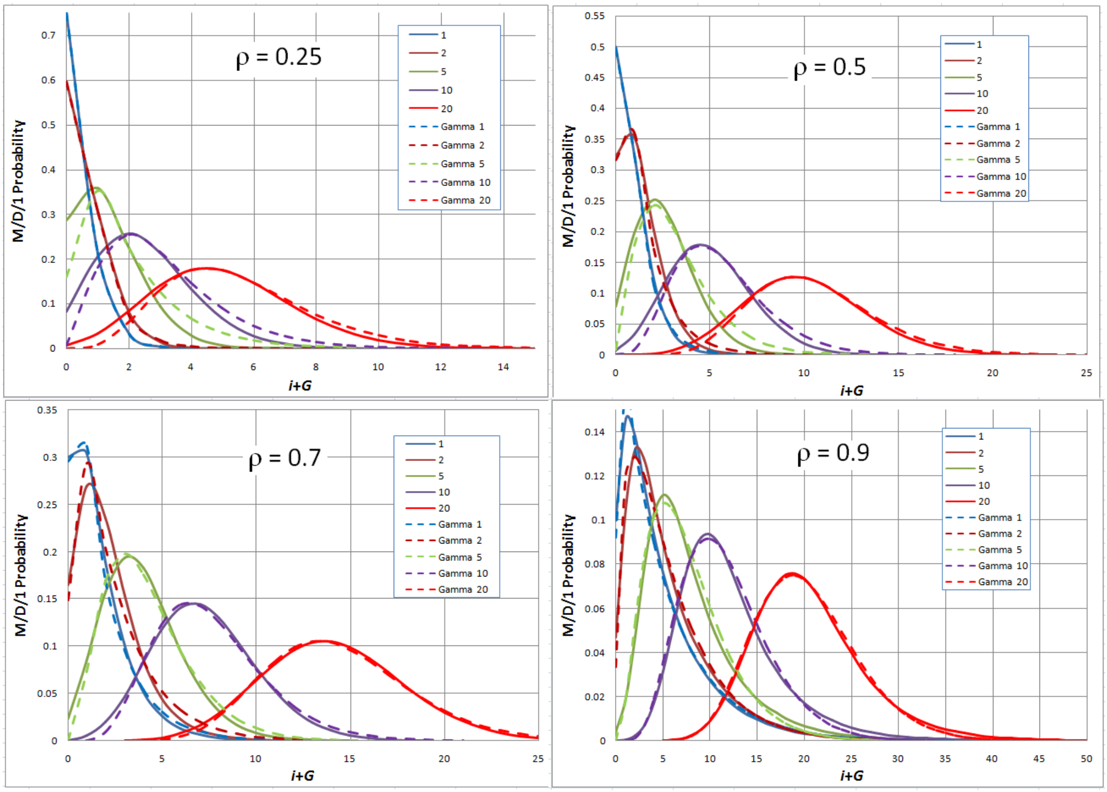


Figure 5.4.7 Simulated extended M/D/1[G] and adjusted Gamma distributions

An alternative to adjustment could be to match the known p_0 of the queue distribution with the equivalent property of the Gamma distribution, or match the part of the equivalent of average $\bar{p}_0$ to the expected value of (1-ρ). Matching p_0 requires the integral of the Gamma or its cumulative distribution[52], while matching $\bar{p}_0$ requires the first moment of notional terms:

$$p_0 = \int_{x=0}^{G+1} Gamma(x) \qquad (5.4.23)$$

$$\bar{p}_0 = 1 - \rho = p_0 - \frac{1}{G}\int_{x=0}^{G+1} x.Gamma(x) \qquad (5.4.24)$$

Figure 5.4.6 earlier shows that a good match of p_0 is achievable, so this might improve the fit from matching just *Mean* and *Variance*. Figure 5.4.8 shows that the Gamma distribution reproduces the values of $\bar{p}_0$ quite well.

[52] For numerical calculations, Microsoft Excel (spreadsheet) provides a cumulative Gamma function as well as the Gamma function, and this may also be available in other programming tools and languages.

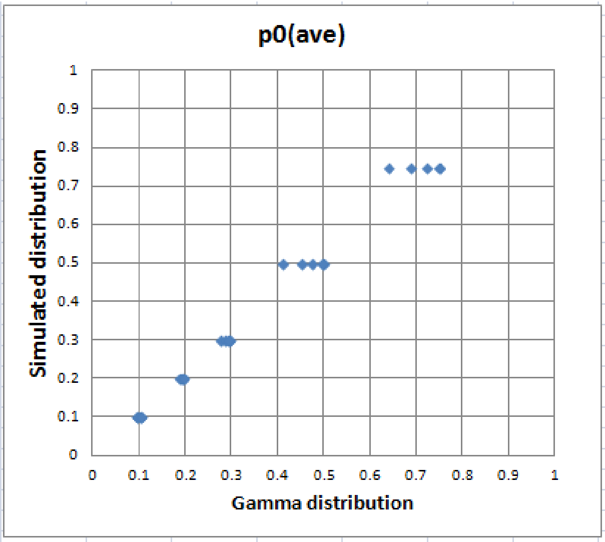


Figure 5.4.8 Fit of simulated average $\overline{p}_0$ with Gamma approximation

5.4.3 M/D/1 growth test with Gamma approximation

Figure 5.4.9 compares Markov simulated distributions (left) at logarithmically advancing times of an M/D/1[5] queue with Gamma approximations (right), where the basic value of ν is the inverse of the mean of the extended Markov distribution. The Markov simulations could be inaccurate for small *t*, as notional probabilities need to be factored up significantly. Significant adjustment is also needed to the Gamma distributions (inset), and the distributions are only broadly matched at small *t*. However, they converge as equilibrium is approached.

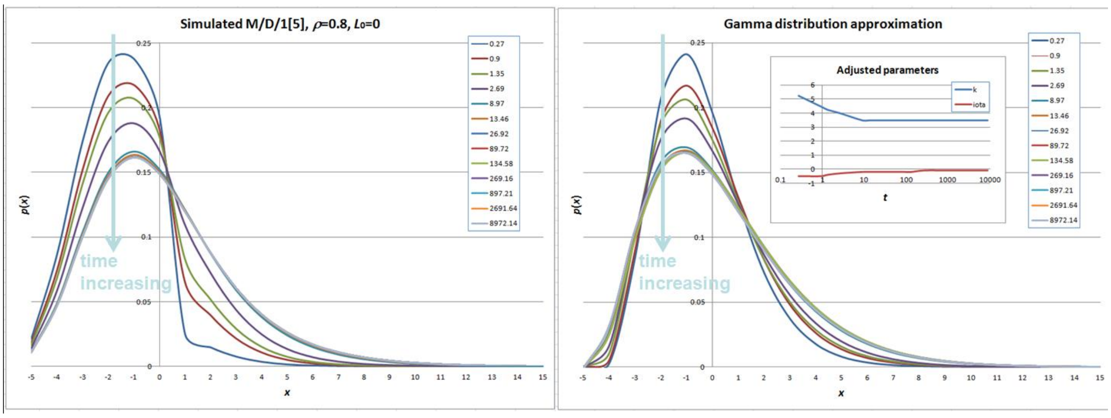


Figure 5.4.9 Comparison of growth to equilibrium with Gamma approximation (Logarithmic time scale - peaks subside increasingly slowly as time progresses)

While the Gamma distribution can fit M/D/1[G] distributions quite well, the adjustments to parameters needed are somewhat erratic, and the results are found to be rather sensitive to them. Since the 'real' distributions have mode zero they may fitted more simply by the doubly-nested geometric. However, the Gamma model may merit further investigation.

5.4.4 Other candidate distributions and their asymptotic properties

From (5.4.14), for large *x* the ratio of Gamma function values of successive integer states is:

$$r_\Gamma = \frac{Gamma(x+1)}{Gamma(x)} = \left(1+\frac{1}{x}\right)^{k-1} e^{-k\nu} \approx \left(1+\frac{(k-1)}{x}\right)e^{-k\nu} \qquad (5.4.25)$$

In the case *k*=1 the Gamma distribution is precisely the exponential distribution (5.4.2). For *k*>1, once *x* is sufficiently large compared to *k* the above ratio tends to a constant value given by equation (5.4.26), so Gamma becomes exponential asymptotically:

$$r_\Gamma \to e^{-k\nu} \qquad \text{as } x \to \infty \qquad (5.4.26)$$

Using similar methods LogNormal and Poisson distributions have been tested as approximations to M/D/1[G] extended distributions, in the former case matching the mean and mode, and in the latter case just the mean.

Comparing Figures 5.4.10 and 5.4.11 (overleaf) with 5.4.7 suggests that these distributions perform worse than Gamma, especially for small *G*. In particular, while the LogNormal performs reasonably well for larger values of the mean (smaller s.d./mean ratio), where it is more similar to the Normal, it performs very poorly for small values of mean where its asymmetrical character is most evident.

For the LogNormal, the ratio has a rather complex form that after declining tends very slowly back up towards 1, and is therefore not representative of an equilibrium distribution:

$$r_L = \left(\frac{x+1}{x}\right)^{-\frac{(\ln(x(x+1))-2\mu)}{2\sigma^2}} \to 1 \text{ as } x \to \infty \qquad (5.4.27)$$

Although the Poisson distribution looks broadly similar, its ratio between integer state probabilities declines to zero, so is also not representative of an equilibrium distribution:

$$r_P = \frac{\lambda e^{-\lambda}}{x+1} \qquad \to 0 \text{ as } x \to \infty \qquad (5.4.28)$$

In addition, neither distribution includes the exponential distribution as a particular case, so neither appears suitable for approximating an equilibrium queue distribution.

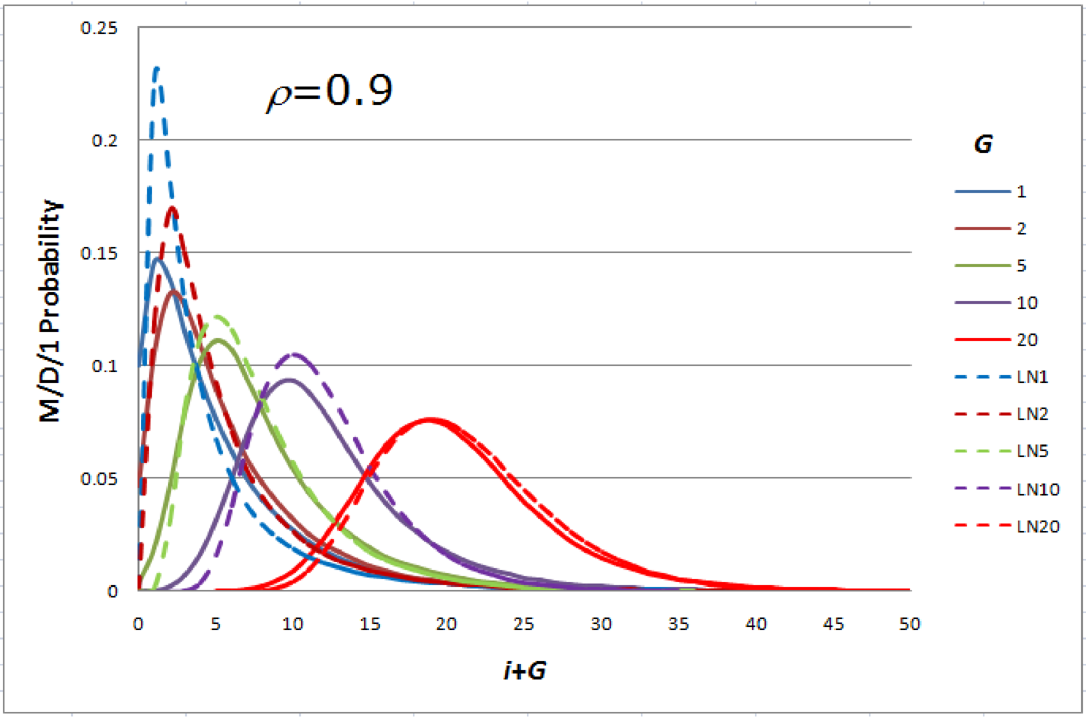


Figure 5.4.10 Simulated M/D/1[G]and LogNormalfor ρ=0.9, mean and variance matched

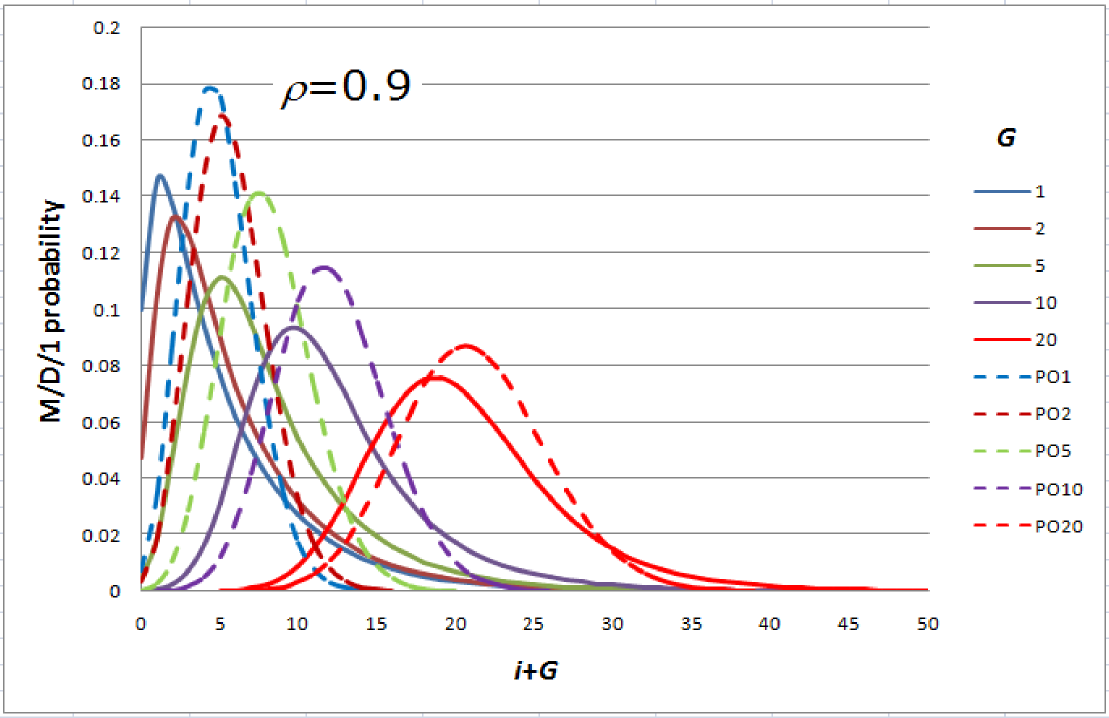


Figure 5.4.11 Simulated M/D/1[G] and Poisson for ρ=0.9, mean matched

More exotic distributions include the Conway-Maxwell-Poisson distribution, a generalised form of Poisson which can match non-zero p_0 without needing an origin shift, but has no closed form. Others with a smooth Poisson-like shape include the Gaudin and Tracy-Widom distributions (which turn up as limiting distributions in connection with Random Matrices).

However, neither has a closed form expression, nor an obvious physical connection with simple queues. The convenient properties of the exponential, Gamma and Normal distributions favour their use even should some more complex distribution turn out to be a ‘truer’ representation.

5.4.5 Fitting general time-dependent distributions

It is clearly important that M/M/1 should be covered by the set of equilibrium approximations, and it now appears equivalent to be a limiting case of M/D/1[G] where the service interval has shrunk to zero. With three generating parameters, k, ν and an origin shift, ideally optimised by fitting p_0, L and V, the Gamma distribution may be suitable for approximating other unimodal queue size distributions such as those covered in Chapter 3.

So far, mainly equilibrium distributions have been fitted. Unimodal distributions may have enough degrees of freedom to fit p_0, L and V where a target dynamic distribution is unimodal, but none can reproduce the bimodal distributions seen in Figures 2.5.4 and 5.2.1. This is not a problem for a *combination* of distributions, however. The earlier results of the diffusion approximation suggest that such a combination can consist of two Normal drifting-and-diffusing components and one equilibrating component that retains its basic shape but changes in scale and/or amplitude over time. This is convenient if the statistics of the queue process are assumed to be constant. The combination is then in some sense equivalent to the combination of deterministic and random components in the sheared approximation. Such a quasi-static approach is not without precedent. Rider (1976) defines a variable $\sigma(t)$ such that the system initially has the distribution $p_i=(1-\sigma)\sigma^i$, relaxing to the final distribution $(1-\rho)\rho^i$.

5.5. COMBINATION APPROACHES TO APPROXIMATING DISTRIBUTIONS

The preceding diffusion solutions can be interpreted as an equilibrating component and two relatively shifted diffusing components, but apart from their inherent complexity they strictly describe time development from an initial 'shock', so to describe time development in general needs convolution with an initial distribution. However, for present purposes, a *time-dependent* function for the probability distribution is not needed because time-dependence is handled by the analytical approximations to the queue moments. All that is required is to fit a *time-stamped* static function to these moments. Given the three properties p_0, L and V, the objective of this Section is to fit an approximate distribution that can accommodate all conditions from initial through dynamic to equilibrium. Such a distribution can be defined by:

$$p(x) = A(x) p_E(x,k,\nu) + B(x) p_D(x,m,s) \qquad (5.5.1)$$

subject to:

$$\int_0^\infty p(x) = 1 \qquad (5.5.2)$$

Here it is assumed that the equilibrium component is exponential or Gamma, and the dynamic component is Normal or similar, so each is specified by at most two parameters. As has been shown earlier, 'real' states of an equilibrium distribution, with the possible exception of p_0 and p_1, can be represented by an exponential function, although Gamma might be a more flexible alternative, and a Normal function appears to be the most satisfactory candidate for the dynamic component, although LogNormal could be considered.

The natural choices for $A(x)$ and $B(x)$ are either constants or exponential functions, e.g. $e^{-\theta x}$, both of which can satisfy the basic FPE when combined with an exponential or Normal function since they result in functions of the same form. The known value of $p(0)$, derived from p_0 as discussed in the next Section and elsewhere, can be accommodated specially by appropriate calibration of the weighting factors, or more generally by ensuring:

$$A(0) = 1, \qquad B(0) p_D(0,m,s) = 0 \qquad (5.5.3)$$

The following Sections address specific issues of combining continuous distributions to approximate a discrete queue size distribution and explore some alternative approaches.

5.6. MATCHING CONTINUOUS AND DISCRETE DISTRIBUTIONS

5.6.1 Motivation and approach

While relating a discrete probability distribution to a continuous function has obvious advantages for approximations, calculation and understanding of moments, identification of family resemblances, and dynamic estimation, it raises the questions of how conversion can be done consistently, how the zero of the continuous axis should be interpreted, and whether differences between the impacts on different moments matter.

5.6.2 Relating discrete probabilities to intervals of the continuous function

Alternative ways to relate a discrete distribution to a continuous analogue are given by:

$$p_i = \int_{i+h-0.5}^{i+h+0.5} p(x)dx \tag{5.6.1}$$

$$p_i = p(i+h) \tag{5.6.2}$$

One might think that a sensible choice of *h* would be 0, but this requires *p*(*x*) to be defined for *x*<0, which is likely to be inconvenient. If *h* is set to 0.5, then (5.6.1) leads to the convenient relationships given earlier in sub-section 5.4.1. Equation (5.6.1) converts exactly between the exponential function used there and the discrete geometric distribution. While this may not apply to other equilibrium distributions, the tendency of such distributions to revert to an exponential/geometric form at higher states makes it a reasonable practical assumption.

Equation (5.6.2) is accurate to the extent that *p*(*x*) is symmetrical or linear over a unit interval. If *h* were a function of *i* and corresponded to the centroid of the interval, the two versions would correspond. For exponential, the adjustment *h* is independent of *i* and given exactly by:

$$h_{optimum} = \frac{1}{\ln \rho} \ln\left(\frac{1-\rho}{-\ln \rho}\right) \tag{5.6.3}$$

The value of *h* so defined is always less than 0.5, but does not fall below 0.4 until ρ is below 0.1, so *h*=0.5 can be considered a reasonable practical assumption, given that the distribution matching exercise will in any case be an approximation based on approximate data.

However, what amounts to a shift in the continuous origin leads to differences between the moments, as shown earlier by equations (5.4.4). The corresponding discrete states are in effect shifted by +0.5. This does not affect the sum of the probabilities, but:

$$\tilde{L} = \sum_{0}^{\infty} (i+h) p_i = L + h \tag{5.6.4}$$

$$\tilde{V} + \tilde{L}^2 = \sum_{0}^{\infty} (i+h)^2 p_i = V + L^2 + 2hL + h^2 \quad \text{hence}$$

$$\tilde{V} = V + L^2 + 2hL + h^2 - (L+h)^2 = V \tag{5.6.5}$$

So the only correction needed is to the mean. *Thus, h must be added to the analytically calculated mean before estimating the parameters of a continuous distribution to match it.* When matching to a *simulated discrete distribution* for testing purposes, it is sufficient to place the discrete probabilities at points *i+h*, whence the discrete mean will be augmented automatically.

5.6.3 Interpreting the origin of the continuous function

One further point concerns the nature of $p(0)$, which is intimately connected to the parameter ν of the continuous equilibrium distribution in Section 5.4. For the exponential function, (5.6.1) leads to the relationship:

$$p(0) = \nu = -\ln(\rho) = \ln(1 - p_0) \tag{5.6.6}$$

This transformation is essential to generate the correct continuous distribution, but it means that $p(0)$ and p_0 as graphed will differ. Going back to the previous Section, *estimation of the continuous approximation will be much simplified if only the equilibrium component is assumed to exist at x=0.* Therefore its parameter ν can be calculated from the p_0 of the analytical approximation, analogous to the quasi-static utilisation *x* in the sheared approximation. This requires the dynamic continuous component to be zero at *x*=0, which can be assured by suitable choice of weighting functions. It does mean, however, that this component does not vanish entirely on [0,1], so in principle, p_0 is influenced not only by the form of the equilibrium component but by the dynamic component too. In practice this can probably be ignored, again appealing to the approximate nature of the exercise.

5.6.4 Effect of weighting on the Gamma distribution used as an equilibrium function

If p_E is chosen to be an exponential equilibrium distribution, exponential weighting is straightforward, as described later.

If p_E is chosen to be a Gamma distribution, the shape parameter k can produce a peaked unshifted distribution. For M/D/1[G] (signal-type) processes, the distribution will then need to be shifted by G places (approximately), and p_0 will be calculated by summing notional state probabilities where relevant. For G/M/r processes (i.e. infinitesimal service period), G=0. All variables except x and k can be functions of time. Finite moment integrals of the Gamma distribution can be expressed in terms of the Cumulative Gamma with modified parameters.

$$\int_0^y x.Gamma(x,k,\nu)dx = \frac{1}{\nu}.CumGamma\left(y,k+1,\frac{k\nu}{k+1}\right) \qquad (5.6.7)$$

$$\int_0^y x^2.Gamma(x,k,\nu)dx = \frac{k+1}{k\nu^2}.CumGamma\left(y,k+2,\frac{k\nu}{k+2}\right) \qquad (5.6.8)$$

Exponentially weighting the Gamma distribution by the factor $e^{-\theta x}$ modifies its amplitude and scale without change of shape:

$$\tilde{p}_E(x,\theta,k,\nu) \equiv \left(\frac{\nu}{\nu+\frac{\theta}{k}}\right)^k .Gamma\left(x,k,\nu+\frac{\theta}{k}\right) \equiv A(\theta,k,\nu)p_E\left(x,k,\nu+\frac{\theta}{k}\right)$$

(5.6.9)

5.6.5 Using nested exponential/geometric distributions for the equilibrium component

Although the Gamma distribution seems a good approximation to the various possibilities of equilibrium distribution, it may be asked whether it is really necessary for the present purpose given that the time-dependent approximation of queue evolution will provide instantaneous (end of time-slice) values for $p_{0,}$ L and V, and the known queue statistics will define the corresponding asymptotic equilibrium values for the current time slice. Then the doubly-nested geometric approximation can provide a discrete asymptotic equilibrium distribution with the same degrees of freedom as a shifted Gamma. Figure 5.6.1 confirms that in general the exponential functions that fit the three parts of a doubly-nested geometric distribution, p_0, p_1, and $\{p_i\}$ (i>1), are too different to be merged into even a piecewise-continuous single function.

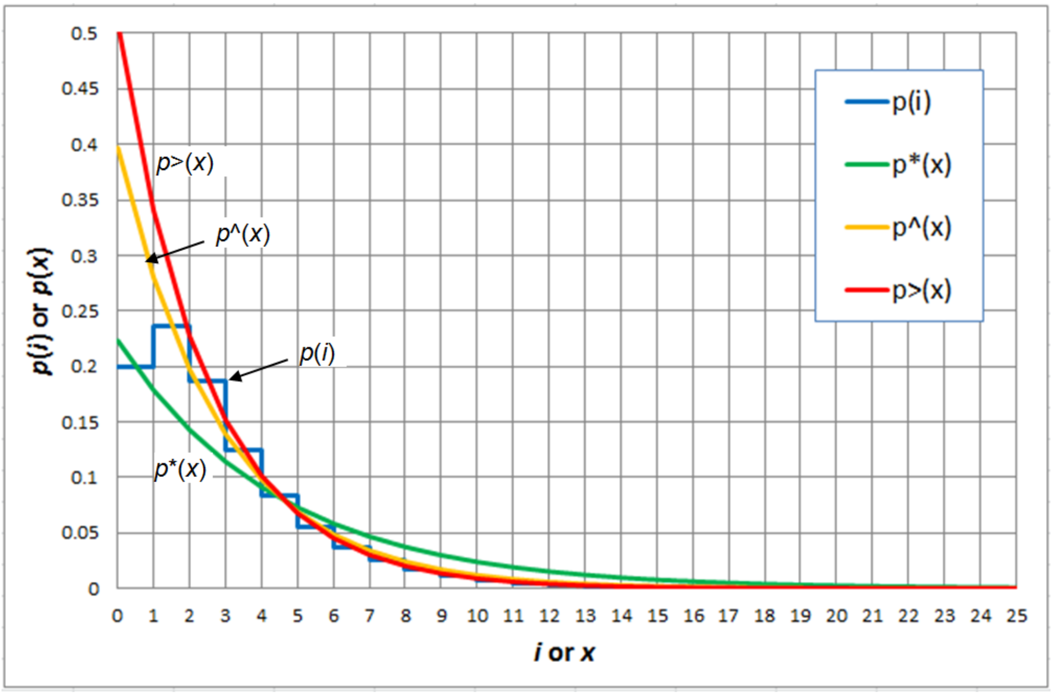


Figure 5.6.1 Fitting exponential functions to parts of doubly-nested geometric distribution

Where p_0 and possibly p_1 depart from the geometric form, it seems most manageable to adjust the mean and variance of the continuous approximation so that only an ‘underlying’ geometric distribution, as represented by the parameter $\vec{\rho}$ introduced earlier in Chapter 3, need be matched. If the shape of the nested distribution is assumed fixed through time, then the other parameters can be calculated from ratios of the asymptotic equilibrium values as determined by the actual ρ and queue statistics. The contributions of the other parts to the mean and variance need to be assessed. Working with the second and third components separately is more complicated than appears justified by the approximate nature of the exercise and the moderate contribution to the moments made by p_1. Thus it is practical to work with a full geometric distribution based on the third $\vec{\rho}$ parameter, which is the default in case one or no levels of nesting are involved.

Since p_0 does not contribute to the continuous mean or variance, only the contribution of p_1 has to be accommodated by them. To do this the exact difference from the contribution of $\vec{\rho}$ could be calculated, but the expressions are complicated because both ends of the integral are finite. By separating out the first two discrete states of the distribution, and treating the rest as continuous, and assuming an exponential weighting factor $e^{-\theta x}$, an approximation to the sum or integral of the probabilities is:

$$\tilde{P}_E \cong (1-\rho^*) + \rho^*(1-\hat{\rho})e^{-\theta} + \rho^*\hat{\rho}\left(\frac{\vec{\nu}}{\vec{\nu}+\theta}\right)e^{-2\theta} \qquad (5.6.10)$$

There is no need to involve *h* here since the discrete components already incorporate this shift. It can be seen that $\tilde{P}_E$ reduces to 1 when θ=0. If the exponential weighting function is treated as quasi-linear in the neighbourhood of *x*=1, then the adjustment for the weighted mean is:

$$\tilde{L}_E \cong \frac{\rho^* \hat{\rho}}{\vec{\rho}^2} \tilde{\vec{L}}_E + \rho^* \left(1 - \frac{\hat{\rho}}{\vec{\rho}}\right) e^{-\theta(1+h)} \qquad (5.6.11)$$

Since 1^2=1, the adjustment to the weighted second cumulative moment V_E+L_E^2 is the same. It is probably simpler to make this adjustment than a more complex adjustment to V_E alone.

There is an implementation issue with these equations in that they involve θ. In order to get a solution, the target value of *L* must be specified, but this now depends, usually to a small extent, on the solution. To avoid a circular reference, the value of θ must be fixed, and this would normally be the initial estimate. In principle, the calculation should be iterated using the value of the previously solved θ. In practice this may not be worthwhile, as the non-geometric nature of the probability distribution is unlikely to be evident unless the queue is close to equilibrium, in which case θ will be small anyway.

5.7. LINEAR COMBINATION APPROACH

This Section considers a linear combination of exponential and Normal distributions using constant weighting factors, as perhaps the simplest approach to a combined distribution:

$$p(x) = Ap_E(x,\nu) + Bp_N(x,m,s) \tag{5.7.1}$$

where the component distributions are defined respectively by:

$$p_E(x,\nu) = -\rho^x \log\rho = \nu e^{-\nu x} \tag{5.7.2}$$

$$p_N(x,s,m) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{s}\right)^2} \tag{5.7.3}$$

Using some weighting factors *A* and *B* the constraints imposed by the moments are:

$$A + \frac{B}{2} erfc\left(-\frac{m}{s\sqrt{2}}\right) = 1 \tag{5.7.4}$$

$$\frac{A}{\nu} + Bs\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{m}{s}\right)^2} + \frac{m}{2s} erfc\left(-\frac{m}{s\sqrt{2}}\right)\right) = L \tag{5.7.5}$$

$$\frac{2A}{\nu^2} + Bs^2\left(\frac{m}{s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{m}{s}\right)^2} + \frac{\left(1+\left(\frac{m}{s}\right)^2\right)}{2} erfc\left(-\frac{m}{s\sqrt{2}}\right)\right) = V + L^2 \tag{5.7.6}$$

$$A\nu + \frac{B}{s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{m}{s}\right)^2} = p(0) \tag{5.7.7}$$

As per equation (5.4.3), p_0 for the exponential component should be calculated as:

$$p_{E0} = \int_0^1 p_E(y)dy = 1 - e^{-\nu} \tag{5.7.8}$$

As an alternative to integrating the Normal over the range [0,1], it can be approximated by the median value, assuming local monotonicity:

$$p_{N0} \equiv \int_0^1 p_N(y)dy \approx p_N\left(\frac{1}{2}\right) = \frac{e^{\left(\frac{4m-1}{8s^2}\right)}}{s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{m}{s}\right)^2} \qquad \text{hence} \tag{5.7.9}$$

$$A\left(1-e^{-\nu}\right)+\frac{Be^{\left(\frac{4m-1}{8s^2}\right)}}{s\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{m}{s}\right)^2}=p_0 \qquad (5.7.11)$$

The following substitutions can now be made, the last two depending on whether $p(0)$ or p_0 is to be fitted according to (5.7.7) or (5.7.10):

$$\frac{1}{2}erfc\left(-\frac{m}{s\sqrt{2}}\right)=\frac{(1-A)}{B} \qquad \text{and} \qquad (5.7.11)$$

$$\frac{1}{s\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{m}{s}\right)^2}=\frac{(p(0)-A\nu)}{B} \qquad \text{or} \qquad (5.7.12)$$

$$\frac{1}{s\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{m}{s}\right)^2}=\frac{\left(p_0-A\left(1-e^{-\nu}\right)\right)e^{-\left(\frac{4m-1}{8s^2}\right)}}{B} \qquad (5.7.13)$$

Noticing that (5.7.12-13) can be used to eliminate B, (5.7.5-6) can be rewritten to give simple weighted sums of the primitive moments with an adjustment term:

$$\frac{A}{\nu}+(1-A)m+E=L \qquad (5.7.14)$$

$$\frac{2A}{\nu^2}+(1-A)\left(m^2+s^2\right)+Em=V+L^2 \qquad (5.7.15)$$

where, depending on the choice between (5.7.12-13), the adjustment term E is either of:

$$E_{(1)}=(p(0)-A\nu)s^2 \qquad \text{or} \qquad (5.7.16)$$

$$E_{(2)}=\left(p_0-A\left(1-e^{-\nu}\right)\right)s^2e^{-\left(\frac{4m-1}{8s^2}\right)} \qquad (5.7.17)$$

Thus there are two equations to fit four unknowns $\{\nu, m, s, A\}$! A possible solution approach is to take account of the dynamics described earlier, the constant rates of drift of the mean and diffusion of the variance of the Normal component, but this raises the question what should be the time origin, as in the case of time-shifting the diffusion approximation. The relaxation behaviour of the exponential component could also be assumed. In principle, fitting both initial and final states would then mean four equations for five unknowns (time origin being the fifth) since the values of ν, m and s would be linked through time, although there seems no natural link between the factors A over time. Despite this improvement, this approach seems to present as many difficulties as the diffusion approximation.

5.8. LOGNORMAL APPROACH

It was argued earlier that the LogNormal does not satisfy the FPE and is unsuitable as an equilibrium queue distribution because of its relatively heavy tail, which does not tend towards exponential. However, it does have two desirable properties for approximating the *dynamic* component, namely being zero at *x*=0 (if unshifted), so naturally accommodating the reflecting barrier, and approximating Normal when the standard deviation is much less than the mean. If the queue size distribution can be divided into separate equilibrium and dynamic components then an explicit solution may be arrived at. Since the unshifted LogNormal vanishes at *x*=0 the amplitude of the equilibrium component there can be determined. However, this component must be weighted so that its integral is < 1, to leave room for the LogNormal component.

If $p_E(0)$ must equal $p(0)$, its weight must be a function of *x* and must equal 1 at *x*=0. Assuming that p_E is restricted to the exponential (*k*=0), an exponential weighting function (as proposed earlier) satisfies the requirements. Since the LogNormal vanishes at *x*=0, its normalising factor can be a constant. Equation (5.5.1) reduces to:

$$p(x) = e^{-\theta x} p_E(x,k,\nu) + n p_L(x,m,s) \qquad (5.8.1)$$

$$p_E(x,\nu) = -\rho^x \log\rho = \nu e^{-\nu x} \qquad (5.8.2)$$

$$\tilde{P}_E = \int_0^\infty e^{-\theta x} p_E(x) = \frac{\nu}{\nu+\theta} \qquad (5.8.3)$$

$$\tilde{L}_E = \int_0^\infty x e^{-\theta x} p_E(x) = \frac{\nu}{(\nu+\theta)^2} \qquad (5.8.4)$$

$$\tilde{V}_E + \tilde{L}_E^2 = \int_0^\infty x^2 e^{-\theta x} p_E(x) = \frac{2\nu}{(\nu+\theta)^3} \qquad (5.8.5)$$

Since the LogNormal is defined only over [0,∞), its contribution to the mean and variance can be calculated directly, and the constant normalising factor *n* is calculable directly from the integral of (5.8.1) which is fixed at 1:

$$n = 1 - \frac{\nu}{\nu+\theta} = \frac{\theta}{\nu+\theta} \qquad (5.8.6)$$

The LogNormal is specified with two parameters (with location parameter = 0):

$$p_L(x,m,s)=\frac{e^{-\left(\frac{\ln(x/m)^2}{2s^2}\right)}}{xs\sqrt{2\pi}} \tag{5.8.7}$$

Its mean has to equal:

$$\tilde{L}_L=\int_0^{\infty} xp_L(x)=\frac{1}{n}\left(L-\frac{\nu}{(\nu+\theta)^2}\right)=\frac{(\nu+\theta)}{\theta}\left(L-\frac{\nu}{(\nu+\theta)^2}\right)=me^{\frac{1}{2}s^2} \tag{5.8.8}$$

and its variance must satisfy:

$$\begin{aligned}\tilde{V}_L+\tilde{L}_L^2&=\int_0^{\infty} x^2p_L(x)=\frac{(\nu+\theta)}{\theta}\left(V+L^2-\frac{2\nu}{(\nu+\theta)^3}\right)\\&=m^2\left(e^{s^2}\left(e^{s^2}-1\right)+e^{s^2}\right)=m^2e^{2s^2}\end{aligned} \tag{5.8.9}$$

leading to the explicit solution:

$$m=\frac{\tilde{L}_L}{\sqrt{\frac{\tilde{V}_L}{\tilde{L}_L^2}+1}}\qquad\qquad s=\sqrt{\ln\left(\frac{\tilde{V}_L}{\tilde{L}_L^2}+1\right)} \tag{5.8.10}$$

Equations (5.8.9-10) together allow *m* and *s* to be expressed in terms of modelled moments:

$$m=\frac{\left(L(\nu+\theta)^2-\nu\right)^2}{(\nu+\theta)\theta^{\frac{3}{2}}\sqrt{\left(\left(V+L^2\right)(\nu+\theta)^3-2\nu\right)}} \tag{5.8.11}$$

$$e^{s^2}=\theta\frac{\left(\left(V+L^2\right)(\nu+\theta)^3-2\nu\right)}{\left(L(\nu+\theta)^2-\nu\right)^2} \tag{5.8.12}$$

If θ=0, *m* and *s* become essentially undefined, and *n*=0 from (5.8.6). If ν=0, the exponential distribution component vanishes, and (5.8.11-12) reduce to $\tilde{L}=L$, $\tilde{V}=V$. Once θ is chosen, the distributions are explicitly determined. The simulated distributions of a relatively heavy peak case, J3P9, have been fitted by selecting the values of θ that minimise the sum-of-squares error between the distributions, where the *i*th discrete term is compared with the continuum probability at *x* = *i*+0.5. The results given in Figure 5.8.1 show that θ never rises much above 0.1 and the error is moderate in all time slices.

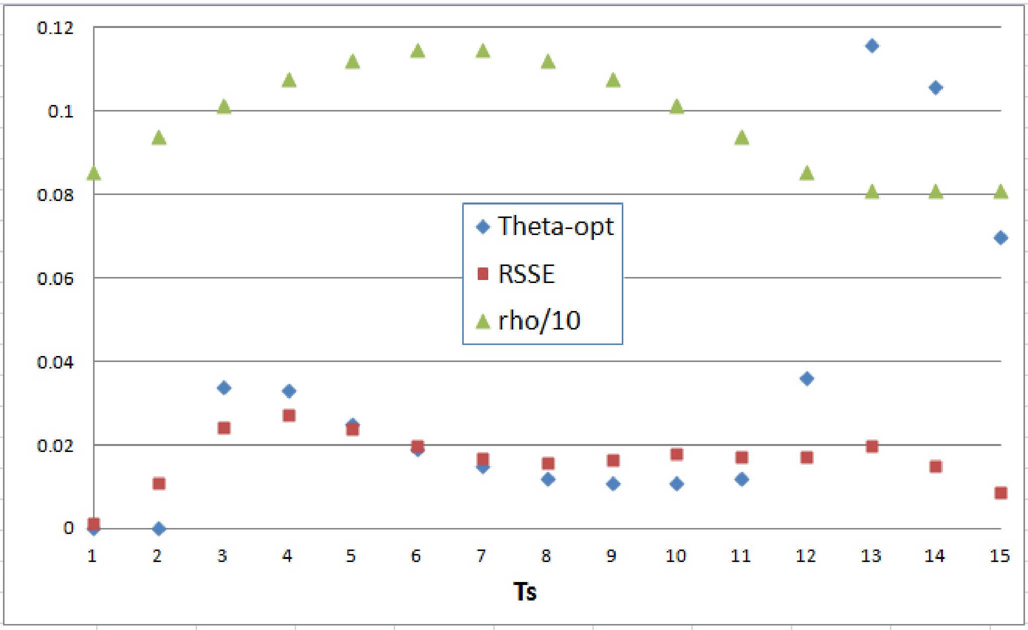
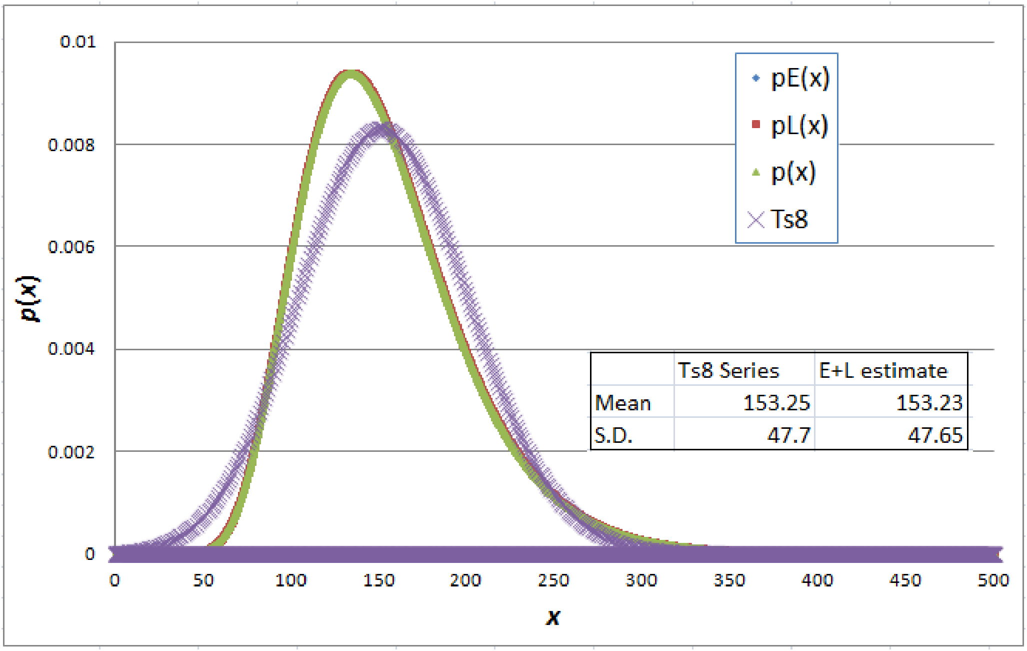


Figure 5.8.1 Exponential and LogNormal distribution fit results for J3P9

The LogNormal function is convenient and easy to work with. Sadly, added to its failure to satify FPE as noted earlier, the match between shapes of estimated and simulated distribution is poor. Figure 5.8.2 gives the example of Ts 8, just after the peak, where the inset shows that while the mean and s.d. agree closely, the distributions visibly do *not* match, because the simulated distribution is close to Normal while the LogNormal is substantially skewed. This situation is inevitable where the real distribution is dominated by the dynamic component.



	Ts8 Series	E+L estimate
Mean	153.25	153.23
S.D.	47.7	47.65

Figure 5.8.2 Simulated and exponential+LogNormal distributions for J3P9 post-peak

5.9. EXPONENTIALLY-WEIGHTED NORMAL APPROACH

5.9.1 Formulation and its consequences

A weighting factor linear in *x* or *t* cannot cover an infinite range. The simplest usable weighting function running from 1 to 0 over the range [0,∞) is an exponential function of *x*, as already applied in earlier Sections, whose parameter θ could be a function of *t*. If a complementary factor is used for the dynamic component, this ensures that only the equilibrating component is non-zero at *x*=0. In general, the parameters in the weighting factors could differ, though in practice they will be equated to reduce the degrees of freedom. Even then the problem is overspecified because there are four parameters to be fitted to three moments. The combined extended probability distribution can be written:

$$p(x) = e^{-\theta x} p_E(x,k,\nu) + n\left(1 - e^{-\phi x}\right) p_D(x,m,s) \qquad (5.9.1)$$

Now assuming the dynamic part to be Normal, exponential weighting acts to shift and scale the distribution, leaving its Normal form unchanged:

$$\tilde{p}_N(x,\phi,m,s) \equiv e^{-\phi\left(m-\frac{\phi}{2}s^2\right)} p_N\left(x, m-\phi s^2, s\right) \equiv B(\phi,m,s) p_N\left(x, m-\phi s^2, s\right) \qquad (5.9.2)$$

Equation (5.9.1) is thus transformed into a linear combination of three components, such that the last two cancel at *x*=0, leaving the weighted exponential $p_E(0)$ to represent (most of) p_0.

$$p(x) = A p_E\left(x, k, \nu + \frac{\theta}{k}\right) + n\left[p_N(x,m,s) - B p_N\left(x, m-\phi s^2, s\right)\right] \qquad (5.9.3)$$

This has a similarity to the form of the diffusion solution (5.3.5), and the second Normal term can be considered to represent a reflected component.

Compared to the linear combination, the contributions of the equilibrating component to the mean and variance of the whole distribution are no longer simply factored by a constant, because the exponential weighting factor enters into the integrals when calculating the first and second moments. Numerical evaluation of the estimated distribution should present no fundamental difficulties since exponential, Gamma, cumulative Gamma and error (erf/c) function approximations are available in most programming languages.

Moments require integration over [0,∞). As the exponential or Gamma equilibrium components are not truncated, their standard mean and variance can be used, except when calculating p_0 (unless the range is short enough that the integral can be approximated using the median value of the function). The Normal functions do need to be truncated, so error functions will be involved, and it is no longer possible to get closed-form solutions for the Normal parameters as it was for the LogNormal approximation.

5.9.2 Formulae for weighted exponential distribution as equilibrium component

Accepting the exponential distribution as a quasi-static approximation to part of a time-dependent distribution, the first step is to equate the first (or only) traffic parameter ρ or ρ^* with the utilisation, to give the parameter of the continuous analogue:

$$\nu = -\ln \rho^* \equiv -\ln u = -\ln\left(1 - \overline{p}_0\right) \qquad (5.9.4)$$

Starting again with the exponential distribution (assuming M/M/1 for simplicity):

$$p_E(x) = -\rho^{*x} \ln \rho^* = \nu e^{-\nu x} \qquad (5.9.5)$$

$$\tilde{P}_E = \int_0^\infty e^{-\theta x} p_E(x) = \frac{\nu}{\nu + \theta} \qquad (5.9.6)$$

$$\tilde{L}_E = \int_0^\infty x e^{-\theta x} p_E(x) = \frac{\nu}{(\nu + \theta)^2} \qquad (5.9.7)$$

$$\tilde{V}_E + \tilde{L}_E^2 = \int_0^\infty x^2 e^{-\theta x} p_E(x) = \frac{2\nu}{(\nu + \theta)^3} \qquad (5.9.8)$$

For a doubly-nested distribution the correction (5.6.10-11) can be applied to the moments provided that ν is replaced by $\vec{\nu}$ calculated from the asymptotic geometric parameter $\vec{\rho}$.

5.9.3 Formulae for weighted Normal distribution as dynamic component

As the Normal function is truncated, its integrals involve error functions:

$$\int_0^y p_N(x,m,s)dx = \frac{1}{2}\left[erf\left(\frac{y-m}{s\sqrt{2}}\right)+erf\left(\frac{m}{s\sqrt{2}}\right)\right] \qquad (5.9.9)$$

$$\int_0^y xp_N(x,m,s)dx = \frac{m}{2}\left[erf\left(\frac{y-m}{s\sqrt{2}}\right)+erf\left(\frac{m}{s\sqrt{2}}\right)\right]+\frac{s}{\sqrt{2\pi}}\left(e^{-\frac{1}{2}\left(\frac{m}{s}\right)^2}-e^{-\frac{1}{2}\left(\frac{y-m}{s}\right)^2}\right) \qquad (5.9.10)$$

$$\int_0^y x^2 p_N(x,m,s)dx = \frac{\left(m^2+s^2\right)}{2}\left[erf\left(\frac{y-m}{s\sqrt{2}}\right)+erf\left(\frac{m}{s\sqrt{2}}\right)\right]+\frac{ms}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{m}{s}\right)^2}-\frac{(y+m)s}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y-m}{s}\right)^2}$$

(5.9.11)

Specific moment integrals over [0,∞) are:

$$\int_0^\infty p_N(x,m,s)dx = \frac{1}{s\sqrt{2\pi}}\int_0^\infty e^{-\frac{1}{2}\left(\frac{x-m}{s}\right)^2}dx = \frac{1}{2}erfc\left(-\frac{m}{s\sqrt{2}}\right) \qquad (5.9.12)$$

$$\int_0^\infty xp_N(x,m,s)dx \equiv M(m,s) = \frac{s}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{m}{s}\right)^2}+\frac{m}{2}erfc\left(-\frac{m}{s\sqrt{2}}\right) \qquad (5.9.13)$$

$$\int_0^\infty x^2 p_N(x,m,s)dx \equiv S(m,s) = \frac{ms}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{m}{s}\right)^2}+\frac{\left(s^2+m^2\right)}{2}erfc\left(-\frac{m}{s\sqrt{2}}\right) \qquad (5.9.14)$$

Using the shorthand:

$$N(m,s) \equiv \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{m}{s}\right)^2} \quad E(m,s) \equiv \frac{1}{2}erfc\left(-\frac{m}{s\sqrt{2}}\right) \quad F(\theta,m,s) \equiv e^{-\theta\left(m-\frac{\theta}{2}s^2\right)}$$

(5.9.15)

equations (5.9.12-14) can be rewritten:

$$\int_0^\infty p_N(x,m,s)dx = E(m,s) \qquad (5.9.16)$$

$$\int_0^\infty xp_N(x,m,s)dx \equiv M(m,s) = sN(m,s)+mE(m,s) \qquad (5.9.17)$$

$$\int_0^\infty x^2 p_N(x,m,s)dx \equiv S(m,s) = msN(m,s)+\left(s^2+m^2\right)E(m,s) \qquad (5.9.18)$$

5.9.4 Combining the distributions

Combining the integrated components in accordance with (5.9.1) and (5.9.3), the moments of the full distribution approximation are:

$$\int_0^\infty p(x)dx = 1 = \frac{\nu}{(\nu+\theta)} + n\left[E(m,s) - F(\theta,m,s)E\left(m-\theta s^2, s\right)\right] \tag{5.9.19}$$

$$L = \frac{\nu}{(\nu+\theta)^2} + n\left[M(m,s) - F(\theta,m,s)M\left(m-\theta s^2, s\right)\right] \tag{5.9.20}$$

$$V + L^2 = \frac{2\nu}{(\nu+\theta)^3} + n\left[S(m,s) - F(\theta,m,s)S\left(m-\theta s^2, s\right)\right] \tag{5.9.21}$$

Most of the error function terms in (5.9.20), and thanks to (5.9.3) all the straight exponential terms, can be eliminated to give an alternative expression for *L* from which equation (5.9.19) can be used to eliminate *n* if required:

$$L = \frac{\nu}{(\nu+\theta)^2} + \frac{m\theta}{\nu+\theta} + e^{-\theta\left(m-\frac{\theta}{2}s^2\right)}\left(\frac{n\theta s^2}{2}\right)erfc\left(-\frac{\left(m-\theta s^2\right)}{s\sqrt{2}}\right) \tag{5.9.22}$$

Equation (5.9.22) confirms that the mean tends to 1/ν for small θ, and to *m* for large θ and large *m*, with an additional term that vanishes at both extremes and is linked with the ‘reflected’ Normal component.

5.10. SOLVING EXPONENTIALLY-WEIGHTED DISTRIBUTION COMBINATION

5.10.1 Motivation and approach

The purpose of this Section is to show that solutions exist for time-stamped queue size distributions, that satisfy the constraints of time-dependent moments and are a reasonable match to simulated distributions. It is not proposed to define a solution algorithm since standard numerical methods exist. While these methods may not be computationally efficient enough to be embodied ‘on-line’ in an analytical traffic modelling program, they should be usable ‘off-line’ as a post-processing output. More efficient methods, e.g. to enable the probability of spillback to be incorporated into an assignment algorithm, are left for future research, though ‘heuristic’ estimations based directly on the queue moments might be possible.

5.10.2 Practical problems affecting analytical solution methods

If the normalising factor *n* is eliminated this leaves just three degrees of freedom represented by θ, *m* and *s*. Since ν is fixed by p_0, there are just two moments to be fitted, *L* and *V*. Hence the problem is overspecified unless some other constraint, such as on shape, can be imposed.

In the LogNormal approach, the explicit expressions for *m* and *s* reduce the solution problem to one of choosing θ. As an explicit solution for *m* and *s* appears impossible in this case, even where θ is specified, it is natural to ask first whether a solution can be found by minimising an objective function. Differentials of the elementary functions in (5.9.15) include:

$$dN = ms^{-2}N\left(ms^{-1}ds - dm\right) \qquad dE = s^{-1}N\left(ms^{-1}ds - dm\right)$$
$$dF = \theta^2 sFds - \theta Fdm - \left(m - \theta s^2\right)Fd\theta \tag{5.10.1}$$

where *N* represents the Normal component, $N(x,m,s)$ or $N(m,s)$ with x=0 understood as appropriate. Since ν is assumed to be fixed by p_0, the differential of the integral (5.9.19) is:

$$\begin{aligned}
0 &= -\frac{\nu}{(\nu+\theta)^2}d\theta \\
&+ \left[E(m,s) - F(\theta,m,s)E\left(m-\theta s^2,s\right)\right]dn \\
&+ nN(m,s)\left(ms^{-2}ds - s^{-1}dm\right) \\
&- nF(\theta,m,s)N\left(m-\theta s^2,s\right)\left(\left(ms^{-2}+\theta\right)ds - s^{-1}dm + sd\theta\right) \\
&- n\left(\theta^2 sFds - \theta Fdm - \left(m-\theta s^2\right)Fd\theta\right)E\left(m-\theta s^2,s\right)
\end{aligned} \tag{5.10.2}$$

This expression alone has 12 terms, of which only one can be eliminated by fixing θ, and the equivalents for *dL* and *dV* would be even more complicated. Alternatively the differentials of (5.9.1) can be integrated:

$$\begin{aligned} dp &= x(nN-p)d\theta + n\left(1-e^{-\theta x}\right)N\left[\frac{(x-m)}{s^2}\left(dm+\frac{ds}{s}\right)+\frac{dn}{n}\right] \\ &= x(nN-p)d\theta + \left(p-\nu e^{-(\nu+\theta)x}\right)\left[\frac{(x-m)}{s^2}\left(dm+\frac{ds}{s}\right)+\frac{dn}{n}\right] \end{aligned} \qquad (5.10.3)$$

$$\begin{aligned} 0 &= \left(n\int xN - L\right)d\theta + \left[\frac{(L-m)}{s^2}\left(dm+\frac{ds}{s}\right)+\frac{dn}{n}\right] \\ &-\nu\left(\frac{1}{s^2}\left(dm+\frac{ds}{s}\right)+\frac{dn}{n}\right)\int(x-m)e^{-(\nu+\theta)x} \\ &= \left(nM(m,s)-L\right)d\theta + \left[\frac{(L-m)}{s^2}\left(dm+\frac{ds}{s}\right)+\frac{dn}{n}\right] \\ &-\nu\frac{(m+\nu+\theta)}{(\nu+\theta)^2}\left(\frac{1}{s^2}\left(dm+\frac{ds}{s}\right)+\frac{dn}{n}\right) \end{aligned} \qquad (5.10.4)$$

This is somewhat more manageable, as would be $\int xdp$ and $\int x^2 dp$ from which *dL* and *dV* can be obtained. However, it is still a fairly awkward expression, encouraging the use of a standard numerical solution method for the purpose of practical demonstration.

5.10.3 Selecting initial parameter values and target criteria

Numerical methods generally require an initial approximate solution, the choice of which may affect the final result. The simplest starting parameters that would not cause an instant error are θ=0, *m*=0, *s*=1. This would produce an initial exponential distribution. Following some experimentation, it has been found that except in equilibrated (or nearly) pre-peak time slices, where the exponential is a natural initial choice, a useful initial estimate of θ is given by the following, obtained simply from (5.9.19) by ignoring the contribution of the second Normal component, assuming *n*=1, and setting initial *m* and *s* as if the distribution were Normal:

$$m = \tilde{L}, \qquad s = \sqrt{V} \qquad (5.10.5)$$

$$\theta \approx \nu\left(\frac{2}{erfc\left(-\frac{\tilde{L}}{\sqrt{2V}}\right)}-1\right) \qquad (5.10.6)$$

$\tilde{L}$ is the target mean queue adjusted according to equation (5.6.4), and v is calculated from p_0 according to equation (5.9.4). The possibility of solution can be tested by using a numerical tool such Solver available with Microsoft Excel, whose default solution method is Newton.

The error in the estimated distribution (‘*dis*’) is not naturally normalised, but the sum of squared differences (5.10.7) will converge and become nearly constant for sufficiently large *N*.

$$Error(p) = \sqrt{\sum_{0}^{N} \left(p_{i(sim)} - p_{i(dis)}\right)^2} \qquad (5.10.7)$$

However, minimising this error cannot be an objective because a simulated distribution will not normally be available. Therefore the target for solution must be based on fitting moments of the distribution. However, since there are two moments that in general cannot be fitted simultaneously, so they must be combined in some way. A simple choice is to combine the errors in the mean and standard deviation unweighted:

$$Error(L,S) = \sqrt{\left(\tilde{L}_{(tar)} - L_{(dis)}\right)^2 + \left(S_{(tar)} - S_{(dis)}\right)^2} \qquad (5.10.8)$$

Alignment of the discrete and continuous distribution was discussed earlier in Section 5.6. In these tests the discrete probability values taken from the simulated distributions are placed at half-interval points, i.e. p_0 at x=0.5, p_1 at x=1.5 etc, amounting to assumption of a constant displacement h=0.5. As pointed out earlier, this is quite accurate for higher values of ρ, and not too far off for lower values. The discrete mean is thus automatically adjusted before being compared with the continuous mean (variances being unaffected). When fitting a distribution to a *calculated* mean queue size, the latter needs to be adjusted explicitly, the earlier argument again justifying the simple addition of 0.5.

5.10.4 Early unsuccessful attempts at efficient solution

In principle an explicit solution method tailored to the problem would be preferable to a proprietary ‘black box’, or numerical calculation of uncertain duration. However, an attempt to create an *ad hoc* solution for the parameters by extrapolating the results of varying them by small amounts, e.g. ±10%, proved unreliable. Improving upon standard but not necessarily efficient, e.g. Newton, solution methods would be likely to be a major piece of work in itself and a digression from the present purpose. Computationally efficient implementation of the estimation of probability distributions, as such, is not an aim of this work, so for the purposes of demonstration a method available with proprietary software has been used.

5.10.5 Solution results using Excel Solver

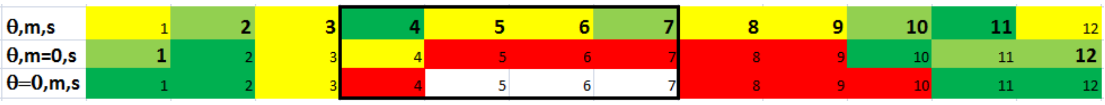
Two peak cases have been used in tests, J2P4 and J3P9, the latter being the more demanding because it represents one of the heaviest peaks. Steps of estimation are given in Table 5.10.1, Figure 5.10.1 plotting values of θ and errors for 15 time slices of J3P9, each 10 minutes long.

Table 5.10.1 Excel Solver estimation schemes for distribution parameters

Setup	Define criterion, e.g. mimimise error in L, $S=\sqrt{V}$ as defined by (5.10.8)	
1	Set initial θ, m, s	Solve for θ,m, s
2	Force m=0, set initial θ,s	Solve for θ,s
3	If ρ<1, Force θ=0, set initial m, s	Solve for m, s
Output	Select the result that minimises the error criterion	

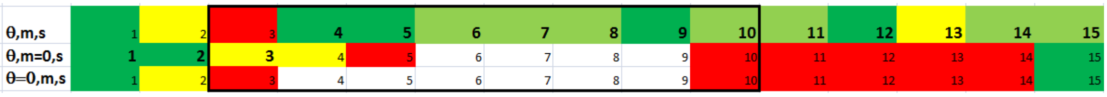
The initial parameter estimates (5.10.5-6) are used in each step - there is no recursion. Forcing a parameter to zero means that it is not only initialised to zero but is held to zero in the solution. The logic of Step 3 is that an equilibrium distribution cannot occur if ρ≥1, but Step 2 is also aimed at ρ<1 since a dynamic distribution is likely to have a significant Normal component with m>0. It is found in practice that Step 3 does not improve the solution, so can be omitted.

In testing the method, there is an issue concerning what is being compared with what. Initial test sought to establish that it could reproduce the queue size probability distributions given their actual moments $\{p_0,L,V\}$. In the 'heat-maps', Figures 5.10.1-2, the minimum error solutions are indicated by bold time-slice numbers. The subjective fit to the distribution is indicated by a colour code: full green is good, light green acceptable, yellow indicates a defect though not necessarily serious, red is unacceptable, and blank means not tested.



θ,m,s	🟨 1	🟢 **2**	🟨 **3**	🟩 **4**	🟨 **5**	🟨 **6**	🟢 **7**	🟨 **8**	🟨 **9**	🟢 **10**	🟩 **11**	🟨 12
θ,m=0,s	🟢 **1**	🟩 2	🟨 3	🟨 4	🟥 5	🟥 6	🟥 7	🟥 8	🟥 9	🟩 10	🟢 11	🟢 **12**
θ=0,m,s	🟩 1	🟩 2	🟨 3	🟥 4	5	6	7	🟥 8	🟥 9	🟥 10	🟩 11	🟩 12

Figure 5.10.1 Verification of distribution fit by time-slice for J2P4 moderate peak



θ,m,s	🟩 1	🟨 2	🟥 3	🟩 **4**	🟩 **5**	🟢 **6**	🟢 **7**	🟢 **8**	🟩 **9**	🟢 **10**	🟢 **11**	🟩 **12**	🟨 **13**	🟢 **14**	🟢 **15**
θ,m=0,s	🟩 **1**	🟩 **2**	🟨 **3**	🟨 4	🟥 5	6	7	8	9	🟥 10	🟥 11	🟥 12	🟥 13	🟥 14	🟩 15
θ=0,m,s	🟩 1	🟨 2	🟥 3	4	5	6	7	8	9	🟥 10	🟥 11	🟥 12	🟥 13	🟥 14	🟩 15

Figure 5.10.2 Verification of distribution fit by time-slice for J3P9 heavy peak

Significant results are that:

- No pure equilibrium solution with θ=0 is optimal, though the difference can be small;
- Modified equilibrium solutions with m=0 may be optimal outside the peak;
- In most post-peak periods all three parameters are optimally non-zero.

Figures 5.10.3-4 plot errors for the same two peak cases as Figures 5.10.1-2. Errors in the distribution (equation 5.10.7) are absolute since the distribution is normalised, and are modest, the greatest being a little over 4%. Errors in the moments according to the criterion equation (5.10.8) are shown relative to the mean queues and are very small.

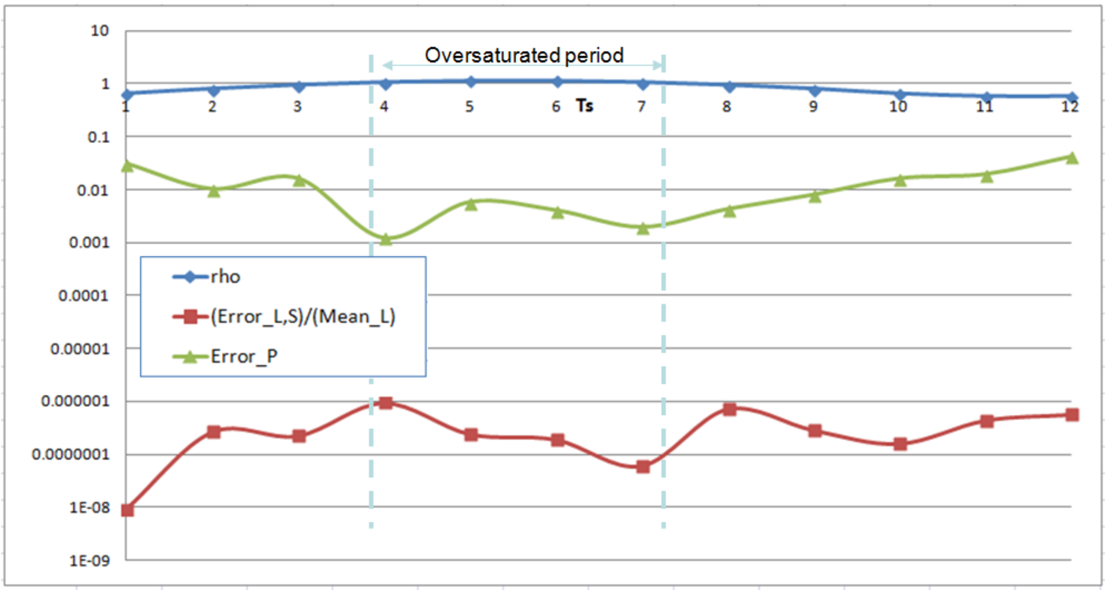


Figure 5.10.3 Intrinsic distribution fit errors by time-slice for J2P4 moderate peak

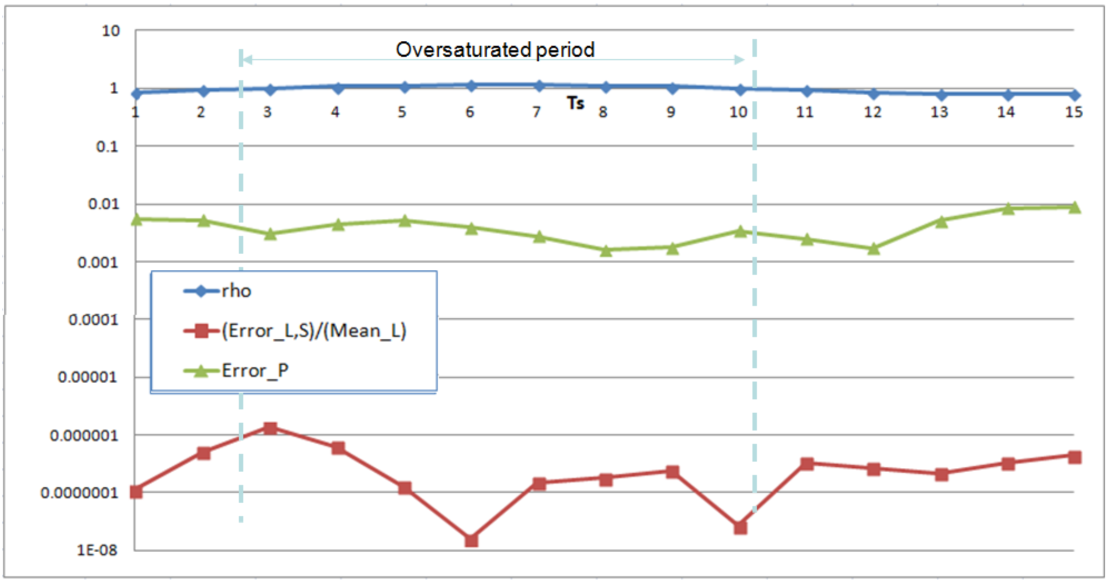


Figure 5.10.4 Intrinsic distribution fit errors by time-slice for J3P9 heavy peak

Figures 5.10.5-6 compare estimated and simulated probability distributions for four time slices in each of the peak cases, where it can be seen that the fits are close. In these plots the green and red curves represent the exponential and Normal components respectively. The resultant estimated distribution in blue and the simulated distribution is purple and marked by crosses. These results give confidence in the method as such.

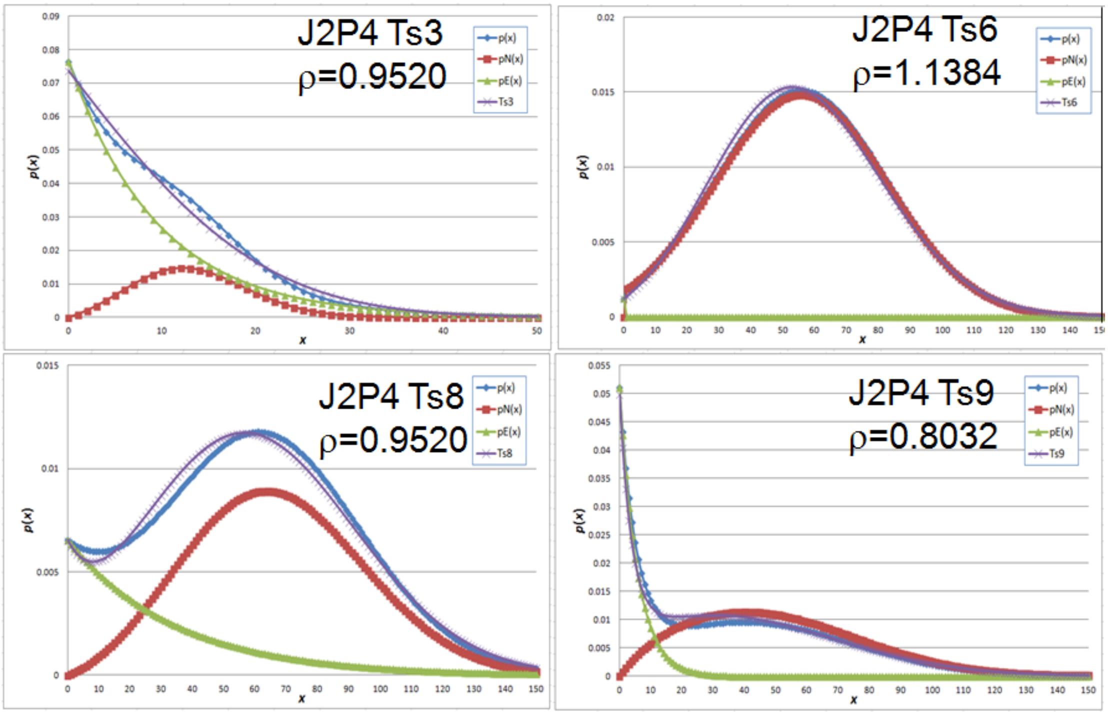


Figure 5.10.5 Intrinsic distribution fits in four time slices of J2P4 moderate peak

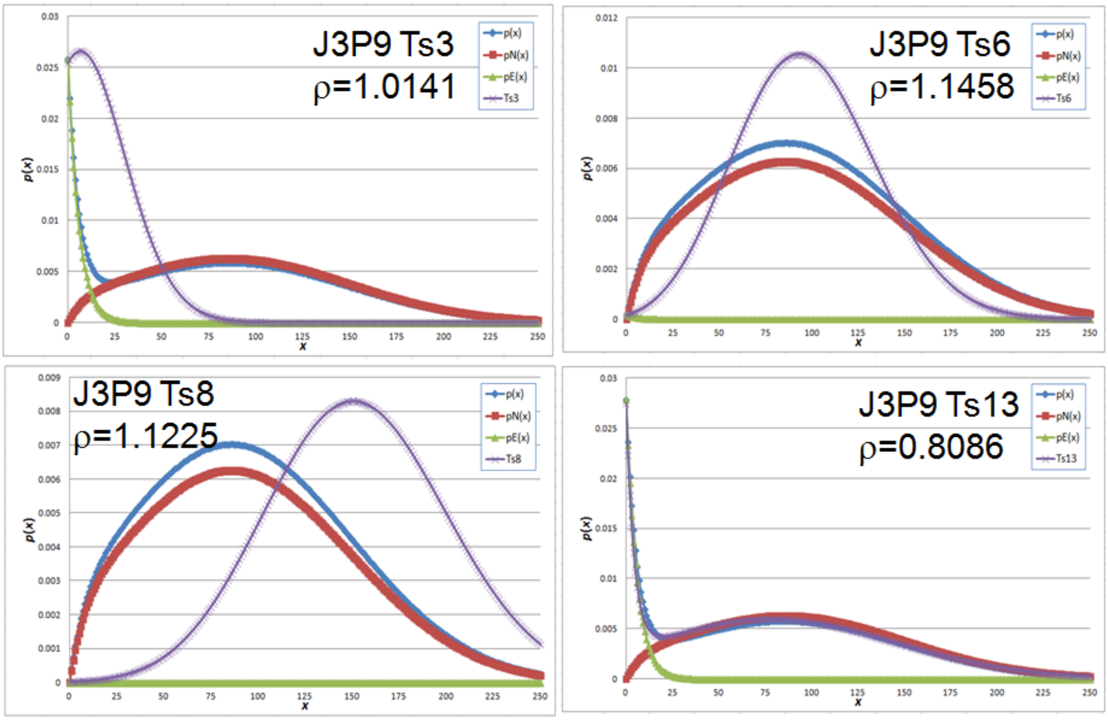


Figure 5.10.6 Intrinsic distribution fits in four time slices of J3P9 heavy peak

However, in practice the simulated distributions will not be available, and the estimated queue moments will be developed over many time periods allowing errors to creep in. Therefore to compare estimated with simulated distributions time-slice by time-slice is not entirely fair, but is nevertheless necessary to establish usefulness of the method. Figures 5.10.7-8 plot errors in estimated distributions on this basis.

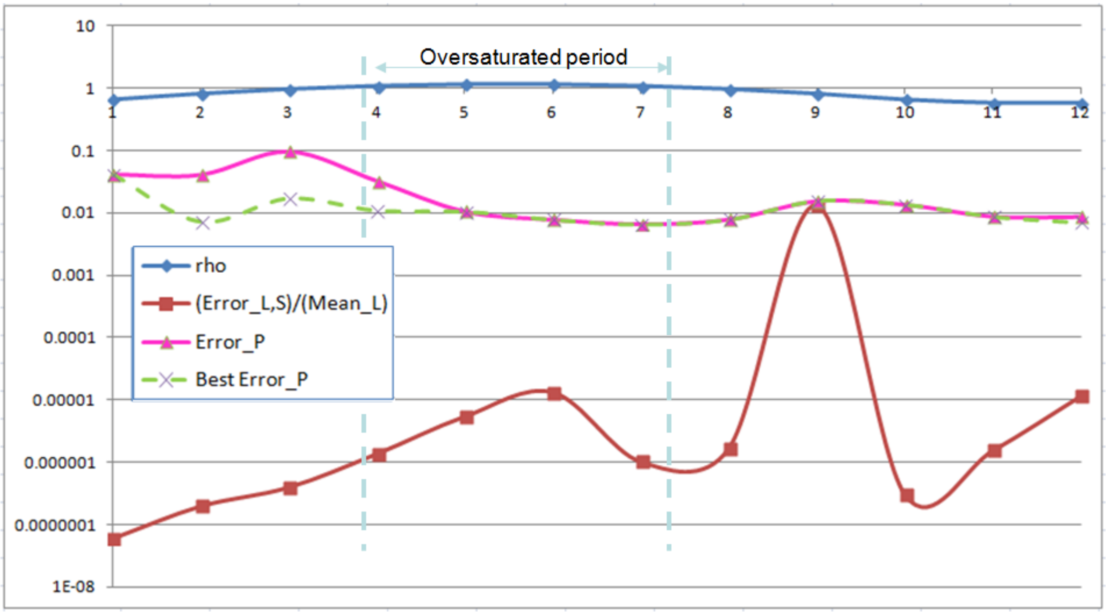


Figure 5.10.7 Distribution fit errors through time-slices for J2P4 moderate peak

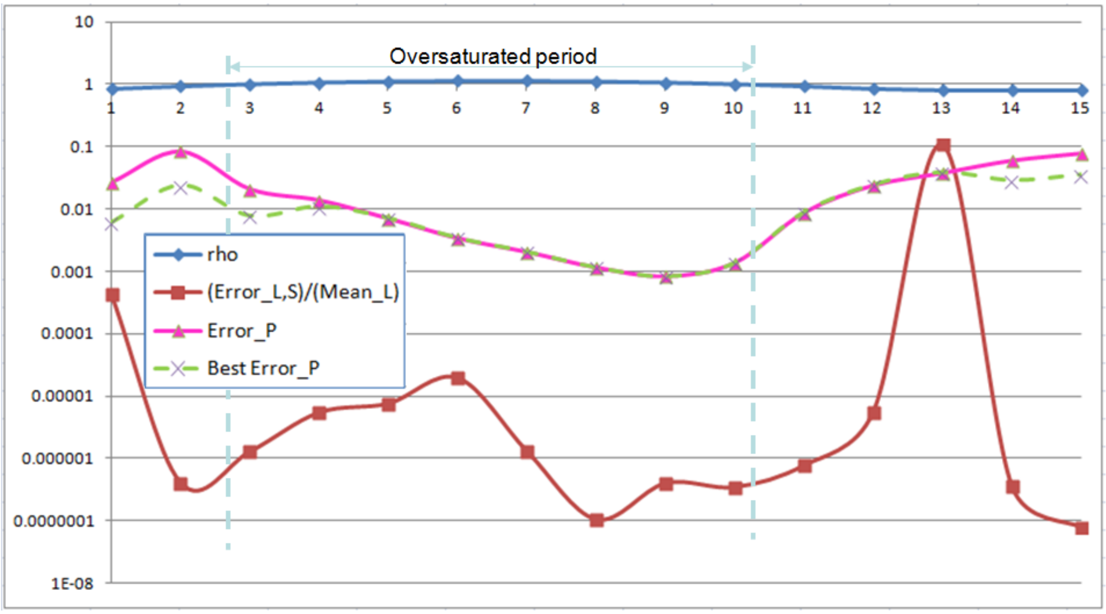


Figure 5.10.8 Distribution fit errors through time-slices for J3P9 heavy peak

Errors in the distribution itself are highly variable but only in isolated cases reach 10%. The larger errors reflect difficulty in reproducing simulated distributions either because of their shape or because p_0 is not accurately estimated, bearing in mind that its estimated value is highly sensitive to estimated utilisation near to 1. In some time-slices, the fit to the simulated distribution could be improved (broken lines) but only at the expense of worsening the fit to the estimated moments. Results are necessarily a compromise, but Figures 5.10.9-10 show that the distribution shapes are in most cases similar, a notable exception being in Ts 3 of case J2P4, which lies in the pre-peak transition period just before oversaturation.

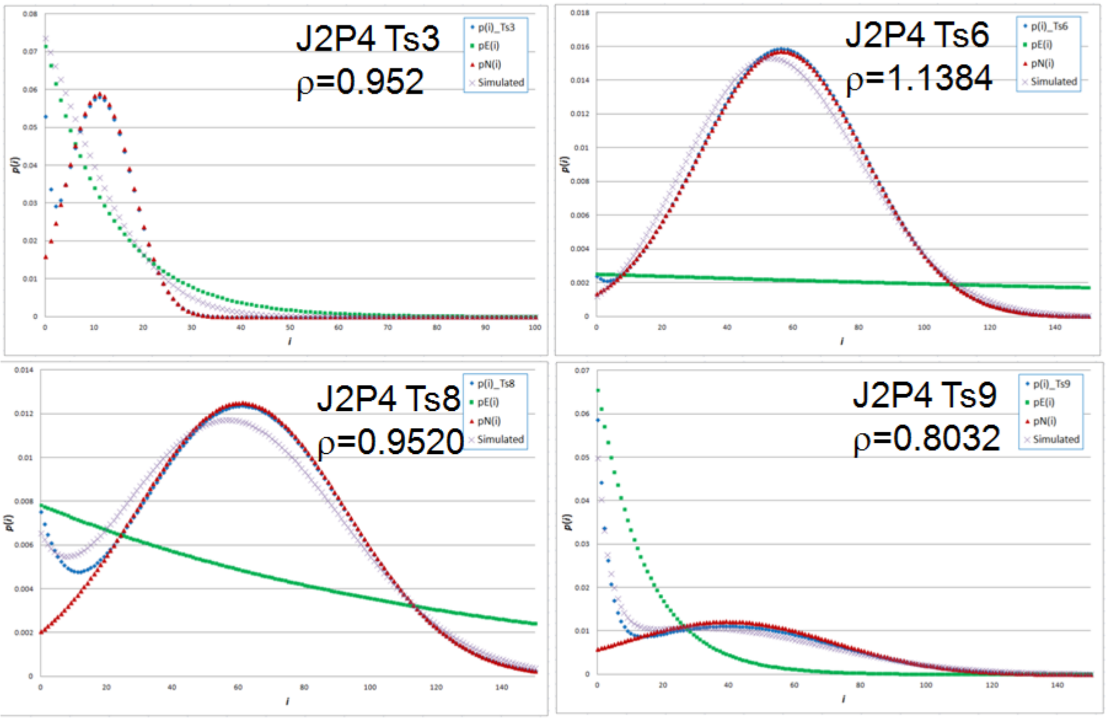


Figure 5.10.9 Example distribution fits in four time slices of J2P4 moderate peak

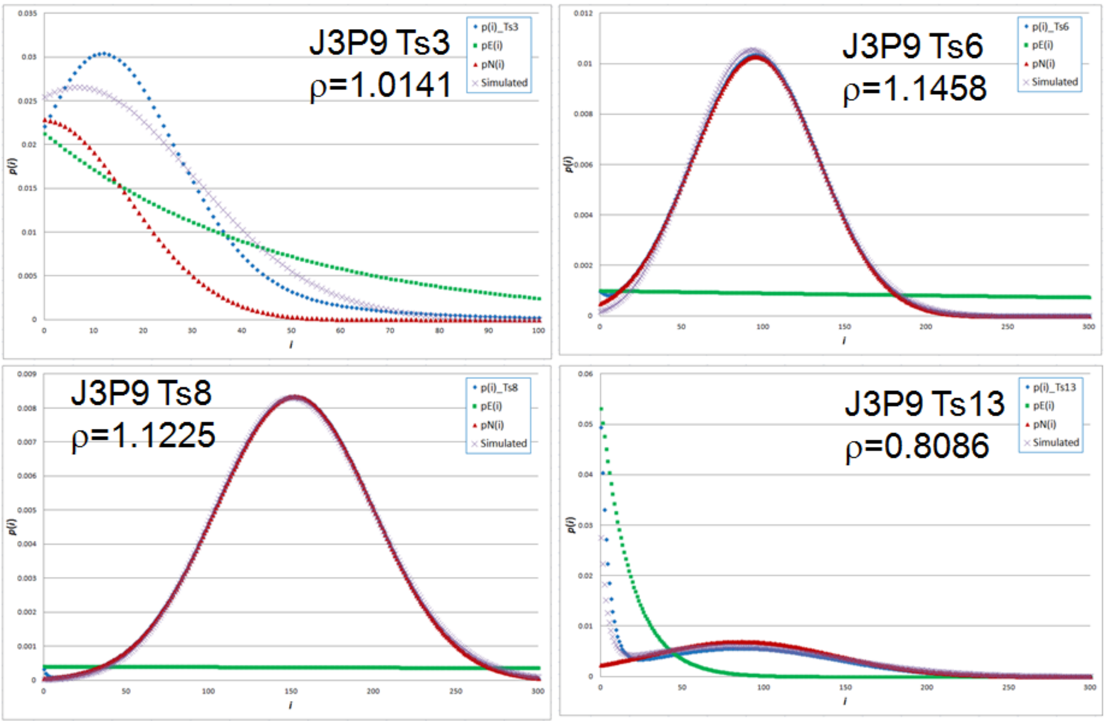


Figure 5.10.10 Example distribution fits in four time slices of J3P9 heavy peak

Table 5.10.2 gives some parameters of the peak cases to indicate scale.

Table 5.10.2 Throughput parameters for the peak cases tested

Case	**Average capacity**	**Time slice duration**	**Ave. throughput capacity in Ts**	**Max. traffic intensity**	**Duration of oversaturation**
J2P4	14.7 units/min	9 minutes	132	1.1384	36 minutes
J3P9	22.5 units/min	10 minutes	225	1.1458	80 minutes

While the Normal distribution makes a good approximation around the height of a peak in these oversaturated cases, it is evident that both components are needed post-peak once demand falls below saturation. As indicated by Figures 5.10.1-2, no distribution is adequately fitted by a purely exponential distribution - there are no optimal solutions with θ exactly zero. This result is less surprising in hindsight, given that the queuing processes are dynamic. Values of *m* tend to be similar to the mean queues unless θ is very small, while values of *s* vary quite considerably. Solver would be expected to find a small value of θ where the moments are near to equilibrium. This is consistent with the view that Step 3 in Table 5.10.1 can be omitted.

Figures 5.10.8-9 show that problems occur mostly in the pre-peak growth period and the period of maximum post-peak decay rate where it is most difficult to estimate the moments reliably. Since the queue estimation procedure gives quite accurate values of mean, and reasonably accurate values of standard deviation, some of the difference between can be ascribed to a difference between the simulated and calculated values of p_0, which is critical for ‘pinning’ the left-hand end of the distribution. This is an aspect of the macroscopic time-dependent queue approximation which might be addressed in further research.

From the viewpoint of risk and resilience an important practical question is how accurately the tails of queue size distributions can be estimated, though this depends on where the critical size is placed. These questions could be addressed in further research. The results given suggest that the error in estimation of tail probabilities, assuming these to begin some way above the distribution mode, is likely to be small except in a few cases and then mostly pre-peak, while concern is likely to be focused post-peak when extended distributions are most likely.

5.11. CONCLUSIONS ON ESTIMATING PROBABILITY DISTRIBUTIONS

Probability distributions of queue size or delay can be important both for estimating the reliability of travel times and for predicting the likelihood of blocking effects on facilities upstream of a bottleneck. Then the shape of the distribution is important for estimating not only the weight of the tail beyond a specified point, but also its sensitivity to changes in parameters, which is maximised near where the slope of the distribution is greatest. If this point coincides with a critical network location or facility then both large and unpredictable variations in system performance could result. Knowing mean and variance alone is insufficient where probability distributions are far from an equilibrium shape, as the foregoing shows is probable in peak cases. However, distributions can be estimated if utilisation or the probability of the queue beng zero is also known.

This Chapter 5 has discussed the nature of time-dependent queue size distributions and how these can be represented by time-stamped[53] continuous functions with equilibrium or diffusion characteristics. Standard diffusion solutions, while informative, are considered too difficult to work with for present purposes, that seek only to fit distributions statically to analytically derived time-dependent moments, rather than to obtain fully evolving time-dependent distribution functions. Having considered various methods, it is concluded that an exponentially-weighted combination of an equilibrium distribution, most simply exponential but possibly Gamma or doubly-nested geometric/exponential, with a Normal distribution, can give sufficient flexibility to match distributions generated by simulation. The problem is underspecified, and satisfactory results have been obtained in tests using a generally available standard solution method and a simple objective based on the target moments.

Development of equilibrium and time-dependent deterministic variance formulae, correction or substitution of time-dependent mean queue formulae, enlargement of the range of queue processes to which these can be applied, and development of approximations to equilibrium and dynamic probability distributions, complete the main aims of this work. Computationally efficient implementations require an efficient solution algorithm, but specifying this is not an objective since standard methods exist. Implementation for demonstration purposes and computational issues are discussed later in Chapter 7.

[53] 'Time-stamped' signifies that the probability distribution is evaluated for a particular point in time during the evolution of a queue, from time-dependent moments, but is not explicitly a time-dependent function.

CHAPTER 6: QUEUING ON MULTIPLE LANES

6.1. INTRODUCTION

Queues on several interacting lanes, and the particular case of two lanes with turning movements, are explored. Apart from extending the range of cases accessible to time-dependent approximations, and potentially leading to enhanced junction modelling methods, this may give insight into some features of queuing processes. Efficient approximation methods are sought for cases complicated by the interaction of several traffic streams, situations not covered in the foregoing Chapters. Multi-lane queues raise issues about how the statistics of arrivals and departures should be represented, because from the myopic viewpoint of a particular lane, arrivals may appear to be 'censored' if they select another lane either randomly or when the queue there appears to be shorter. If lanes are assumed to share the capacity of a common service channel, effectively their service processes become correlated. The extent to which multi-lane processes can be accommodated by the Pollaczek-Khinchin model is considered, and properties of lane probability distributions are investigated. A paper reporting some of this work was presented at the UTSG 2011 conference (Taylor 2011).

6.2. MULTI-LANE FORMULATION AND ANALYSIS OF TWO LANES

Many junctions, particularly larger ones including major roundabouts, have several entry lanes. Although some lanes are dedicated to particular turning movements, others may be shared by different movements, and movements may have a choice of lanes. Short flares may be present which interact with main approach lanes, but only main lanes of unrestricted length are considered in this idealised analysis. If choice of lane is available, it may be random or involve rational selection, for example choosing the lane that appears to contain the shortest queue. The question is then how this affects the queue sizes. Although M/M/*c* and G/G/*c* multi-server queues are covered by standard works, they tend to deal with short queues on a large number of non-interacting channels, whereas the type of queues of interest here are likely to be longer on fewer channels, and lane-changing either within or on departure from the queue may occur.

6.2.1 Apparent problem with queuing on two lanes

For a single lane where the traffic intensity is ρ the M/M/1 equilibrium queue size is:

$$L_e = \frac{\rho}{1-\rho} \tag{6.2.1}$$

Suppose a junction approach is divided into two lanes, and arrivals can choose a lane either randomly or according to some rule, and there is sharing of service, as in Figure 6.2.1.

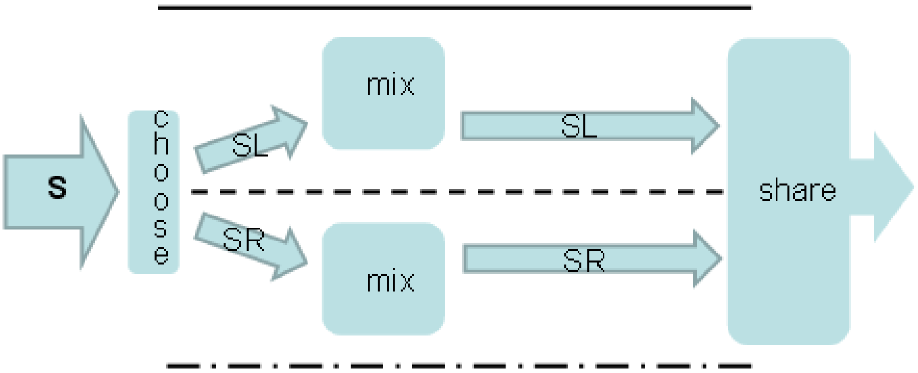


Figure 6.2.1 Two lane approach with possibility of lane choice and service sharing

If there is no lane-changing after arrival and no sharing of capacity, then the lane queues are independent apart from the initial choice of lane. In the steady state, queue size depends only on the demand intensity, not the arriving volume or capacity separately. If each lane has half the total capacity and arrivals are divided with rates $\gamma\lambda$ left and $(1-\gamma)\lambda$ right, and the process in *each* lane is M/M/1, then the total equilibrium queue is:

$$L_e(\gamma) = \frac{2\gamma\rho}{1-2\gamma\rho} + \frac{2(1-\gamma)\rho}{1-2(1-\gamma)\rho} = \frac{2\rho(1-2\gamma\rho(1-\gamma))}{(1-2\gamma\rho)(1-2(1-\gamma)\rho)} \qquad (6.2.2)$$

Because the capacity of each lane is fixed, total queue size is maximised when γ is 0 or 1, and a lane can become oversaturated even if $\rho<1$. Total queue size is minimised when $\gamma=0.5$, but is then twice what it would be in a single lane, namely:

$$L_e = \frac{2\rho}{1-\rho} \qquad \text{(separate and symmetrical lanes)} \qquad (6.2.3)$$

This does not seem realistic. If either lane can be chosen freely, arrivals might choose the lane they perceive to have the shorter queue[54], but this does not seem to be essential since any symmetrical strategy, from choosing lanes alternately to choosing a lane randomly, should equalise the mean lane queues. If the centre line is removed, allowing free movement between lanes, it can be replaced in simulation by free choice between lane servers, if this is possible. Lane-changing as such is only of interest in as much as it may affect throughput capacity.

[54] An element of mis-perception can be introduced by adding to the actual queue size a random amount proportional to the square-root of queue size. This effectively treats the space occupied by customers as a Poisson variable. In practice, it is the split ratio which matters, not the method used to achieve it.

If these various processes are imagined to be hidden inside a ‘black box’, they become equivalent to a single lane (6.2.1). By symmetry, each lane queue must then be half the total. Yet the demand intensity on each lane is still ρ, since both the arrivals (assumed symmetrical) and the nominal capacity in each lane are halved. Therefore the statistics of the queue process in either lane can no longer be M/M/1. If lane choice by arrivals is (relatively) unimportant, as implied above, then it is the service process that is critical.

If two lanes ultimately feed into a single stream, for example where a multi-lane stop line at a roundabout feeds into an unmarked circulating section, some turbulence in the merge might be expected. In practice such movements may be difficult to observe in detail, making it difficult to verify a description except in aggregate. Nevertheless, one can imagine a range of possibilities from no sharing of service to perfect sharing, with any actual case approximated somewhere in that range. This is of some interest for its own sake, and is approached here in those terms, but there is also the possibility of shedding light on non-M/M/1 processes and on how they can be approximated.

6.2.2 Simulation of multi-lane queues and their probability distributions

A simple microscopic simulation program has been constructed, where arrival and service events have exponential headway distributions generated using a standard pseudo-random generator[55], with sufficient events to allow equilibrium to be approached. Options include:

- Lane selected randomly on arrival and for shared service
- Shortest lane queue selected on arrival and lane selected randomly for shared service
- Independent arrivals and service on each lane with same total traffic.

Queue sizes are given in Table 6.2.1 for a four-lane case, with ρ=0.9, capacity μ=1 for shared service or 0.25 for independent service where each lane receives 1/4 of the flow, and 1 million simulated events. The difference between the simulated and expected mean queues, 8.838 and 9.0 respectively, is consistent with an analysis of ‘standard’ error given in Appendix F[56]. The shared service mechanism in the first two cases selects a lane at random from those *where a queue is present*. Therefore the process is not truly FIFO, since queuing ‘customers’ are not represented explicitly. That the effect of this is small has been confirmed by a small difference in results using an alternative simulation that maintains FIFO in each lane. However, differences in the software and its efficiency *are* significant because individuals are identified. As is to be expected, there is substantial correlation between lane queues with shared service.

[55] Random number generator supplied with the Silverfrost Fortran 95 compiler.
[56] I am grateful to one of the Examiners for pointing out a need to increase the number of events simulated.

Table 6.2.1 Queue sizes in multi-lane simulations (ρ=0.9, μ=1.0, 1M events)

No. of	**Capacity μ and**	**Total**	**Individual Lane Queues**				**Correlation of**
Lanes	**Type of process**	**Queue**	**Lane 1**	**Lane 2**	**Lane 3**	**Lane 4**	**Queue Sizes(r)**
1	1.00 random	8.838	8.838	-	-	-	n/a
4	0.25 indep.	36.786	9.211	9.254	9.128	9.193	0.009
4	1.00 shared	9.178	2.310	2.271	2.344	2.253	0.373
4	1.00 shortest	9.178	2.295	2.292	2.295	2.296	0.933

Table 6.2.2 verifies that the statistics of arrival and service headway sequences are close to exponential (coefficient of variation ≅ 1) and uncorrelated in a single queue. There is some correlation (Pearson r) between arrivals and service on lanes with shared service, as may be expected since arrivals correlate with queue size, and queue size correlates with service since a larger queue is more likely to ‘borrow’ service from another lane.

Table 6.2.2 Arrival and service intervals in simulations (ρ=0.9, μ=1.0 shared, 1M events)

	Arrivals					**Service**				
	Single	**Multi-Lane Queuing**				**Single**	**Multi-Lane Queuing**			
	Queue	**Lane1**	**Lane2**	**Lane3**	**Lane4**	**Queue**	**Lane1**	**Lane2**	**Lane3**	**Lane4**
Mean	1.112	4.453	4.454	4.439	4.451	0.999	4.003	4.001	3.990	4.002
S.d.	1.114	4.452	4.465	4.462	4.461	1.002	4.045	4.058	4.041	4.037
Co.Var	1.002	1.000	1.002	1.005	1.002	1.002	1.010	1.014	1.013	1.009
ρ_{eff} :$r_{a\leftrightarrow s}$	0.899	0.899	0.898	0.899	0.899	0.079	0.543	0.547	0.543	0.544

Table 6.2.3 shows how the probability of the queue being zero varies. Here it is important to distinguish between the value of p_0 for all lanes combined and the average of the individual lane values. The total and individual lane queues are consistent with the corresponding M/M/1 values of p_0, i.e. $1/(L_e+1)$, except for the case where arrivals select the shortest queue, when the individual queues are less likely to be zero, although the combined arrival and service events are still uncorrelated because they are generated by the same independent mechanisms.

Table 6.2.3 Probability of zero queue in simulations (ρ=0.9, μ=1.0, 1M events)

No. of	**Capacity μ and**	**Average**	**Combined p_0**	**Individual Lane p_0**			
Lanes	**Type of process**	**Lane p_0**	(~1-ρ)	**Lane 1**	**Lane 2**	**Lane 3**	**Lane 4**
1	1.00 random	0.100	0.100	0.100	-	-	-
4	0.25 indep.	0.099	n/a	0.100	0.098	0.098	0.099
4	1.00 shared	0.377	0.105	0.376	0.378	0.375	0.378
4	1.00 shortest	0.250	0.105	0.251	0.250	0.249	0.248

Table 6.2.4 gives the variances of the queue sizes in the four tests (nominal value = 90). The difference between the values in the first two cases can be ascribed to the high sensitivity of the M/M/1 equilibrium value to the effective ρ value ($dV/d\rho$=1900 @ ρ=0.9). Since simulated ρ is typically 'out' by 0.001-0.002, around half the variability in V is accounted for by this. However, the lane variances are much smaller values in the last two cases, where lanes share service. The mechanism of this is analysed later (for the case of two lanes).

Table 6.2.4 Variance of queue size in multi-lane cases (ρ=0.9, μ=1.0, 1M events)

No. of Lanes	**Capacity μ and Type of process**	**Average Lane Variance**	**Individual Lane Variance**			
			Lane 1	**Lane 2**	**Lane 3**	**Lane 4**
1	1.00 random	82.413	82.413	-	-	-
4	0.25 indep.	94.790	95.934	94.355	97.743	91.129
4	1.00 shared	11.662	11.428	10.941	12.721	11.556
4	1.00 shortest	6.505	6.515	6.502	6.490	6.512

Kleinrock (1975) states[57] that where a queue is present the queue size distribution is geometric for any G/M/c process, c being number of parallel channels. A multi-lane system is not G/M/c, but Table 6.2.5 gives 'effective rho' values, the average ratio between successive probabilities η on the *assumption* that the distribution is geometric, i.e. $p_i = p_0\eta^i$. This assumption would imply p_0 = 1-η. The average η from simulation in the first two cases can be considered practically equal to ρ, and p_0 is consistent with geometric distribution.

Table 6.2.5 Geometric ratio in probability distribution (ρ=0.9, μ=1.0, 1M events)

No. of Lanes	**Capacity μ and Type of process**	**Average Lane p_0**	**Average η**	**Individual Lane η**			
				Lane 1	**Lane 2**	**Lane 3**	**Lane 4**
1	1.00 random	0.100	0.899	0.899	-	-	-
4	0.25 indep.	0.099	0.909	0.903	0.896	0.899	0.898
4	1.00 shared	0.377	0.761	0.759	0.756	0.762	0.768
4	1.00 shortest	0.250	0.691	0.688	0.696	0.692	0.689

Where arrivals choose the shortest queue, the probability distributions are no longer truly geometric in shape since p_0 is relatively reduced, as seen in Figure 6.2.2, where the distributions are sampled showing some variability, and Figure 6.2.3 where the distributions are equilibriated (aggregated between 10,000 and 1M events, compared to the theoretical relaxation period for ρ=0.9 of ~380 service events) and there is no visible difference between the lanes.

[57] Kleinrock (1975), p246-249.

In the two shared service cases the average η from simulation is substantially different from ρ, but with random selection the distribution has a geometric appearance. Where the shortest queue is selected the distribution is no longer geometric, and an appropriate approximation would be nested geometric (see Chapter 3 earlier). However, since p_0 does not contribute to moments, they can still be estimated by substituting an 'effective rho' into M/M/1 formulae.

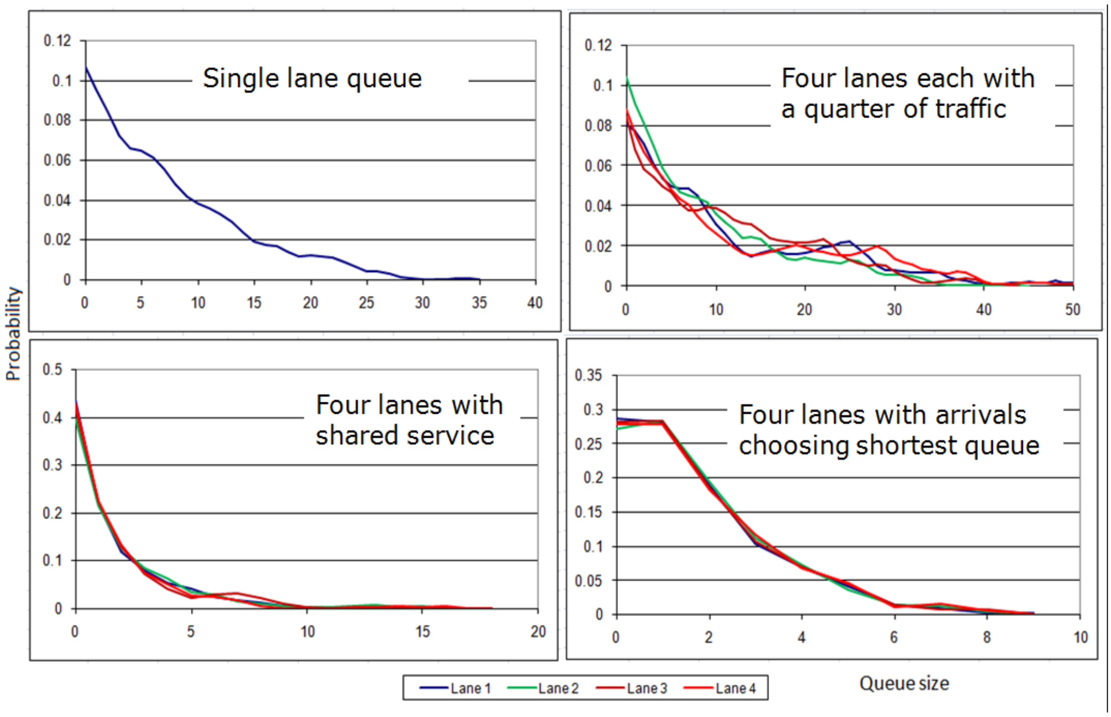


Figure 6.2.2 Sample queue size distributions for 4 lanes (ρ=0.9, μ=1.0, 9000 events)

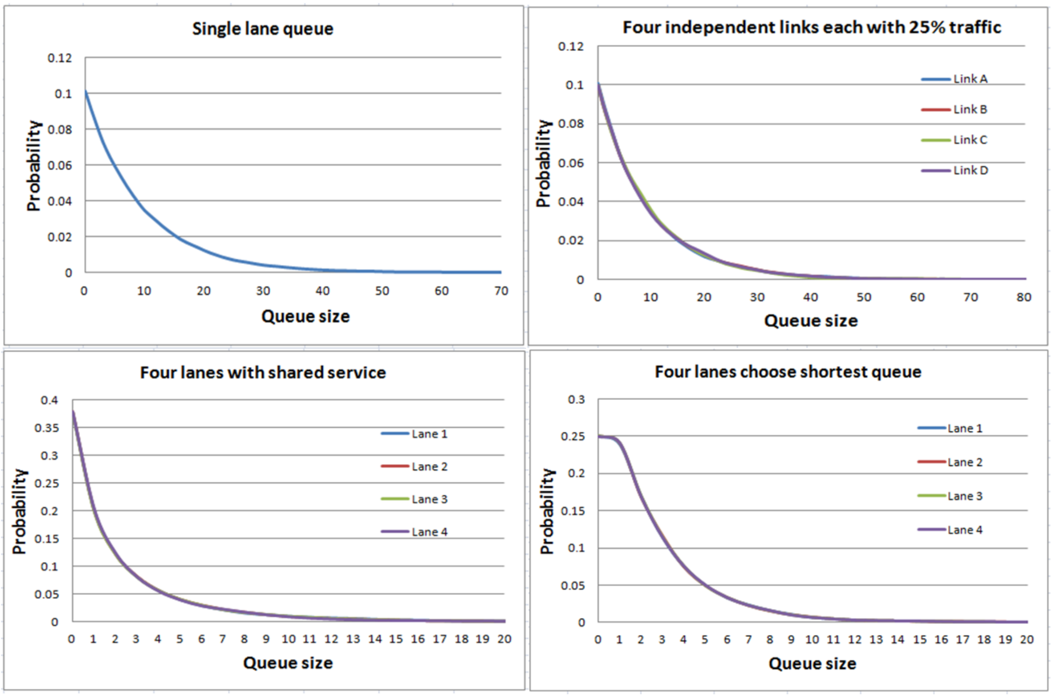


Figure 6.2.3 Equilibriated distributions for 4 lane cases (ρ=0.9, μ=1.0, 1M events)

To produce on each lane a mean equilibrium queue size that equals the total queue divided by the number of lanes n, assuming that p_0 fits the pattern, the effective demand intensity on each lane should be given by equation (6.2.4). Figure 6.2.4, comparing estimated and simulated moments of various shared exit cases with ρ=0.7 or 0.9, and shortest queue selection with ρ=0.9, shows that this works fairly well given the simplicity of the approximation[58].

$$\eta = \frac{\rho}{\rho + n(1-\rho)} \tag{6.2.4}$$

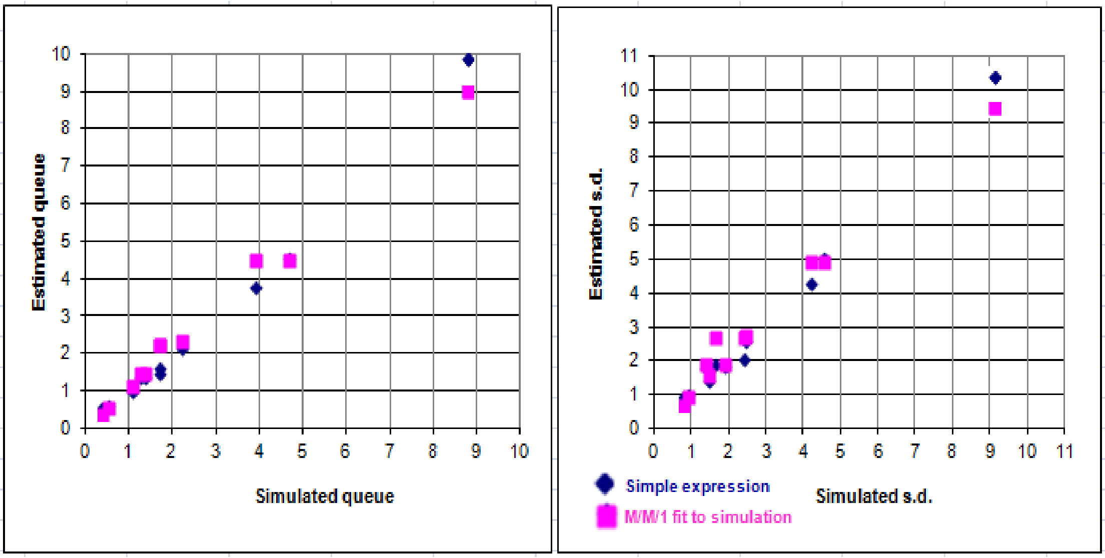


Figure 6.2.4 Comparing queue moments based on estimated η with simulation

6.2.3 Markov model of multi-lane queuing

If the system has n identical lanes, then the mean size of each queue is L_e/n and the probability distributions of the individual lane queues must combine to reproduce the combined M/M/1 geometric distribution. However, the lane queues are not independent. Even if arrivals and departures are randomly chosen, choosing one lane necessarily affects the other(s). First define:

$p^{(n)}_{i,j,k,l\ldots}$ = absolute probability of the n lane queues being in state $i,j,k,l\ldots$
$p^{(n)}_{i(m)}$ = prob. that lane queue indexed m of n is in state i (6.2.5)
P_i or $p_{i(n)}$ = the probability that the queue size on a typical lane of n is i.

A 'Markovian' process is memoryless, so its future state depends only on the current state. A multi-lane queuing process can be modelled as a continuous-time process:

[58] Subject to the largest queue (n=1), corresponding to the largest value of η (=ρ), being simulated by a sufficiently large number of events to bring it near to equilibrium, as mentioned earlier.

$$\frac{1}{\mu}\frac{dp_{ijk..}}{dt} = M\left(\left\{p_{i'j'k'..}(\mu t)\right\}, i, j, k..\rho, \alpha_1, \alpha_2, ..\right) \tag{6.2.6}$$

$$p_{ijk..} \leftarrow p_{ijk..} + \frac{dp_{ijk..}}{dt}\delta t \tag{6.2.7}$$

where M is some function, primed indices are dummies representing all probabilities of the current state at time t, and $\{\alpha_m\}$ are various option settings. Since μ is always associated with t it can be convenient work in terms of μt instead of t. The Markov process can be implemented by a computer program using a finite step size $\mu\delta t$, and will be stable if this is not too large, although a very small value will increase computation time with little benefit to accuracy. The formulation of recurrence relations will be described and discussed later. To establish relevant behaviour at the outset, some results of simulation are given first.

6.2.4 Queue size probability distributions from Markov simulation of 2 lanes

Figure 6.2.5 shows a clear difference between the two-server (broken line) and M/M/1 (nearby solid line) distributions, while the distributions on individual lanes are similar regardless of how generated (three solid lines with higher p_0). The dip in p_0 in the selective case is present though less evident than in the previous 4-lane simulations. These results confirm the earlier view that the critical factor is the service process, and that at least in symmetrical cases and where the primary objective is to estimate moments, the individual lane processes can be approximated by M/M/1 with a modified ‘effective ρ’, opening the way to simplified approximations amenable to standard time-dependent methods.

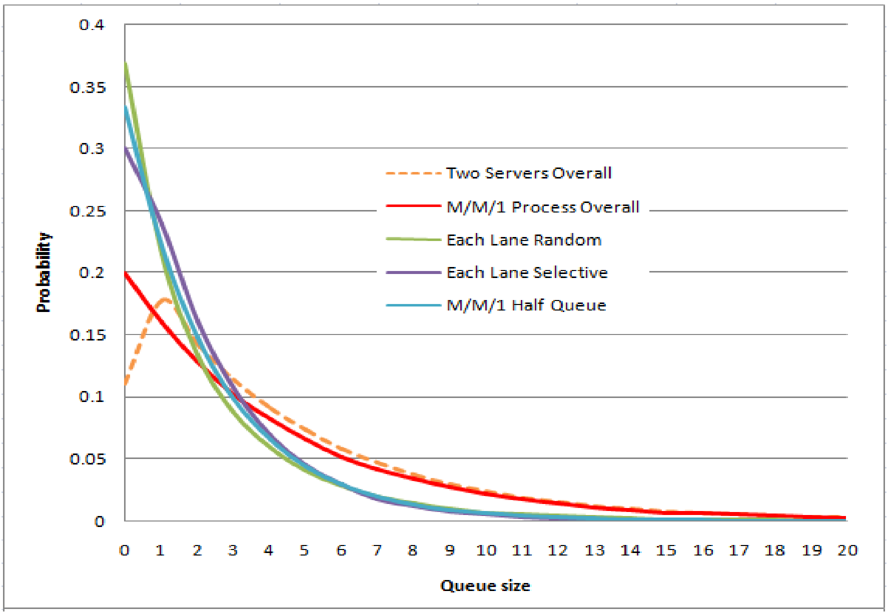


Figure 6.2.5 Markov simulated queue size distributions for some two-lane cases with ρ=0.8

The first parts of the probability distributions, simulated for ρ=0.8, μ=1, and δ*t*=0.01 in equation (6.2.7), and 50,000 iteration cycles, are tabulated in Table 6.2.6, showing that the queue size probability distribution in each lane is not dissimilar to that of an M/M/1 distribution with half the mean queue size. Some invariants are given in Table 6.2.7. Tables 6.2.8 and 6.2.9 later give more detail about probabilities for a two lane system with random selection of lanes.

Table 6.2.6 Probability distributions in whole system or each of 2 lanes, ρ=0.8

Queue Size *i*	**Whole System**	**Each Lane Random**	**Each Lane Selective**	**M/M/1 half mean queue**
0	0.200	0.369	0.300	0.333
1	0.160	0.215	0.239	0.222
2	0.128	0.134	0.161	0.148
3	0.102	0.088	0.107	0.099
4	0.082	0.059	0.069	0.066
5	0.066	0.040	0.045	0.044
6	0.052	0.028	0.029	0.029
7	0.042	0.020	0.018	0.020
8	0.034	0.014	0.012	0.013
9	0.027	0.010	0.008	0.009
10	0.022	0.007	0.005	0.006
11	0.017	0.005	0.003	0.004
12	0.014	0.004	0.002	0.003
13	0.011	0.003	0.001	0.002
14	0.009	0.002	0.001	0.001
15	0.007	0.001	0.001	0.001
16	0.006	0.001	0.000	0.001

Table 6.2.7 Invariant values for the Markov simulated two-lane cases

Queue Size *i*	**Two Servers Total**	**M/M/1 Process Total**	**Each Lane Random**	**Each Lane Selective**	**M/M/1 Half Queue**
p_0	0.111	0.200	0.369	0.300	0.333
L_e	4.444	4.000	2.000	2.000	2.000

The probability P_0 of either lane being empty regardless of the other is the sum of the probabilities that one lane is empty and the other lane has any possible value, i.e. the row-0 or column-0 sum in Table 6.2.8:

$$P_0 = \sum_{i=0}^{i=\infty} p_{0i} \qquad (6.2.8)$$

If the queue in one lane is zero, knowing that the probability of both lanes being empty is 0.2 (=1-ρ), the probability that the other lane is also empty has a value much greater than that of an independent lane, viz:

$$P_0^{(i)}\left|\left(j=0\right)=\frac{0.2}{0.3686}=0.5426\right. \qquad (6.2.9)$$

In Table 6.2.8, the probability of both lane queues being empty is not the product of the probabilities of each lane being empty. If the two lanes were independent, and each had on average half the total arrivals and half the total service, then each would still have the same ρ=0.8 and probability 0.2 of being zero, so the probability of both being zero would be only 0.2^2 = 0.04. However, *each* lane queue would then have the same mean size as the actual two-lane system. This violation of the 'product form' (Koenigsberg 1991) is an indication of non-independence of the lane queues, which naturally results in diagonal joint probabilities being greater than the products (squares) of the corresponding lane probabilities. The differences might be seen as a measure of inter-lane correlation (if the lanes were perfectly correlated each diagonal joint probability would be equal to the corresponding individual lane probability).

Table 6.2.8 Markov simulated joint and lane state probabilities, random selection

ρ=0.8	**Sums**	0.3686	0.2148	0.1344	0.0877	0.0588	0.0402	0.0278	0.0195
Sums	***i,j***	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0.3686	**0**	0.2001	0.0800	0.0375	0.0198	0.0113	0.0069	0.0043	0.0028
0.2148	**1**	0.0800	0.0530	0.0314	0.0186	0.0113	0.0070	0.0045	0.0029
0.1344	**2**	0.0375	0.0314	0.0221	0.0147	0.0096	0.0063	0.0041	0.0028
0.0877	**3**	0.0198	0.0186	0.0147	0.0107	0.0075	0.0051	0.0035	0.0024
0.0588	**4**	0.0113	0.0113	0.0096	0.0075	0.0055	0.0040	0.0028	0.0020
0.0402	**5**	0.0069	0.0070	0.0063	0.0051	0.0040	0.0030	0.0022	0.0016
0.0278	**6**	0.0043	0.0045	0.0041	0.0035	0.0028	0.0022	0.0017	0.0013
0.0195	**7**	0.0028	0.0029	0.0028	0.0024	0.0020	0.0016	0.0013	0.0010
0.0137	**8**	0.0019	0.0019	0.0019	0.0017	0.0014	0.0012	0.0009	0.0007
0.0097	**9**	0.0012	0.0013	0.0013	0.0011	0.0010	0.0008	0.0007	0.0005
0.0069	**10**	0.0008	0.0009	0.0009	0.0008	0.0007	0.0006	0.0005	0.0004
0.0050	**11**	0.0006	0.0006	0.0006	0.0006	0.0005	0.0004	0.0004	0.0003
0.0036	**12**	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003	0.0002
0.0026	**13**	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	0.0002	0.0002
0.0019	**14**	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001
0.0014	**15**	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
0.0010	**16**	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
0.0007	**17**	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
0.0005	**18**	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000	0.0000	0.0000
0.0004	**19**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0003	**20**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

6.2.5 Interpretation of shared service as ‘borrowing’ of capacity between lanes

The interpretation that unused capacity in one lane is ‘borrowed’ by the other is justified by the state transition model in Figure 6.2.6, where it is made explicit when one lane is empty.

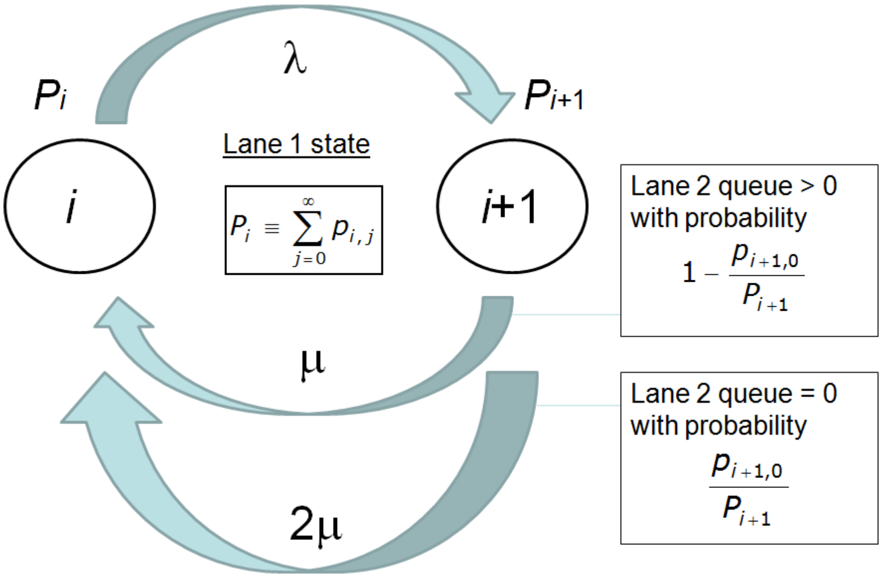

P_i
λ
P_{i+1}
Lane 1 state
i
$P_i \equiv \sum_{j=0}^{\infty} p_{i,j}$
i+1
Lane 2 queue > 0 with probability $1-\frac{p_{i+1,0}}{P_{i+1}}$
μ
Lane 2 queue = 0 with probability $\frac{p_{i+1,0}}{P_{i+1}}$
2μ

Figure 6.2.6 State transition diagram for two lanes

The state probabilities P_i are for Lane 1 alone, not technically ‘in isolation’ but as if Lane 2 and correlations were hidden from an observer. The probability ratio in the diagram follows from the fact that, in considering the transition on Lane 1 from queue state i+1 to state i through a service event, what is required is the probability of Lane 2 being empty *given* that the queue in Lane 1 is known to be i+1. The transition balance equation is therefore:

$$\lambda P_i = \mu P_{i+1}\left(1-\frac{p_{i+1,0}}{P_{i+1}}\right) + 2\mu P_{i+1}\left(\frac{p_{i+1,0}}{P_{i+1}}\right) = \mu\left(P_{i+1} + p_{i+1,0}\right) \qquad \text{so}$$

$$P_{i+1} - \rho P_i = -p_{i+1,0} \qquad (6.2.10)$$

Full recurrence relations for this and other cases, and what can (or cannot) be extracted from them, are presented and discussed next.

6.2.6 Evolution of probabilities where arrivals select one queue at random

For n=2, the following formulae can be derived by considering each possible initial state leading to a given final state, where one index is allowed to change at a time:

$$\frac{1}{\mu}\frac{dp_{0,0}}{dt} = p_{0,1} + p_{1,0} - \rho p_{0,0}$$

$$\frac{1}{\mu}\frac{dp_{0,j}}{dt} = .5p_{1,j} + p_{0,j+1} - (1+\rho)p_{0,j} + .5\rho p_{0,j-1} \qquad (j>0)$$

$$\frac{1}{\mu}\frac{dp_{i,0}}{dt} = .5p_{i,1} + p_{i+1,0} - (1+\rho)p_{i,0} + .5\rho p_{i-1,0} \qquad (i>0) \qquad (6.2.11)$$

$$\frac{1}{\mu}\frac{dp_{i,j}}{dt} = .5p_{i+1,j} + .5p_{i,j+1} - (1+\rho)p_{i,j} + .5\rho p_{i-1,j} + .5\rho p_{i,j-1} \qquad (i,j>0)$$

RHS of equations (6.2.11) are expanded in Tables 6.2.9 for low values of i and j. In the steady state all the cell sums are identically zero.

Table 6.2.9a Rho-factored (arrival) components of dp_{ij}/dt=0

$i \backslash j$	0	1	2	3
0	$-p_{00}$	$-p_{01}+.5p_{00}$	$-p_{02}+.5p_{01}$	$-p_{03}+.5p_{02}$
1	$.5p_{00}-p_{10}$	$.5p_{01}+.5p_{10}-p_{11}$	$.5p_{02}+.5p_{11}-p_{12}$	$.5p_{03}+.5p_{12}-p_{13}$
2	$.5p_{10}-p_{20}$	$.5p_{11}+.5p_{20}-p_{21}$	$.5p_{12}+.5p_{21}-p_{22}$	$.5p_{13}+.5p_{22}-p_{23}$
3	$.5p_{20}-p_{30}$	$.5p_{21}+.5p_{30}-p_{31}$	$.5p_{22}+.5p_{31}-p_{32}$	$.5p_{23}+.5p_{32}-p_{33}$

Table 6.2.9b Unit-factored (departure) components of dp_{ij}/dt=0

$i \backslash j$	0	1	2	3
0	$p_{10}+p_{01}$	$.5p_{11}+p_{02}-p_{01}$	$.5p_{12}+p_{03}-p_{02}$	$.5p_{13}+p_{04}-p_{03}$
1	$p_{20}+.5p_{11}-p_{10}$	$.5p_{21}+.5p_{12}-p_{11}$	$.5p_{22}+.5p_{13}-p_{12}$	$.5p_{23}+.5p_{14}-p_{13}$
2	$p_{30}+.5p_{21}-p_{20}$	$.5p_{31}+.5p_{22}-p_{21}$	$.5p_{32}+.5p_{23}-p_{22}$	$.5p_{33}+.5p_{24}-p_{23}$
3	$p_{40}+.5p_{31}-p_{30}$	$.5p_{41}+.5p_{32}-p_{31}$	$.5p_{42}+.5p_{33}-p_{32}$	$.5p_{43}+.5p_{34}-p_{33}$

Noting the symmetry $p_{ij} = p_{ji}$, summing the rows or columns of the elementary probabilities p_{ij} gives the following relationships, where P_i represents the i^{th} sum, the probability that the queue size in *one* lane is i:

$$
\begin{aligned}
&P_1 - \rho P_0 = -p_{01} = -.5\rho p_{00} \\
&P_2 - (1+\rho)P_1 + \rho P_0 = p_{01} - p_{02} \qquad (6.2.12)\\
&P_3 - (1+\rho)P_2 + \rho P_1 = p_{02} - p_{03}
\end{aligned}
$$

etc., or summing cumulatively

$$
\begin{aligned}
&P_1 - \rho P_0 = -p_{01} \\
&P_2 - \rho P_1 = -p_{02} \qquad (6.2.13)\\
&P_3 - \rho P_2 = -p_{03}
\end{aligned}
$$

So equation (6.2.10) is recovered. The LHS above are identical to those for the M/M/1 queue, but the RHS are now non-zero. Summing the equations, using (6.2.8), leads only to the known result p_{00}=(1-ρ).

6.2.7 Estimation of two lane moments and probability distributions

The first moment L_j of equations (6.2.13) can be related to the ensemble mean by:

$$L_j + \ell_0 = L_e \qquad \text{where} \quad \ell_0 = \frac{\sum_0^\infty i p_{0i}}{1-\rho} \qquad (6.2.14)$$

where L_j is the actual equilibrium queue in the jth lane, L_e is the equilibrium queue of the whole M/M/1 system ρ/(1-ρ) (the subscript e could be interpreted here as standing for 'ensemble'), and ℓ_0 represents a 'shadow queue' contributed by the other lane. Taking the second moment:

$$V_j + L_j^2 + \upsilon_0 + \ell_0^2 + 2L_e\ell_0 = V_e + L_e^2 \qquad \text{where} \quad \upsilon_0 = \frac{\sum_0^\infty i^2 p_{0i}}{1-\rho} - \ell_0^2 \qquad (6.2.15)$$

Here, in addition to a 'shadow second moment' there is a term that could be described as a 'decoupling' or anti-correlation term (a correlation term would be negative). The 'shadow queue's probability distribution is not normalised, not even in the second of equations (6.2.14), because the sum of its probabilities is P_0, not 1 or (1-ρ). However, there is no obvious way to get an expression for P_0 from these expressions. As this is a specific system and no statistical parameters have been 'hidden', there is no logical prohibition of a closed-form solution, yet it appears not to be available even for a simple physical system like this. Nevertheless, it is interesting that all that is needed, to describe the system fully, is the state of one queue when the other is empty. By symmetry the two lane queues are equal, therefore from (6.2.14-15):

$$\sum_{0}^{\infty} p_{0i} = P_0 \qquad \sum_{0}^{\infty} i p_{0i} = \frac{\rho}{2} \tag{6.2.16}$$

$$L_j = \tfrac{1}{2} L_e \qquad \ell_0 = \tfrac{1}{2} L_e \qquad V_j + \upsilon_0 = V_e - \tfrac{1}{2} L_e^2 \tag{6.2.17}$$

Now $V_j \neq \upsilon_0 \neq \tfrac{1}{2} V_e$, and while L_j and ℓ_0 are equal their underlying probability distributions are different. These results can be confirmed using the data in Table 6.2.8. In all M/M/1 cases:

$$p_{00} = 1 - \rho \qquad \text{and for any number of lanes} \qquad p_{0\ldots0} = 1 - \rho \tag{6.2.18}$$

Therefore, the simplest estimate of the whole distribution $\{p_{0i}\}$ would be the unnormalised singly-nested distribution:

$$p_{0i} \approx (1-\rho)\eta^i \qquad (i \geq 0) \tag{6.2.19}$$

In practice this performs poorly, and in any case forces a relationship between p_{01} and η. In fact it is possible to obtain p_{01} explicitly. For any number of lanes, if the total queue is 1 then only one lane can contain the queue, but all lanes are equally likely, so $p_{0\ldots1}$ must be $1/n$ of the M/M/1 ensemble probability of a queue of size 1, hence:

$$p_{0\ldots1} \approx \frac{\rho(1-\rho)}{n} \qquad \text{and in particular} \qquad p_{01} \approx \frac{\rho(1-\rho)}{2} \tag{6.2.20}$$

With p_{00} and p_{01} known, the simplest progression of the distribution is geometric:

$$p_{0i} \approx p_{01}\eta^{i-1} \qquad (i > 1) \tag{6.2.21}$$

The solution for η to fit the mean (6.2.16) in this case is:

$$\eta = 1 - \sqrt{\frac{2p_{01}}{\rho}} = 1 - \sqrt{1-\rho} \tag{6.2.22}$$

An early attempt at estimating the whole distribution $\{p_{0i}\}$ was the empirical formula (6.2.23), which while consistent with (6.2.19-20) when $i \leq 1$, has no theoretical justification and would not be easy to extend to n>2, but is included for completeness:

$$p_{0i} \approx \frac{(1-\rho)\rho^i}{\left[2+(i-1)\sqrt{\frac{1}{2}i+1}\right]} \qquad (6.2.23)$$

Summing terms of (6.2.21) together with (6.2.18-19) gives the following value for $P_{0(2)}$:

$$P_{0(2)} \approx 1-\rho+\frac{\rho\sqrt{1-\rho}}{2} \qquad (6.2.24)$$

An earlier attempt to devise an empirical formulae for $P_{0(2)}$ is also included for completeness:

$$P_{0(2)} \approx 1-\frac{\rho(1+\rho)}{2(1+\rho-\rho^2)} \quad \text{(constructed from } \rho \text{ terms to 2}^{\text{nd}} \text{ order)} \qquad (6.2.25)$$

Figure 6.2.7 compares estimates of P_0 using different methods.

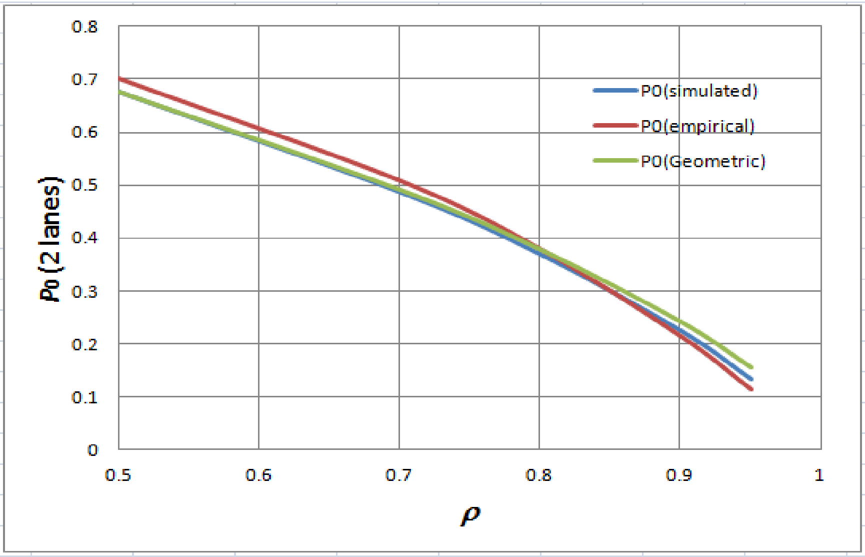


Figure 6.2.7 Estimates of P_0 for two lanes and a range of traffic intensities

The mean value of $\{p_{0i}\}$ according to the formulation (6.2.20-22) is, as required:

$$\sum_0^\infty ip_{0i} = \frac{p_{01}}{(1-\eta)^2} = \frac{\rho}{2} \qquad (6.2.26)$$

To complete a practical description it remains to determine the lane queue variance V_j. Using (6.2.20-22) the second moment of $\{p_{0i}\}$ can now be estimated as (6.2.27), and thence υ_0 from

(6.2.15), from which V_j can be obtained using the last of (6.2.17) (best done numerically since neither expression reduces to a suggestive form).

$$\sum_0^\infty i^2 p_{0i} = \frac{(1-\rho)\rho(1+\eta)}{2(1-\eta)^3} \qquad (6.2.27)$$

Alternative estimated probability distributions for ρ=0.8, including a doubly-nested distribution based on three moments including variance, are compared with simulation in Figure 6.2.8. The empirical distribution appears to fit the simulation data very well, which may make it worthy of future investigation. The similarity of estimated (6.2.19-21) and doubly-nested distributions is not surprising since they have similar forms each determined by three parameters.

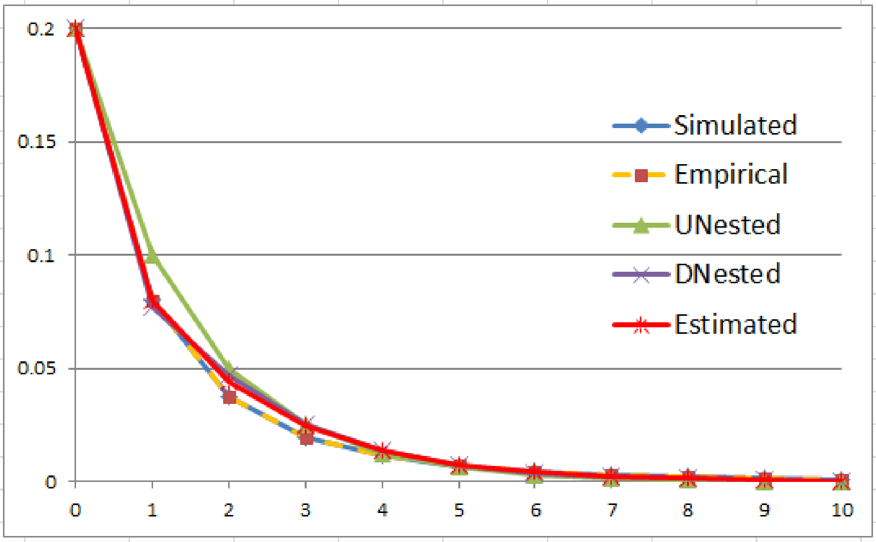


Figure 6.2.8 Simulated and estimated probability distributions $\{p_{0i}\}$, two lanes, ρ=0.8

Figure 6.2.9 plots the variance elements as functions of ρ. The estimated value of υ_0 goes negative when ρ = 0.95, which is allowed since it is not a true variance, and it remains small so V_j is generally around half V_e. If the lanes were independent, each with half the arrival and capacity rates, and hence the same demand intensity equal to the ensemble value ρ, the total queue size and variance would both be *twice* the ensemble values. On the other hand, to produce the same *total* queue with independent lanes, each lane would need a reduced demand intensity ρ/(2-ρ), and the total variance would be approximately *half* the ensemble variance. This shows that the variance of a lane queue more or less tracks the mean queue size as the relationship between the lane processes is varied.

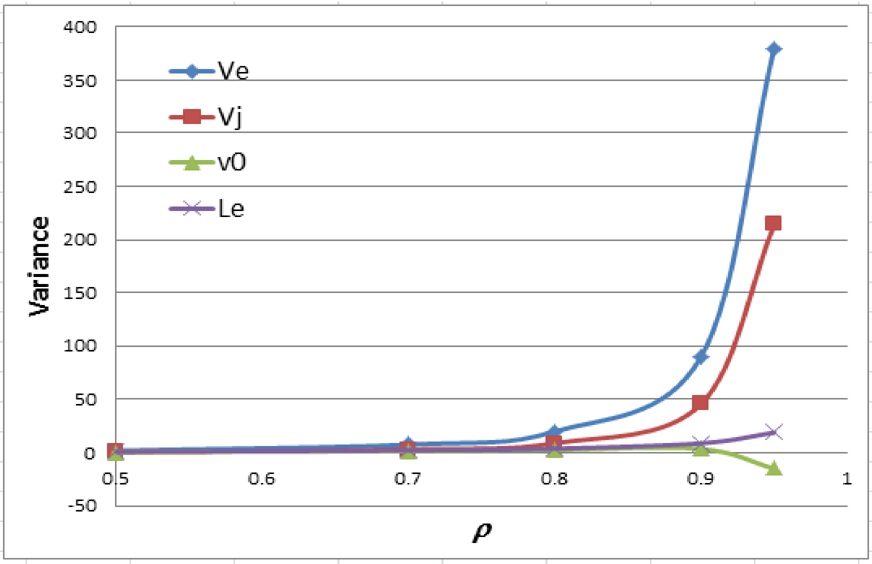


Figure 6.2.9 Calculated lane queue variances for two lanes and range of ρ values

6.2.8 Evolution of probabilities where arrivals select shorter of two queues

In the case of selection of the shorter queue between two lanes (or randomly only if the queues are exactly equal), equations (6.2.11) are modified as below, some terms being evaluated in Tables 6.2.10a,b:

$$\frac{1}{\mu}\frac{dp_{0,j}}{dt} = .5p_{1,j} + p_{0,j+1} - (1+\rho)p_{0,j} \qquad (j \geq 2)$$

$$\frac{1}{\mu}\frac{dp_{i,0}}{dt} = .5p_{i,1} + p_{i+1,0} - (1+\rho)p_{i,0} \qquad (i \geq 2)$$

$$\frac{1}{\mu}\frac{dp_{i,j}}{dt} = .5p_{i+1,j} + .5p_{i,j+1} - (1+\rho)p_{i,j} + \rho p_{i-1,j} + .5\rho p_{i,j-1} \qquad (i=j\text{-}1)$$

$$\frac{1}{\mu}\frac{dp_{i,j}}{dt} = .5p_{i+1,j} + .5p_{i,j+1} - (1+\rho)p_{i,j} + .5\rho p_{i-1,j} + \rho p_{i,j-1} \qquad (i=j+1)$$

$$\frac{1}{\mu}\frac{dp_{i,j}}{dt} = .5p_{i+1,j} + .5p_{i,j+1} - (1+\rho)p_{i,j} + \rho p_{i-1,j} + \rho p_{i,j-1} \qquad (i=j)$$

$$\frac{1}{\mu}\frac{dp_{i,j}}{dt} = .5p_{i+1,j} + .5p_{i,j+1} - (1+\rho)p_{i,j} + \rho p_{i-1,j} \qquad (i<j\text{-}1)$$

$$\frac{1}{\mu}\frac{dp_{i,j}}{dt} = .5p_{i+1,j} + .5p_{i,j+1} - (1+\rho)p_{i,j} + \rho p_{i,j-1} \qquad (i>j+1)$$

(6.2.28)

Table 6.2.10a Rho-factored (arrival) components of dp_{ij}/dt=0

$i\setminus j$	0	1	2	3	4
0	$-p_{00}$	$-p_{01}+.5p_{00}$	$-p_{02}+.5p_{01}$	$-p_{03}+.5p_{02}$	$-p_{04}+.5p_{03}$
1	$.5p_{00}-p_{10}$	$p_{01}+p_{10}-p_{11}$	$p_{02}+.5p_{11}-p_{12}$	$p_{03}-p_{13}$	$p_{04}-p_{14}$
2	$.5p_{10}-p_{20}$	$.5p_{11}+p_{20}-p_{21}$	$p_{12}+p_{21}-p_{22}$	$.5p_{22}+p_{13}-p_{23}$	$p_{14}-p_{24}$
3	$.5p_{20}-p_{30}$	$p_{30}-p_{31}$	$.5p_{22}+p_{31}-p_{32}$	$p_{23}+p_{32}-p_{33}$	$.5p_{33}+p_{24}-p_{34}$
4	$.5p_{30}-p_{40}$	$p_{40}-p_{41}$	$p_{41}-p_{42}$	$.5p_{33}+p_{42}-p_{43}$	$p_{34}+p_{43}-p_{44}$
5	$.5p_{40}-p_{50}$	$p_{50}-p_{51}$	$p_{51}-p_{52}$	$p_{52}-p_{53}$	$.5p_{44}+p_{53}-p_{54}$

Table 6.2.10b Unit-factored (departure) components of dp_{ij}/dt=0

$i\setminus j$	0	1	2	3	4
0	$p_{10}+p_{01}$	$.5p_{11}+p_{02}-p_{01}$	$.5p_{12}+p_{03}-p_{02}$	$.5p_{13}+p_{04}-p_{03}$	$.5p_{14}+p_{05}-p_{04}$
1	$p_{20}+.5p_{11}-p_{10}$	$.5p_{21}+.5p_{12}-p_{11}$	$.5p_{22}+.5p_{13}-p_{12}$	$.5p_{23}+.5p_{14}-p_{13}$	$.5p_{24}+.5p_{15}-p_{14}$
2	$p_{30}+.5p_{21}-p_{20}$	$.5p_{31}+.5p_{22}-p_{21}$	$.5p_{32}+.5p_{23}-p_{22}$	$.5p_{33}+.5p_{24}-p_{23}$	$.5p_{34}+.5p_{25}-p_{24}$
3	*Etc – same as Table 6.2.9b*				

Unlike in the non-selective case, these formulae do not lend themselves to a repetitive simple relationship, but there is some regularity. Other terms besides p_{0i} now appear on the RHS. Eliminating P_0 from the first two equations gives the next pair relationship, but subsequent relationships become increasingly messy.

$$\begin{aligned}
&P_1-\rho P_0=-p_{01}\\
&(P_2-P_1)-2\rho(P_1-P_0)=p_{01}-p_{02}-\rho(p_{11}+2p_{10}-p_{00}) \qquad (6.2.29)\\
&(P_3-P_2)-2\rho(P_2-P_1)=p_{02}-p_{03}-\rho(p_{22}+2p_{20}-p_{11}+2p_{21}-p_{10})
\end{aligned}$$

... or ...

$$P_2-(2\rho-1)P_1=-p_{01}-p_{02}-\rho(p_{11}+2p_{10}-p_{00}) \qquad (6.2.30)$$

It is possible that arrivals gain some advantage from ‘jockeying’, that is changing lane whenever there appears to be an advantage in doing so. This could lead to almost infinite variation in the details of lane selection and service sharing mechanisms. Nevertheless, if the whole system is treated as a ‘black box’ with random arrivals and service, whose n internal lanes are identical, then symmetry requires that the means of all the lane queues be equal to $1/n$ of the M/M/1 queue, their probability distributions are the same, and the convolution of all the lane probability distributions must give the usual M/M/1 queue size probability distribution. What appears free to vary within these constraints is the *correlation* between lane queues.

6.3. EXTENDING TO ANY NUMBER OF LANES

6.3.1 Relationship to multi-channel queue process

It is natural to ask whether queues in multi-lane systems could be represented by using or adapting standard queue processes or modifying their parameters. Multi-channel queues with independent servers are described by the M/M/*n* or G/G/*n* processes, the first of which was summarised earlier in Section 3.5, representing something like a supermarket checkout or communications system with *n* channels. Arrivals choose an idle server if one is available and otherwise choose randomly, and there is no interaction between channels' service. In standard works the focus tends to be on evaluating waiting times in cases where individual queues are short, implying that server provision can be matched to demand. In transport the focus is more on evaluating queue sizes and delays where there are a few heavily loaded servers, since it may be difficult or impossible to provide extra capacity. The standard processes do not provide for explicit service sharing.

If ρ is the demand intensity on the whole ensemble relative to the sum of the channel capacities, the M/M/*n* steady-state queue (e.g. Medhi 2003) is:

$$L_{e(n)} = n\rho + C(n,\rho)\frac{\rho^2}{1-\rho} \qquad \text{where} \qquad C(n,\rho) = \frac{(n\rho)^n p_0^{(n)}}{n!\rho(1-\rho)} \qquad (6.3.1)$$

and $$p_{0(n)} = \left[\sum_{i=0}^{n-1}\frac{(n\rho)^i}{i!} + \frac{(n\rho)^n}{n!(1-\rho)}\right]^{-1} \qquad p_{i(n)} = \frac{n}{\min(i,n)}\rho p_{i-1}^{(n)} \quad (i>0) \qquad (6.3.2)$$

Here p_0 is the probability that the whole system is empty, and the coefficient of variation of service is assumed to equal 1 for each server, as implied by (6.3.1) when *n*=1. The probability distribution becomes geometric when *i*≥*n*. When *n*=1, the usual M/M/1 formulae are recovered. For *n*=2:

$$p_{0(2)} = \frac{1-\rho}{1+\rho} \qquad \text{and} \qquad L_{e(2)} = \frac{2\rho}{1-\rho^2} \qquad (6.3.3)$$

The mean queue size (6.3.3 right) lies between the extremes (6.2.1) and (6.2.3), which range would allow for channel selection by arrivals, but service sharing is excluded since the channels are physically independent. If the first term of the queue formula in (6.3.1) can be interpreted as total units-in-service, dividing it by *n* gives the expected value *I*=1 for each channel. However, as Figure 6.3.1 shows, *C* decreases with *n*, so if the average queue on each lane is got by dividing (6.3.1) by *n*, and *n* is large, the randomness term virtually disappears.

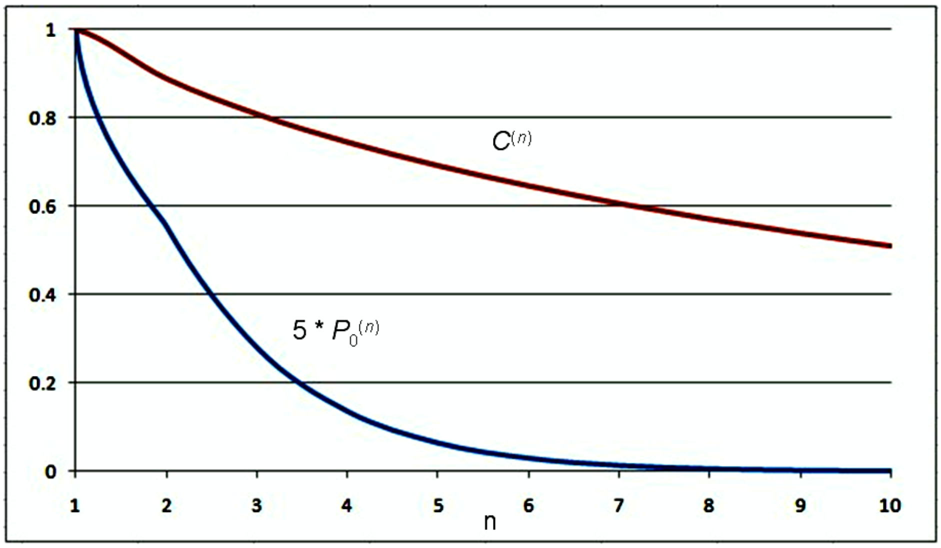


Figure 6.3.1 Variation of M/M/*n* quantities with number of channels *n*, ρ=0.8

Since over time arrivals are equally likely to select any channel, and the service processes are independent, each channel experiences demand intensity ρ. Although the arrival process seen by each channel is initially non-random, once sufficient queues have built up it becomes random and the statisitics of each channel process will be not far from M/M/1, while for the ensemble they *are* far from M/M/1. So this is not a good approximation to a multi-lane system.

6.3.2 Can multiple lanes be represented by P-K formula with modified parameters?

If a multi-lane ensemble process is M/M/1 and there is symmetry between lanes, the mean queue on each lane must be reduced by the factor of the number of lanes *n*:

$$L_{e(n)} = \frac{\rho}{n(1-\rho)} \qquad (6.3.4)$$

The P-K queue formulae derived earlier (equation 3.2.13 or 3.2.16 with 3.2.10), notwithstanding the arguments against the use of c_a, is:

$$L_e = I\rho + \frac{(I_a - 1)\rho}{2(1-\rho)} + \frac{\frac{1}{2}\left(1+c_b^2\right)\rho^2}{(1-\rho)} \approx I\rho + \frac{\frac{1}{2}\left(c_a^2 + c_b^2\right)\rho^2}{(1-\rho)} \qquad (6.3.5)$$

It might be expected that where arrivals choose the shorter of two queues, this could affect the dispersion of arrivals in each queue separately. In practice, the difference from random selection is likely to be small, since symmetry requires that the lane queues tend to become equal anyway. So to first approximation the dispersion-related terms in (6.3.5) can be eliminated by setting I_a=1 or c_a=1. To account for service sharing there are now only the parameters I and c_b. It is not obvious that I can meaningfully be subdivided, as it represents the unavoidable average time needed to service each customer. Subdividing the randomness coefficient in (6.3.5) appears impossible unless c_a is adopted and can be subdivided as well as c_b. This might just be possible for two lanes, but no more. In conclusion, manipulating P-K statistical coefficients whether in M/M/1 or M/M/n appears insufficient.

6.3.3 Are Censored Poisson or Erlang processes relevant to multi-lane queues?

In a Censored Poisson process every rth arrival is accepted and the rest discarded. Even though arrivals are random this is subtly different from M/M/r. The probability density function of accepted-arrival intervals is given by equation (6.3.6) (e.g. Trabka and Marchand 1970):

$$f(x) = \frac{k^r x^{r-1} e^{-kx}}{(r-1)!} \qquad \text{so that} \quad \text{mean}[f]=r/k,\ \text{var}[f]=r/k^{\,2} \qquad (6.3.6)$$

This is equivalent to the Erlang formula with substitutions of variables:

$$a(t) = \frac{rq(rqt)^{r-1} e^{-rqt}}{(r-1)!} \qquad x \leftrightarrow kt,\ q \leftrightarrow k/r \qquad (6.3.7)$$

An interpretation of (6.3.7) is that instead of (r-1) arrivals being discarded in each interval up to the rth arrival, they are parked and then combined with the rth in a bunch. The form of (6.3.7) is misleading since q is not the mean arrival rate but the inverse of the mean arrival interval, the true mean arrival rate of the bunched arrivals being k=rq. However, the mean arrival rate of the censored arrivals *is* q. Therefore the censored process can be represented as Erlang-r arrivals with arrival rate reduced by the factor r.

If distributing arrivals between r lanes is equivalent to 'censoring', and capacity equally divided between the lanes so ρ is the same for each lane as for the ensemble, then in each lane, from (3.2.16):

$$L_e = \frac{\rho}{1-\rho}\left[I(1-\rho) + \frac{1-r}{2r} + \rho \right] \qquad \text{(per lane)} \qquad (6.3.8)$$

Assuming the 'unit-in-service' coefficient I remains 1 independently of r, the results are as shown in Table 6.3.1.

Table 6.3.1 Queue size in each of *r* lanes predicted by equation (6.3.8)

r	**1**	**2**	**3**	**..**	**∞**
$L_{e(r)}$	$\left(\frac{\rho}{1-\rho}\right)$	$\frac{3}{4}\left(\frac{\rho}{1-\rho}\right)$	$\frac{2}{3}\left(\frac{\rho}{1-\rho}\right)$	..	$\frac{1}{2}\left(\frac{\rho}{1-\rho}\right)$

The total queue $rL_{e(r)}$ increases without limit, but is always less than it would be if the lanes were totally independent, which in this case means arrivals selecting lanes at random, since there is no interaction between the service processes. According to equation (3.2.16), including the Erlang service coefficient *m*, the only way to get the correct result is if *I*=*r*=*m*. This shows that multi-lane queuing is not equivalent to censoring arrivals.

6.3.4 Emergent patterns from simulation of two or more lanes

Markov simulation has been applied for up to four lanes with several values of ρ. Because of the somewhat inefficient technique of generating explicit transition tables for each initial state, run time escalates rapidly with the number of lanes, four being the current practical limit. Also, the reliability of the distributions cannot be guaranteed except that their convolution can be checked to equal the geometric distribution of the ensemble queue. This constraint applies regardless of how arrivals select a lane, e.g. randomly or selectively, provided that the combined arrivals and service are random, i.e. exponentially distributed.

If simulated lane queue probability distributions for *n*>1 are normalised to the same amplitude $p_{0(1)}$ and their state axes suitably transformed then they can be overlaid approximately as in Figure 6.3.2. Although the matches are approxmate and their quality is difficult to assess rigorously, there appears to be a simple transformation that links the distributions, which are expressed here as continuous functions because of the non-integral scaling factor, so interpolation is generally necessary to derive integer state probabilities:

$$P_{(.)}(x) = \frac{P_{(1)}(0)}{P_{(n)}(0)} P_{(n)}\left(\frac{x}{1+(n-1)\rho^2}\right) \qquad (6.3.9)$$

In equation (6.3.9), as previously, capital *P* is used for lane state probabilities to distinguish them from joint probabilities $p_{i...j}$. The presence of ρ^2 appears to be necessary for ρ<0.9, although not obtained by anything more rigorous than experiment and noting that differences between the moments of the transformed distributions are broadly minimised.

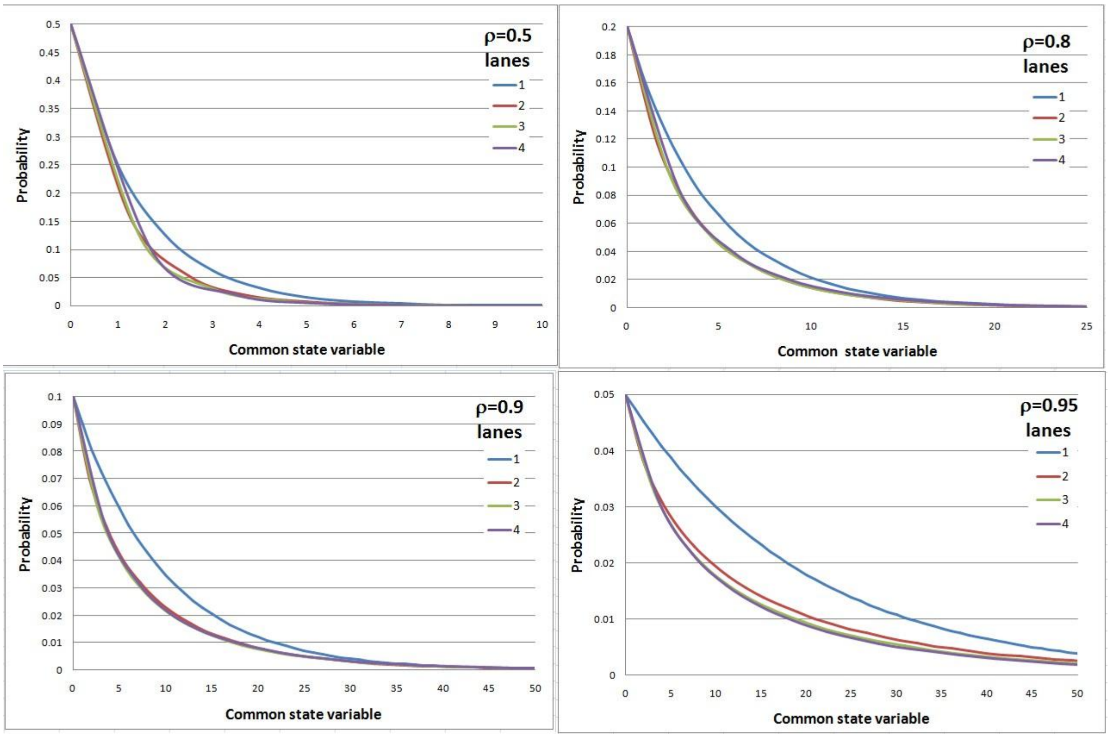


Figure 6.3.2 Markov multi-lane distributions against transformed state variable
(The graphs for a single lane are visibly separate from the multi-lane graphs)

The pattern does not extend to a single lane, as shown by the separate (light blue) curves. This is unavoidable, as all symmetrical lane selection strategies must result in the M/M/1 distribution of the total queue provided that the ensemble arrivals and service remain random. However, since different strategies result in different distributions, no one transformation could match the M/M/1 distribution. The rule starts to fail for 2 lanes in the case ρ=0.95, which may mean that it is not appropriate to cases where ρ is close to 1, which are not practical to simulate all the way to equilibrium. The n=1 curves can be moved into the common trend approximately by applying a scaling factor 0.75, though this does not work for ρ=0.95. Referring back to equation (6.3.4), if the distributions *were* M/M/1 their effective parameters would be given by (6.3.10), which is consistent with (6.3.9) when ρ is close to 1.

$$\rho^{/} \approx \frac{\rho}{\rho + n(1-\rho)} \approx \rho^{n} \quad (\rho \to 1) \qquad (6.3.10)$$

The true P_0 values depend on the lane selection process, so there is inevitably a degree of arbitrariness in any system where this is not specified. For practical purposes, adopting (6.3.9) gives results that are 'useful', although significantly inferior to those for 2 lanes obtained in the previous Section, and surprisingly, gives reasonable value for P_0, although by a formula quite different from those found earlier. Table 6.3.2 gives $\rho^{/}$ ('effective rho') values according to equation (6.3.10) and inferred from P_0. Table 6.3.3 shows increasing error in the estimated values of P_0 that is reflected in error in the distribution, despite the increasing *consistency* between the *n*-lane processes. Noting that, unlike the percentage errors in $\rho^{/}$ in Table 6.3.2, the errors in P_0 in Table 6.3.3 (relative sum-of-squares errors for the whole distributions are also shown) vary relatively little and in no consistent direction with *n*, and if only half-integral or integral powers of ρ are allowed, a simple formula for the percentage error is given by:

$$e_{\%} = \frac{\rho}{\left(1-\sqrt{\rho}\right)} \qquad (6.3.11)$$

Table 6.3.2 Comparison of 'Effective rho' values for M/M/1 multi-lane approximation

ρ	**Number of lanes**								
	2			**3**			**4**		
	1-*P*$_0$	$\rho^{/}$	**%error**	**1-*P*$_0$**	$\rho^{/}$	**%error**	**1-*P*$_0$**	$\rho^{/}$	**%error**
0.5	0.3244	0.3333	2.74	0.2421	0.25	3.26	0.1937	0.2	3.25
0.7	0.5138	0.5385	4.81	0.4106	0.4375	6.55	0.3439	0.3684	7.12
0.8	0.6315	0.6667	5.57	0.5278	0.5714	8.26	0.4562	0.5	9.6
0.9	0.7762	0.8182	5.41	0.6874	0.75	9.11	0.6192	0.6923	11.81
0.95	0.8675	0.9048	4.3	0.7935	0.8636	8.83	0.7284	0.8261	13.41

Table 6.3.3 *P*$_0$ values estimated by M/M/1 approximation and their errors

ρ	***P*$_0$ simulated/*estimated* by number of lanes**				***%Error in P$_0$*/%RSSE by lanes**		
	1	**2**	**3**	**4**	**2**	**3**	**4**
0.5	0.5	0.6756	0.7579	0.8063	*1.32*	*1.04*	*0.78*
	0.5	*0.6667*	*0.75*	*0.8*	1.33	1.29	1.09
0.7	0.3	0.4862	0.5894	0.6561	*5.07*	*4.56*	*3.74*
	0.3	*0.4615*	*0.5625*	*0.6316*	3.08	3.61	3.49
0.8	0.2	0.3685	0.4722	0.5438	*9.54*	*9.24*	*8.05*
	0.2	*0.3333*	*0.4286*	*0.5*	4.1	5.31	5.58
0.9	0.1	0.2238	0.3126	0.3808	*18.76*	*20.03*	*19.2*
	0.1	*0.1818*	*0.25*	*0.3077*	4.77	7.02	8.22
0.95	0.05	0.1325	0.2065	0.2716	*28.12*	*33.96*	*35.97*
	0.05	*0.0952*	*0.1364*	*0.1739*	4.46	7.84	10.6

Figure 6.3.3 shows that the error formula (6.3.11) is fairly accurate except for the points at upper right, corresponding to ρ=0.95, but Table 6.3.3 shows that the errors invariably take the form of underestimation, suggesting that a systematic adjustment is possible.

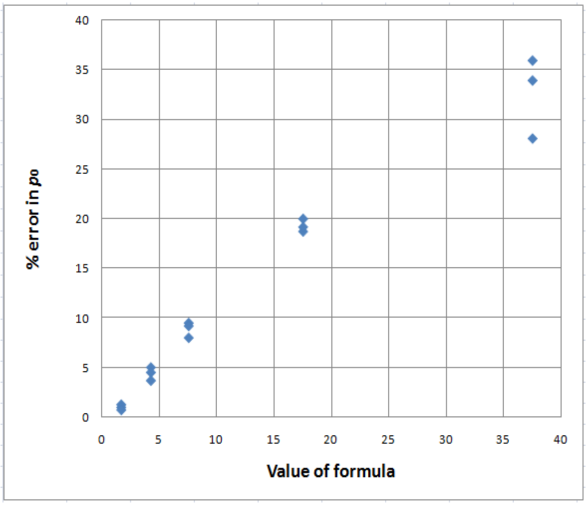


Figure 6.3.3 Performance of formula estimation percentage error in P_0

An adjusted 'effective ρ' for each lane, still retaining the M/M/1 model is thus:

$$\rho^{/} \approx 1-\left(1+\frac{\rho}{\left(1-\sqrt{\rho}\right)}\right)\left(1-\frac{\rho}{\rho+n\left(1-\rho\right)}\right) \qquad P_0 = 1 = \rho' \qquad (6.3.12)$$

Table 6.3.4 shows that this achieves a useful reduction in errors.

Table 6.3.4 P_0 values and errors based on adjusted estimates

ρ	**P_0 simulated/*estimated* by number of lanes**				**%*Error in P_0*/%RSSE by lanes**		
	1	**2**	**3**	**4**	**2**	**3**	**4**
0.5	0.5	0.6756	0.7579	0.8063	*0.36*	*0.65*	*0.91*
	0.5	*0.6780*	*0.7628*	*0.8137*	0.71	0.75	0.88
0.7	0.3	0.4862	0.5894	0.6561	*1.00*	*0.47*	*0.39*
	0.3	*0.4813*	*0.5866*	*0.6586*	1.79	1.95	1.77
0.8	0.2	0.3685	0.4722	0.5438	*2.69*	*2.36*	*1.09*
	0.2	*0.3586*	*0.4610*	*0.5379*	2.49	3.08	3
0.9	0.1	0.2238	0.3126	0.3808	*4.51*	*6*	*5.03*
	0.1	*0.2137*	*0.2938*	*0.3617*	3.1	4.18	4.55
0.95	0.05	0.1325	0.2065	0.2716	*1.15*	*9.19*	*11.94*
	0.05	*0.1310*	*0.1875*	*0.2392*	3.77	4.29	5.01

The resulting fit of P_0 and $L_{e(n)}$ is shown in Figure 6.3.4, the least accurate points in the mean being those associated with the highest demand intensity employed, ρ=0.95.

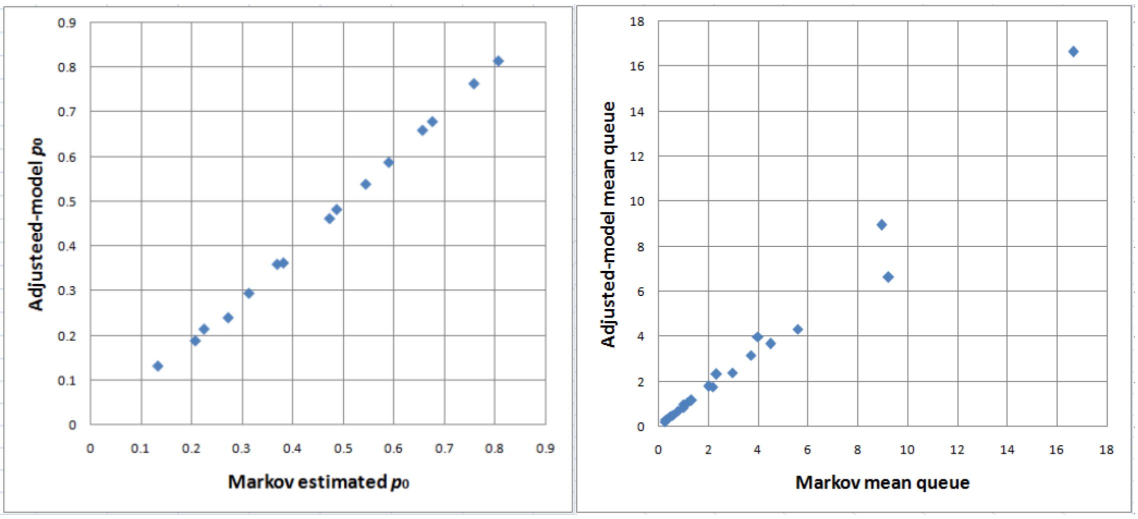


Figure 6.3.4 Fit of adjusted estimates of P_0 and mean queue to simulated values

6.3.5 Conclusions on extension to any number of lanes

The conclusions of this Section are:

- A standard multi-channel process is not appropriate for describing multiple lanes;

- Censored Poisson or Erlang arrivals *might* be able to reflect lane selection but cannot represent service sharing;

- When the ensemble queue is M/M/1, individual lane queue probability distributions are clearly not M/M/1, but the error in assuming them to be so is not great, bearing in mind that the actual processes of lane selection and service could be uncertain;

- Queue distributions for different numbers of lanes appear to be related approximately by a simple transformation. Since the distribution on two lanes can be estimated using the non-M/M/1 distribution described in the previous Section, distributions on more than two lanes can be inferred, although this may fail for $\rho \geq 0.95$;

- These are subject to the proviso that only M/M/1 ensembles have been considered, whereas one might expect multiple lane queues to be equally if not more prevalent at signals. However, up to this point signals queues would seem to require separate treatment.

6.4. MULTI-LANE METHOD WITH TURNS BASED ON UTILISATIONS

6.4.1 Motivation and approach

Attempts to find formulae for probability elements of lane queues have had limited success. With turning movements the situation becomes much more complicated. If each lane queue could be treated as *approximately* M/M/1, it could be expressed in Pollaczek-Khinchin form, even though as shown earlier the coefficients of dispersion and variation in P-K cannot be used to represent the effect of multiple lanes. If so, then all results obtained for the time-dependent sheared approximation can be applied. This Section derives an analytical approach based on an analysis of utilisations, referring back to Section 6.2.

6.4.2 Effect of service sharing and turning movements on utilisation

Figure 6.4.1 extends Figure 6.2.1 to include turning movements[59]. Left turners always use the left lane and right turners the right lane. Straight ahead movers may use either lane according to some choice mechanism, such as selecting the queue that appears to be shorter.

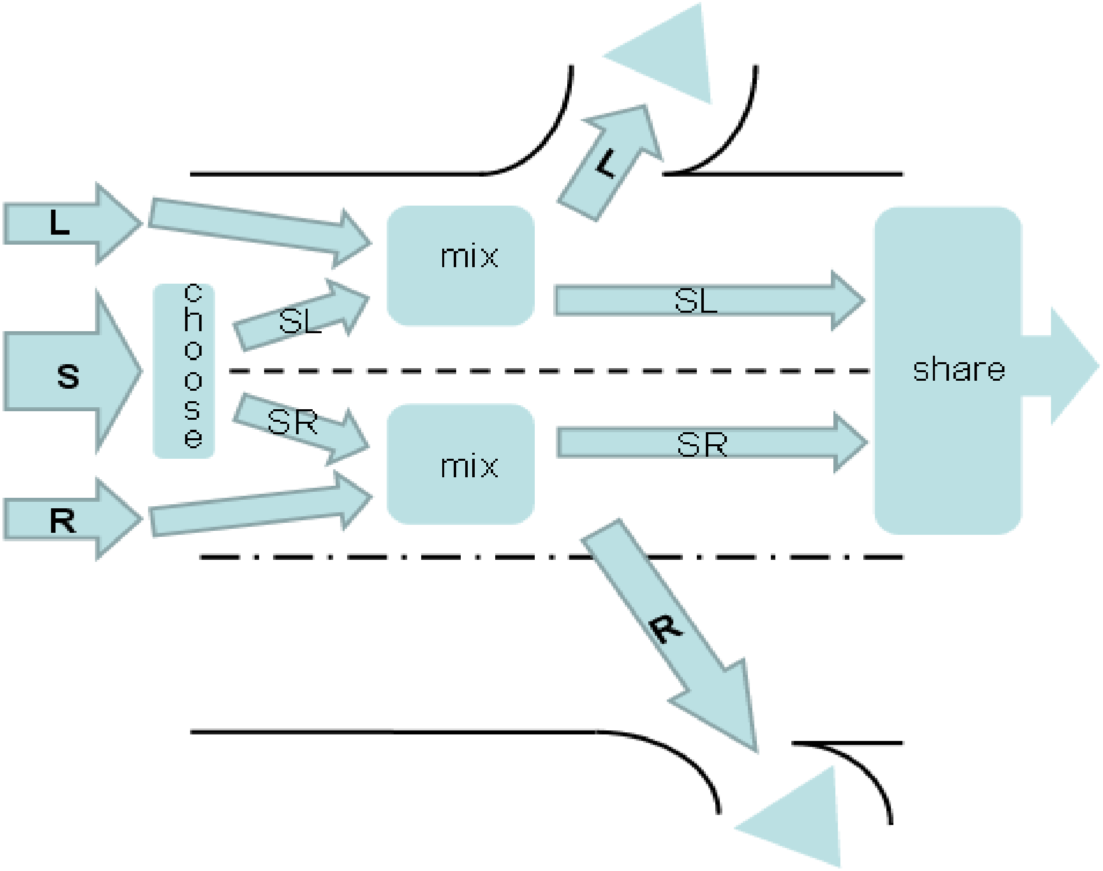


Figure 6.4.1 Two lanes with turning movements and shared service

[59] The layout in the diagram reflects left-side (UK) driving convention but this does not affect the analysis.

In Table 6.4.1 the states of lane queues are divided into empty and non-empty, with the utilisation of each lane being defined as the proportion of *time* for which it is non-empty[60]. Again the convention of lower-case p for elementary probabilities and upper-case P for symmetrical single–lane probabilities is used:

Table 6.4.1 Grouping of elementary probabilities for two lanes with shared service

	Lane 2 queue = 0	**Lane 2 queue > 0**
Lane 1 queue = 0	P_0	
	p_{00}	$\Sigma_1^{\infty} p_{0j} = P_0 - p_{00}$
Lane 1 queue > 0	$\Sigma_1^{\infty} p_{i0} = P_0 - p_{00}$	$\{p_{ij,}\, i,j>0\}$
	$1\text{-}P_0$	

If each lane can use all the capacity available to the other when it is empty, then (also referring back to Figure 6.2.5) the effective factor of gain in capacity for each lane is:

$$f_{u(2)} = \frac{1 - P_0 + P_0 - p_{00}}{1 - P_0} = \frac{1 - p_{00}}{1 - P_0} \qquad (6.4.1)$$

In practical terms, the throughput capacity on a lane doubles during some random time intervals, as in Figure 6.2.5, and the distribution of service headways is therefore no longer perfectly exponential. Therefore, in terms of time, a smaller proportion of total time is needed to service the queue in each lane, so the effective utilisation in terms of time is *reduced*. A similar calculation for three lanes gives the results in Table 6.4.2.

Table 6.4.2 Utilisation construction for three lanes with shared service

Case	**Absolute probability**	**Utilisation factor**
All lanes zero	$p_{000} = 1 - \rho$	0
Both other lanes zero	$\Sigma p_{00i} = P_{00} - p_{000}$	3 (contributes additional 2x)
One other lane zero	$P_0 - 2(\mathrm{P}_{00} - p_{000}) - p_{000}$	2 (contributes additional 1x)
Neither other lane zero	*remainder*	1
Lane 1 non-zero	$1\text{-}\Sigma p_{0ij} = 1\text{-}P_0$	Includes the three cases above

The effective utilisation with three lanes is then:

$$f_{u(3)} = \frac{1 - P_0 + \left(P_0 - 2\left(P_{00} - p_{000}\right) - p_{000}\right) + 2\left(P_{00} - p_{000}\right)}{1 - P_0} = \frac{1 - p_{000}}{1 - P_0} \qquad (6.4.2)$$

[60] This relies on the assumption that the capacity is constant over the measured period.

For any number of lanes *n*, the probability of the total queue being zero $p_{0\ldots0}$ is the same, namely (1-ρ) where ρ is the demand intensity on the whole system. In time-dependent cases ρ should be replaced by the combined utilisation of service *x*, but for the moment only the steady state is considered. To calculate $f_u^{(n)}$ requires an estimate of P_0, the probability that an individual lane is empty. To this end, a 'transfer probability' $P_x^{(n)}$ is first defined by:

$$P_{x(n)} = \frac{P_0 - p_{0..0}}{(n-1)(1-P_0)} \tag{6.4.3}$$

The factor $f_{u(n)}$ can be built inductively as a Binomial expansion (6.4.4), whose components represent the separate contributions to one lane's utilisation when no other lane is empty, one other lane is empty, two are empty etc, up to all other lanes empty. Each contribution has a factor one more than the number of empty lanes: one empty lane, with a certain probability, doubles the utilisation available to the current lane, two empty lanes triple it (if 'cutting up' the middle lane is allowed!) etc, and each case can occur in the combinatorial number of ways:

$$f_{u(n)} = (1-P_x)^{n-1} + 2\binom{n-1}{1}P_x(1-P_x)^{n-2} + \ldots + nP_x^{n-1} \tag{6.4.4}$$

Equation (6.4.4) simplifies naturally to:

$$f_{u(n)} = (1-P_x) + nP_x \equiv 1 + (n-1)P_x \tag{6.4.5}$$

Even though most of the terms in the Binomial formulation cancel out, it has explanatory value because it represents the contributions as mutually independent, so avoiding the need for an explicit representation of correlation between lanes. By symmetry, the mean steady-state queue in each of *n* lanes has to be:

$$L_{e(n)} = \frac{\rho}{n(1-\rho)} \tag{6.4.6}$$

Treating each lane queue as quasi-M/M/1, this implies the *effective* ρ in each lane is given by (6.4.7), matching (6.2.4), but the last equality also follows from the symmetry assumption since in the steady state the utilisation must equal the effective demand intensity:

$$\rho' = \frac{L_{e(n)}}{L_{e(n)}+1} = \frac{\rho}{\rho + n(1-\rho)} = \frac{\rho}{f_{u(n)}} \tag{6.4.7}$$

If, again assuming quasi-M/M/1, it is supposed that approximately:

$$P_0 \approx 1-\rho^{/} \tag{6.4.8}$$

equations (6.4.5) and (6.4.7) give:

$$P_x \approx 1-\rho \tag{6.4.9}$$

Equation (6.4.9) fixes what was previously called the transfer probability. In this approach the transfer contribution from an empty lane is not only independent of the other lanes but is the same as the nominal probability of the lane being empty as if it were isolated. This is convenient because it can potentially be extended to more general cases, using conventional methods. The linearity of (6.4.5) and (6.4.7) suggests that if only a proportion ϕ of an empty lane's capacity is sharable then P_x can simply be reduced by this factor. This will be convenient where a degree of effective sharing arises from weaving between lanes or after departure.

On a *time* basis,it is possible to write the probabilities that when one lane is ready for service the other lane is empty or otherwise not ready, $P_{x0}^{(t)}$, and its complement where the other lane is ready for service, $P_{xx}^{(t)}$. For two lanes:

$$P_{x0}^{(t)} = \phi\left(\frac{P_0 - p_{00}}{1-P_0}\right), \qquad P_{xx}^{(t)} = \frac{1-(1+\phi)P_0 + \phi p_{00}}{1-P_0} \tag{6.4.10}$$

The probabilities that a customer will be served are given by equations (6.4.11), which differ from (6.4.10) because twice as many can be served on one lane while the other lane is not being served. The significance of this distinction is that the probabilities (6.4.11) can be measured directly in simulations by counting the numbers of customers served under different conditions, allowing models to be tested, and possibly the factor ϕ to be calibrated. Of course, the formulae will be more complicated if there is not symmetry between lanes.

$$P_{x0}^{(s)} = \frac{2p_{x0}^{(t)}}{1+p_{xo}^{(t)}} = \frac{2\phi(P_0 - p_{00})}{1-(1-\phi)P_0 - \phi p_{00}}, \qquad P_{xx}^{(s)} = \frac{1-(1+\phi)P_0 + \phi p_{00}}{1-(1-\phi)P_0 - \phi p_{00}} \tag{6.4.11}$$

For the basic two lane case in (Figure 6.2.1), using (6.4.5) and (6.4.6):, the effective and actual demand intensities for straight movements are related by:

$$\rho'_{SL} = \frac{\rho_{SL}}{1+\phi(1-\rho_{SR})}, \qquad \rho'_{SR} = \frac{\rho_{SR}}{1+\phi(1-\rho_{SL})} \tag{6.4.12}$$

These equations allow service sharing to be accounted for approximately when the arrivals and possibly the underlying service are not necessarily symmetrical between the lanes. In the time-dependent case it is assumed that traffic intensities ρ can be replaced by utilisations.

6.4.3 Space sharing in a single lane

Before proceeding to consider turning movements, it is desirable to show that the method is internally consistent where a lane carries two independent movements. This could be called the 'red queue, blue queue' problem, meaning that if arrivals are arbitrarily painted in two colours it should not affect the result. Suppose that total arrivals are λ, capacity is μ, the 'red' and 'blue' component arrival rates are λ_r and λ_b, and u_r, u_b are the corresponding utilisations relative to the capacity. Then since each 'colour' takes capacity away from the other, the partial capacities are:

$$\mu_r^* = \mu_s(1-u_b),\ \mu_b^* = \mu_s(1-u_r) \qquad \text{where } u_r = \frac{\lambda_r}{\mu_s},\ u_b = \frac{\lambda_b}{\mu_s} \qquad (6.4.13)$$

Now define the effective utilisations simply as follows:

$$u_r^* = \frac{\lambda_r}{\mu_r^*} = \frac{\lambda_r}{(\mu_s - \lambda_b)}, \qquad u_b^* = \frac{\lambda_b}{\mu_b^*} = \frac{\lambda_b}{(\mu_s - \lambda_r)} \qquad (6.4.14)$$

The mean quasi-steady-state queues according to the standard M/M/1 formula are:

$$L_r^* = \frac{u_r^*}{(1-u_r^*)} = \frac{\lambda_r}{(\mu_s - \lambda_r - \lambda_b)}, \qquad L_b^* = \frac{u_b^*}{(1-u_b^*)} = \frac{\lambda_b}{(\mu_s - \lambda_r - \lambda_b)} \qquad (6.4.15)$$

Since these queues are non-overlapping the total queue is just their sum:

$$L^* = \frac{(\lambda_r + \lambda_b)}{(\mu_s - \lambda_r - \lambda_b)} \qquad (6.4.16)$$

But since the total volume λ= λ_r + λ_b, in the steady state this is just the same as:

$$L^* = \frac{\lambda}{(\mu - \lambda)} = \frac{\rho}{(1-\rho)} \equiv L_e \qquad (6.4.17)$$

This result is independent of the ‘red/blue’ split, showing that the method is internally consistent. This will not necessarily work for queue processes other than M/M/1, if the presence of one queue modifies the headway distributions of the other.

6.4.4 Averaging capacities of streams sharing a lane

Suppose the traffic population ‘colours’ *are* distinguishable, for example because their capacities differ. How should the capacities then be combined?

Suppose the elementary capacities are μ_r and μ_b, being the capacities each population would experience if the other were absent. These could be the same, or different if for example each type of customer requires a different amount of capacity. The conventional answer is to use the harmonic mean, as this gives the mean service time. However, this neglects the relative proportions of the colours in the arriving flow, as reflected in the respective utilisations.

Remembering that utilisation has been defined as the proportion of *time* that service takes place on a movement, and that the movements are mutually exclusive, at any point in a given time period a customer can be using either the ‘red’ mode or the ‘blue’ mode, or no service is taking place. The combined utilisation is therefore the *sum* of the component utilisations. Therefore, it is reasonable to define the effective combined capacity as the total throughput divided by the sum of the component utilisations:

$$\bar{\mu} = \frac{u_r\mu_r + u_b\mu_b}{u_r + u_b} \tag{6.4.18}$$

Since the throughputs equal arrivals in the steady-state, $\lambda_x = u_x\mu_x$ (compare equation (6.4.14)) and service times t_x are the inverses of capacities, equation (6.4.18) is equivalent to:

$$\bar{t} = \frac{1}{\bar{\mu}} = \frac{\frac{\lambda_r}{\mu_r} + \frac{\lambda_b}{\mu_b}}{\lambda_r + \lambda_b} = \frac{\lambda_r t_r + \lambda_b t_b}{\lambda_r + \lambda_b} \tag{6.4.19}$$

The average capacity is therefore the harmonic mean of the component capacities weighted by flows. This is also the definition used by Knote (2006) based on the German *Handbuch für die Bemessung von Strassenverkehrsanlagen* (handbook for the measurement of road traffic installations, equivalent to the US Highway Capacity Manual), which he then finds gives accurate results when applied in simulation of a signalised junction.

Contrary to what might appear at first sight, the average capacity in (6.4.18) is not dominated by the larger of the component capacities. Suppose for example that the ‘red’ capacity approaches zero while the ‘blue’ capacity stays finite then, remembering that the sum of the utilisations cannot exceed 1, if there is any ‘red’ demand at all its utilisation must approach 1 and the ‘blue’ utilisation is therefore forced towards zero. The combined capacity then approaches the ‘red’ capacity, zero. The conclusion in this extreme case is then the same as if the combined capacity were the unweighted harmonic mean of the component capacities.

6.4.5 Effective capacity on each movement

Effective capacities for all the movements can be calculated using the shared service and shared lane formulae (6.4.12) and (6.4.13), treating straight-left-lane (SL) and straight-right-lane (SR) as separate movements, the latter involving mutual dependence on actual traffic flows through utilisations, the proportions of *time* spent serving a movement. The sharing factor ϕ is omitted as it would complicate formulae while adding nothing to the argument, but it can be reintroduced quite straightforwardly into numerical calculations.

The effective capacities of the movements (asterisked) can be expressed in terms of either absolute (unprimed) or effective (primed) utilisations, assuming that each lane claims half of the total straight-movement capacity μ_s, though not necessarily half the throughput. Each effective capacity is determined by the proportion of time a lane is not being used by another movement. For example, the capacity for straight ahead movements on the left lane, μ^*_{SL} , is the basic left lane capacity, $.5\mu_S$ factored by the proportion of time the left turn movement is *not* running $(1-u_L)$, enhanced by the proportion of time straight ahead service is not being used by the right lane $(2-u_{SR})$, with the 2 representing that the straight capacity has been halved. The effective capacity of the left turn μ^*_L is factored by the proportion of the effective capacity available to the straight ahead movement in the left lane not used by straight movements in the right lane. The other two effective capacities are mirror images of the above. The effective turn capacities define corresponding *effective* straight ahead utilisations, the proportions of time that the lane is *not* serving turners. The alternative expression for effective capacities follow:

$$\mu^*_L = \mu_L\left(\frac{2-u_{SL}-u_{SR}}{2-u_{SR}}\right) \equiv \mu_L\left(1-u'_{SL}\right) \qquad (6.4.20)$$

$$\mu^*_{SL} = .5\mu_S\left(1-u_L\right)\frac{u_{SL}}{u'_{SL}} = .5\mu_S\left(1-u_L\right)\left(2-u_{SR}\right) = \frac{\mu_S\left(1-u_L\right)\left(1-u'_{SR}\right)}{\left(1-u'_{SL}u'_{SR}\right)} \qquad (6.4.21)$$

$$\mu_{SR}^{*} = .5\mu_S\left(1-u_R\right)\frac{u_{SR}}{u_{SR}^{\prime}} = .5\mu_S\left(1-u_R\right)\left(2-u_{SL}\right) = \frac{\mu_S\left(1-u_R\right)\left(1-u_{SL}^{\prime}\right)}{\left(1-u_{SL}^{\prime}u_{SR}^{\prime}\right)} \qquad (6.4.22)$$

$$\mu_R^{*} = \mu_R\left(\frac{2-u_{SL}-u_{SR}}{2-u_{SL}}\right) \equiv \mu_R\left(1-u_{SR}^{\prime}\right) \qquad (6.4.23)$$

Then assuming quasi-M/M/1 processes, the component queues can be calculated from the effective traffic intensities of the component services:

$$\rho_i^{*} = \frac{\lambda_i}{\mu_i^{*}}, \qquad (6.4.24)$$

$$L_i = \frac{\rho_i^{*}}{1-\rho_i^{*}} \qquad \text{(in the steady state)} \qquad (6.4.25)$$

The component queues in each lane are then added get the total queues, in accordance with (6.4.6) and (6.4.7). The formulae then start to get complicated, so the optimal way to proceed is by numerical calculation.

Referring back to the earlier comment about signal junctions requiring multi-lane service sharing treatment, it appears straightforward to extend the preceding simplified approach to more general types of equilibrium queue such as M/D/1 just by replacing the M/M/1 formulae as in (6.4.6) etc. The usefulness of this could be tested by Monte Carlo simulation, because of the complexity of defining recurrence relations compared to an M/M/1 ensemble. The degree to which individual lane queues tend to share the broad statistical properties of the ensemble could be investigated, and whether their statistical properties are relatively insensitive to number of lanes above 2. These could be topics for further research.

6.5. EXTENSION TO TIME-DEPENDENCE

6.5.1 Time-dependent simulation with turning movements

Time-dependence is important not just when service is oversaturated, and where the demand intensity is enough that relaxation times are long compared to the modelled time period, but also when the system as a whole is within capacity, if interaction between movements reduces the effective capacity of one or more movements sufficiently to oversaturate them. To provide a benchmark for analytical methods an event-based simulation has been developed. Figures 6.5.1 shows the results of a simulation of 1100 units/hour demand impinging on a system with total capacity of 1750 units/hour summed over all movements. The graphs represent the average of nine simulations each running to a simulated duration of about 7 hours.

Event-based simulation has the advantage of revealing the variability of queue sizes as well as avoiding an explosion in the dimensions of the probability distributions, though of course it can be difficult to get reliable averages. In Figure 6.5.2, individual simulation runs exhibit much greater variability than the averaged results. Figure 6.5.3 confirms that the statistics of the exponential event generator are accurate. An apparent slowing of the queue growth rates in the early growth part in Figure 6.5.1, and in Figure 6.5.2, is however illusory. Undersaturated cases have the same ragged appearance but without the secular growth trend, in fact variability is expected to be maximised around saturation (Taylor 2005a).

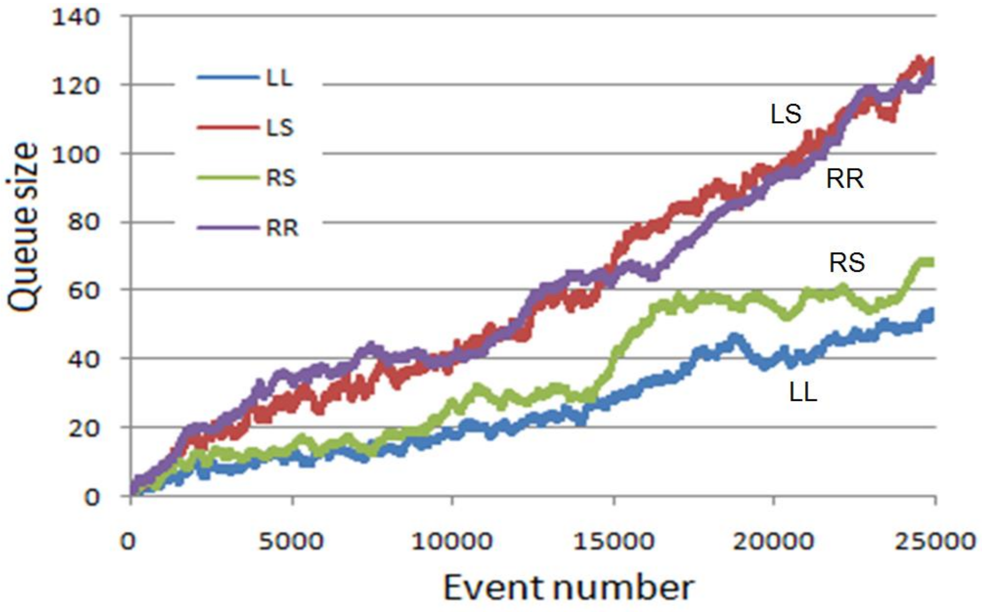


Figure 6.5.1 Averaged oversaturated queues from event simulations on four movements

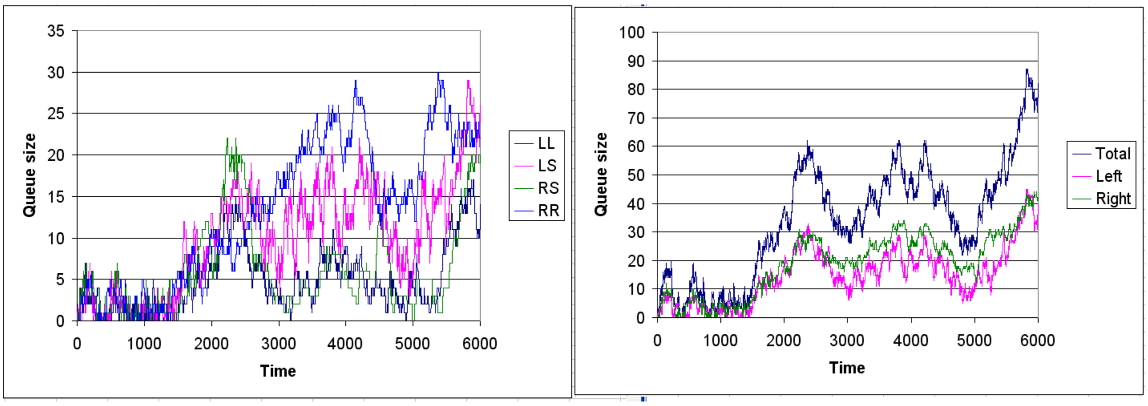


Figure 6.5.2 Detail of one of the event simulations for the above case showing variability

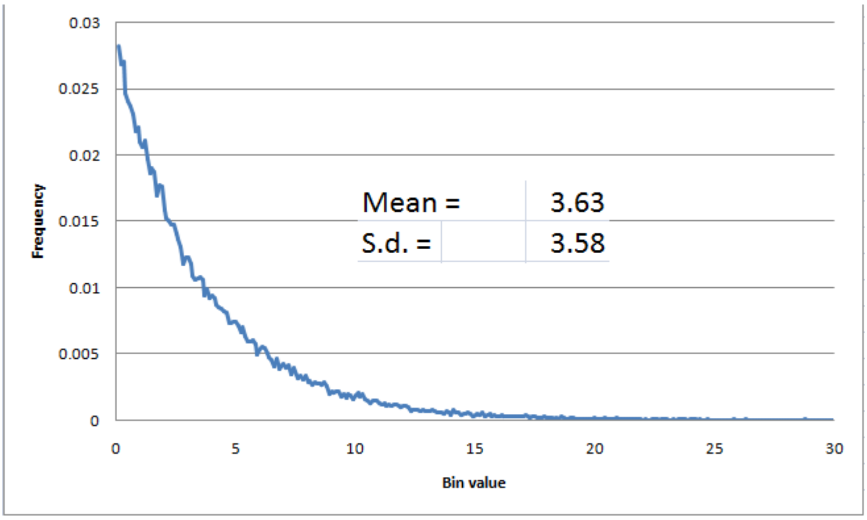


Figure 6.5.3 Verification of exponential generator in simulation (100,000 events)

6.5.2 Time-dependent estimation using the shearing method

In the time-dependent case, demand intensities of arrivals no longer equal utilisations at the service points. The demand intensities are known from data, but it is the utilisations that determine the queue sizes. In the case of a single queue this is modeled by a kind of feedback between the time-dependent deterministic queue formula and the Pollaczek-Khinchin formula, leading to a quadratic solution.

As described by Kimber and Hollis (1979) and elsewhere, shearing transforms the Pollaczek-Khinchin steady-state mean queue formula to be asymptotic to the deterministic queue formula to create a closed-form time-dependent description that handles the transition between under and oversaturation seamlessly (also see earlier in Part B). It can be interpreted functionally as treating the dynamic queue as approximately quasi-static, with the degree of saturation at the service replacing the demand intensity (Taylor 2003). The simplest formulation, where L_0 is the initial queue and C represents the statistical term $(1+c_b^2)$, has an additional parameter a to allow choice between calculating the final queue at time t (a=1), and the average queue over [0,t] (a=2), where these values rely on the fact that with constant parameters the average queue over time is approximately equal to the instantaneous queue at half time (Kimber and Hollis 1979):

$$L(t) \equiv L_0 + (\rho - x)\frac{\mu t}{a} = Ix + \frac{(I_a - 1)x}{2(1-x)} + \frac{Cx^2}{1-x} \qquad (6.5.1)$$

Average values are needed when adjusting average traffic intensities for feasibility and also if the chosen benchmark is of relative stable time-averaged rather than highly variable final simulated queues.

The solution of the modified problem (6.5.1) for the degree of saturation x is:

$$x(t) = \frac{g - \sqrt{g^2 - 4fh}}{2f} \quad (f \neq 0) \qquad , \; x(t) = \frac{h}{g} \quad (f = 0 \text{ and } g \neq 0) \qquad \text{where} \quad (6.5.2)$$

$$f = \frac{\mu t}{a} - (C - I), \quad g = L_0 + I^* + (\rho + 1)\frac{\mu t}{a}, \quad h = L_0 + \rho\frac{\mu t}{a} \qquad (6.5.3)$$

For M/M/1, where I^* =I = C = 1, and L_0=0, the queue is now calculated in three steps:

$$[g/2f] = \tfrac{1}{2}\left(1 + \rho + \frac{a}{\mu t}\right), \quad x = [g/2f] - \sqrt{[g/2f]^2 - \rho}\,, \quad L = \frac{x}{1-x} \qquad (6.5.4)$$

6.5.3 Practical algorithm for time-dependent lane queues with turning movements

Interacting movements lead to complications that appear soluble only iteratively, particularly as it is impractical to solve for an explicit optimal proportion γ of straight-moving arrivals that should use each lane according to whatever selection policy is adopted. Indeed, if component queues fail to grow in step this proportion could also vary with time. However, for the moment the concern is only to show it is possible to arrive at an internally consistent solution.

The algorithm adopted is the following:

1. Given the arrival (demand) rates on each movement, or total arrivals and turning proportions, and basic capacities for each movement, calculate ‘raw’ average utilisations (degrees of saturation) on each movement using the sheared queue method. If steady state is assured, x=ρ, so this step can be omitted.

2. Calculate initial values of the effective utilisations on movements affected by shared service. This affects only the SL and SR movements in Figure 6.4.1.

3. If the sum of effective utilisations on a lane exceeds 1 then they are infeasible and must be adjusted. This can be done by factoring the utilisations so their sum does not exceed the single utilisation derived from the total demand and average capacity on the lane, the latter being estimated from the basic capacities and raw effective utilisations according to (6.4.18), and the average utilisation in the lane estimated using the sheared approximation with a=2.

4. Using adjusted effective utilisations, calculate effective capacities and demand intensities according to equations (6.4.20-25), then use the sheared method to calculate either the time-averaged or final component queues.

This algorithm is somewhat circular, but solving analytically for feasible, mutually consistent component utilisations appears intractable, and it is felt that a stepwise analytical approach is more transparent than numerical solution. Although up to three evaluations of the sheared queue formula per lane are needed rather than just one, the sheared solution is quite efficient.

However, results are very sensitive to the value of the left-right split factor γ for the straight movement, which should normally be adjusted to achieve some pre-determined condition like equalising lane queues (if feasible). In test cases up to five iterations of the entire algorithm have been needed to get this value. The result appears insensitive to the starting estimate of γ, which therefore may as well be 0.5. In each iteration the factor can be adjusted using a descent direction calculated using derivatives of queue size with respect to γ. Based on a spreadsheet program, the main algorithm needs 52 function evaluations, and the descent direction adds 34. So five outer iterations multiply the basic algorithm work by a factor of about 8, possibly leaving room for some practical improvement.

6.5.4 Results and discussion

The method and algorithm described have been implemented in a spreadsheet and compared with event simulation. The test cases, defined in Table 6.5.1, are not intended to reflect typical realistic situations. In Table 6.5.1 and Figure 6.5.4 the estimation has been provided with the outturn simulated arrival rates and capacities rather than the specified values. This is arguably a fair test as results can be very sensitive to the inputs. ‘Reconstruction’ is a simplified application of the method based on outturn statistics such as equations (6.4.11), and therefore gives an indication of how well the method and event simulation correspond at a basic level without, for example, re-adjusting the straight-movement split factor γ.

Table 6.5.1 Results using specified arrival and capacity rates, with γ estimated iteratively

Movement	**Left**	**Straight**	**Right**
Capacities	500	1000	250
Sub-Case	Turning proportions		
A	0	1	0
B	0.2	0.6	0.2
C	0.35	0.5	0.15

Case	**Total demand**	**Event Simulation**				**Reconstruction**		**Modelled from simulation data**[61]		
	Q	Hours	γ	L_L	L_R	L_L	L_R	L_L	L_R	γ
80A	800	19.571	0.555	2.1	2	2.9	1.9	2	2	0.5
80B	800	19.561	0.708	2	3.4	2.5	5	3.2	3.2	0.853
80C	800	19.543	0.487	2.7	2.4	3.9	2.7	3.4	3.4	0.395
95A	950	18.506	0.512	10.1	10.1	12.3	9.6	9.2	9.2	0.5
95B	950	18.49	0.757	5.4	8.9	6.8	18.1	9.9	10	0.856
95C	950	18.49	0.437	7.8	6.9	12.6	8.9	13.2	13.3	0.396
99A	990	18.174	0.502	35.6	35.6	52.1	47.2	38	38	0.5
99B	990	18.164	0.788	9	13.8	12.1	33.5	18	18.1	0.865
99C	990	18.203	0.421	13.4	12.2	23.3	14.2	28.2	28.4	0.381
110A	1100	1.726	0.504	44.4	44.5	47.8	43.3	40	40	0.5
110B	1100	1.732	0.803	27.3	34.1	28.4	61.8	34.2	34.3	0.827
110C	1100	1.69	0.395	37.4	35.2	45.8	41.1	48.3	48.4	0.371

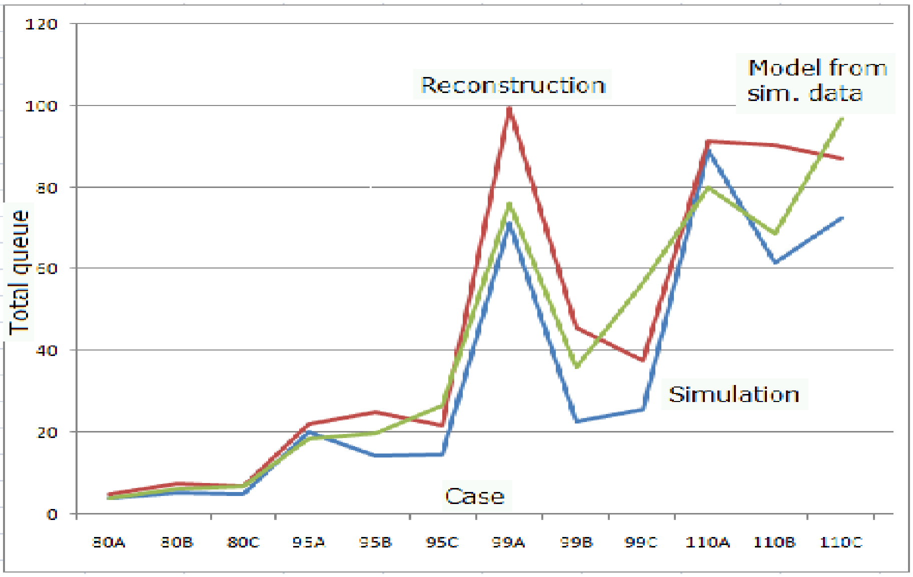


Figure 6.5.4 Total queues in test cases simulated and calculated from simulation data

[61] The original simulated arrival and capacity values are used, but the left-right balance γ has been hand-adjusted.

In Figure 6.5.5, the method is working with the specified data, which can be a few percent different from those actually simulated. There is some error in the estimated queue components relative to event simulation, and in particular the method appears to overestimate turning queues. Figure 6.5.6 plots the standard deviations of the component queues between the nine simulation runs with different random seeds, giving an idea of their inherent uncertainty which is around 30% of the average queue values and is of similar order to the estimation error, suggesting the latter may be in part irreducible.

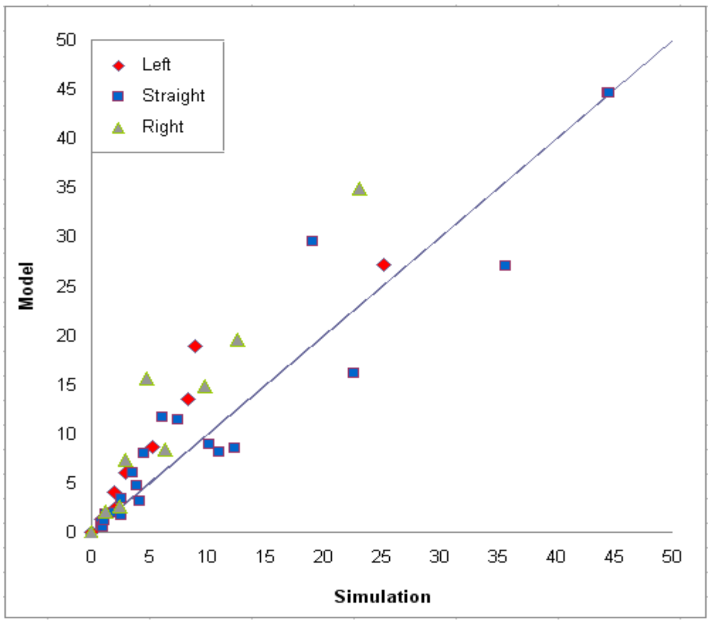


Figure 6.5.5 Test cases calculated from specification v. simulated component queues

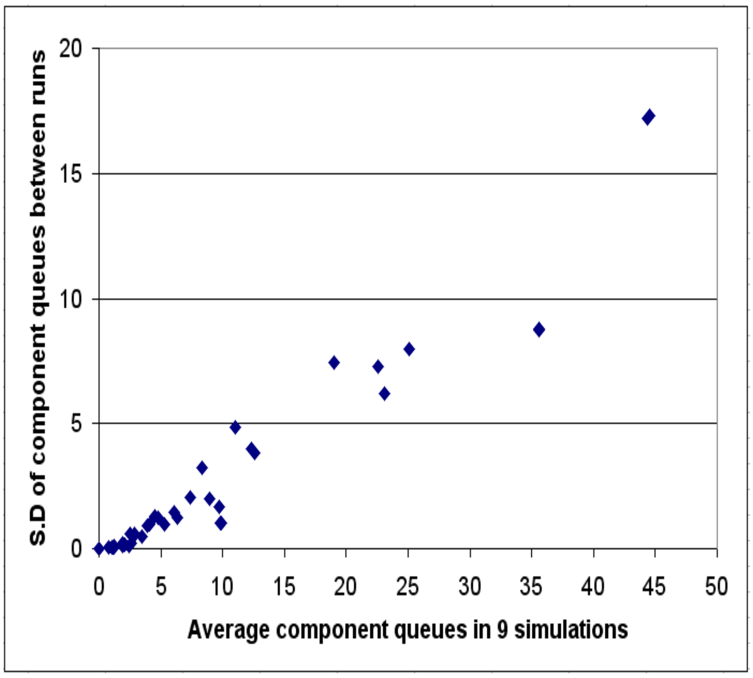


Figure 6.5.6 Standard deviations of simulated component queues between runs

Two particular questions arise concerning realism and possible further investigation.

First, the degree of service sharing ϕ can have a dramatic effect on queue sizes, and the implementation of the method allows it to be specified. What values would be realistic is a matter for further theoretical and possibly empirical study. For two lanes it may be possible to ascribe values to particular cases. For more than two lanes, a single value may not be meaningful, but if appropriate values between lane pairs can be established then the method of adding up independent contributions summarised by equation (6.4.4) should be able to cope.

Second, it has been found in microscopic simulations that even if straight-moving arrivals select the shorter queue at their moment of arrival, lane queues may still end up unequal, the longer queue tending to occur on the lane where the turning flow is most dominant[62]. A possible explanation is that since lane selection affects only straight-movers, as a queue fluctuates they will tend to join it only when it is short, thereby tending to oppose downward fluctuations, but turners, having no choice, will not be put off when the queue is long, so tending to preserve upward fluctuations. Thus, queues with more turners will tend to increase. This effect will be masked by any algorithm that actively *forces* equality on the lane queues.

[62] Personal communication by Dr Helen Gibson, TRL

6.6. CONCLUSIONS ON QUEUING ON MULTIPLE LANES

This Chapter 6 has been focused on the particular topic of multi-lane queues, which they arise frequently in practical urban networks, but it is believed have not previously been treated according to the shared service principle. The approach follows that of the preceding Chapters to the extent of trying to describe the lane queues in terms to which the time-dependent and Pollaczek-Khinchin methods can be applied.

This has been done by developing formulae for the effective capacities of movements, where each movement can have a different exit capacity, and by defining utilisations in terms of the proportion of time available to each movement. This has been formulated in detail for two lanes with turning movements. For any number of lanes, a simple description in terms of service sharing has been proposed.

The results have been derived for equilibrium queues, but are extendable to time-dependence and oversaturation through the sheared method because utilisations always lead to feasible results in the time-dependent queue formulae, subject to consistency checks on the utilisations themselves. The results agree fairly well with event-based simulations, bearing in mind the high variability evident in the latter. This Chapter has therefore extended the scope of the work in a direction that is practically relevant. It has also raised though not necessarily resolved issues that may be of theoretical interest, in particular the role of correlation between lane service processes, and how this can be handled in simple structural terms. These may provide topics for future research.

CHAPTER 7: COMPUTATIONAL ISSUES AND DEMONSTRATOR

7.1. INTRODUCTION

Practical traffic modelling applications generally require efficient computational methods because of the magnitude of problems addressed and the need to test and compare multiple options and to perform sensitivity analysis. What would be appropriate to different applications, and how it may differ from existing computational methods, is considered. Demonstrator software is described together with illustrations of results.

7.2. FIELD OF APPLICATION OF THE METHODS

The specific aim of this work is to predict the evolution of queues and delays for arbitrary traffic profiles of demand and capacity with useful accuracy. As far as can be ascertained from publications, macroscopic modelling currently uses four different methods:

- Analytical static delay-volume functions, as used in traffic modelling suite SATURN (2012). In SATURN delay is estimated by a power function of the ratio of volume to capacity, and demand in the inner simulation is constrained not to exceed capacity.

- Empirical speed-flow relationships, as used in COBA (2012), which also apply only in a static context with demand below capacity. An extension to allow for overcapacity was proposed in 1971 but not generally applied (see Taylor *et al* 2008).

- Time-dependent queuing, as described in this dissertation and used in the CONTRAM dynamic assignment suite (Taylor 2003) and TRL's junction modelling software.

- The Cell Transmission Method (CTM), which models queuing deterministically by considering inflows to and outflows from small network elements (cells) that are chained together to represent traffic streams (TRLSoftware 2012). Similar input-output flow-conservation methods are described in many published papers. A continuous variation of this approach that is widely studied but less used because of its computational difficulty is the Lighthill-Whitham-Richards (LWR) kinematic method, originated by Sir James Lighthill and Gerald Whitham (1955).

What methods are used in practice can be difficult to tell. Since traffic modelling software first appeared in the 1970s, it has become increasingly commercial and proprietary and tends to be sold on features and appearance rather than methodology, as a result of which the inner workings have become increasingly obscure.

The enhancements to queue modelling described here lend themselves mostly to time-dependent macro/mesoscopic tools like CONTRAM for assignment or for the design and modelling of junctions. The random queue calculation is a software module whose inputs are initial queue, demand, capacity, duration of period, and statistical parameters, and whose outputs are final queue size and average utilisation. An enhanced module will have additional inputs and outputs, the initial and final values of p_0 and variance. All other new variables and calculations can be internal to the module. The computational burden for each queue in each time period is likely to be increased substantially compared to mean queue calculation[63], so an increase in run time is also to be expected[64]. However, one would expect a parallel investment in more efficient assignment methods. Note that dropping the variance calculation is not an option if the improved accuracy of mean queue estimation that it enables is to be achieved. Estimation of the queue size probability distribution is likely to be optional because of its relatively high computational effort, but its inputs are simple: demand intensity (demand/capacity), statistical parameters, current values of p_0, mean queue size and variance.

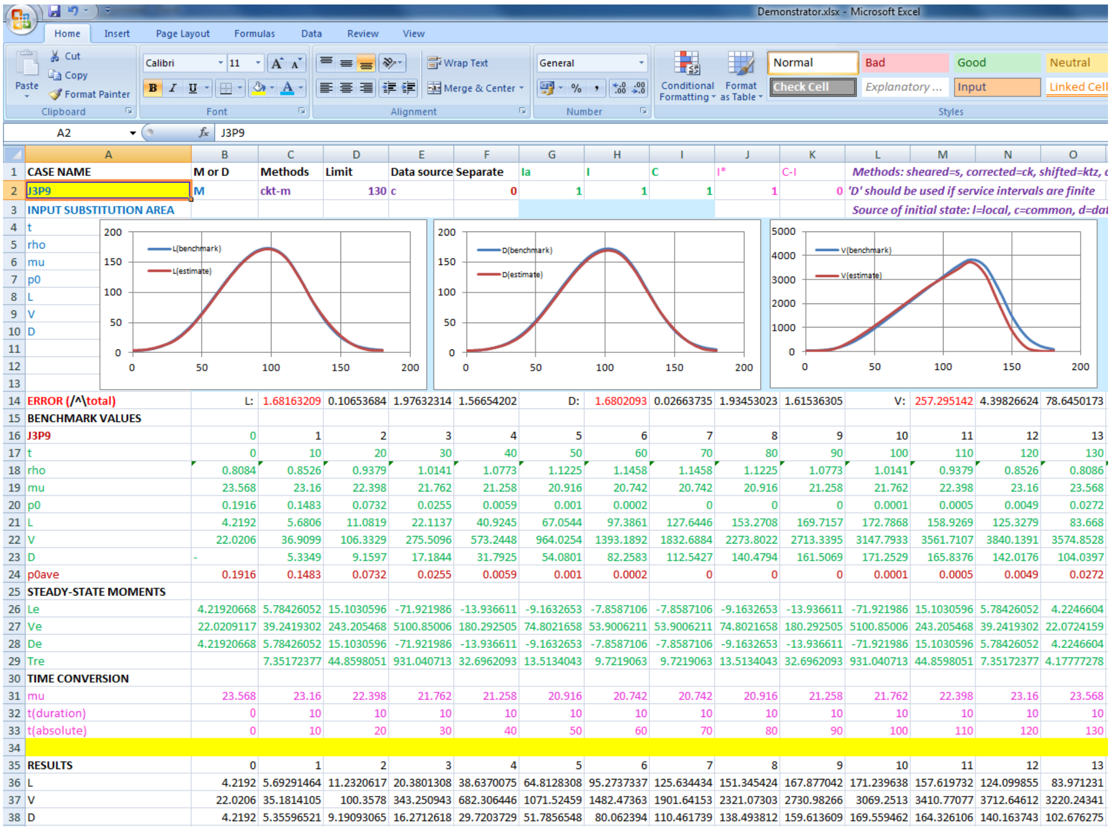
7.3. DEMONSTRATOR

Developing a computationally-efficient module or program to perform the calculations defined in this dissertation is considered to be outside the scope of the work. The reasons are similar to those put forward for not pursuing an independent solution algorithm for fitting queue size distributions. Instead a demonstration program has been constructed enabling scenarios to be tested and results examined. Use of a spreadsheet, while not necessarily computationally efficient, provides a transparent vehicle for computations as well as convenient access to the methods and intermediate results, not normally available from modular software. Comparison of estimation errors across all peak cases required the code to be replicated in a separate worksheet for each case along with its own case data and Markov similulated results, one set for M/M/1 and one set for M/D/1. The Demonstrator 'Case' spreadsheet has one calculating worksheet, which refers by lookup to either the M/M/1 or M/D/1 case database. Formulae are arranged vertically with each time slice occupying one column, this being more convenient for testing. In the transposed 'Demo' worksheet the calculations for each time slice occupy one row, which may be more convenient where the number of time slices is large.

[63] The Demonstrator calculation uses about 70 lines of Excel code compared to 20 for the sheared approximation.

[64] In CONTRAM as much as 80% of run time is devoted to the calculation of link and junction delays.

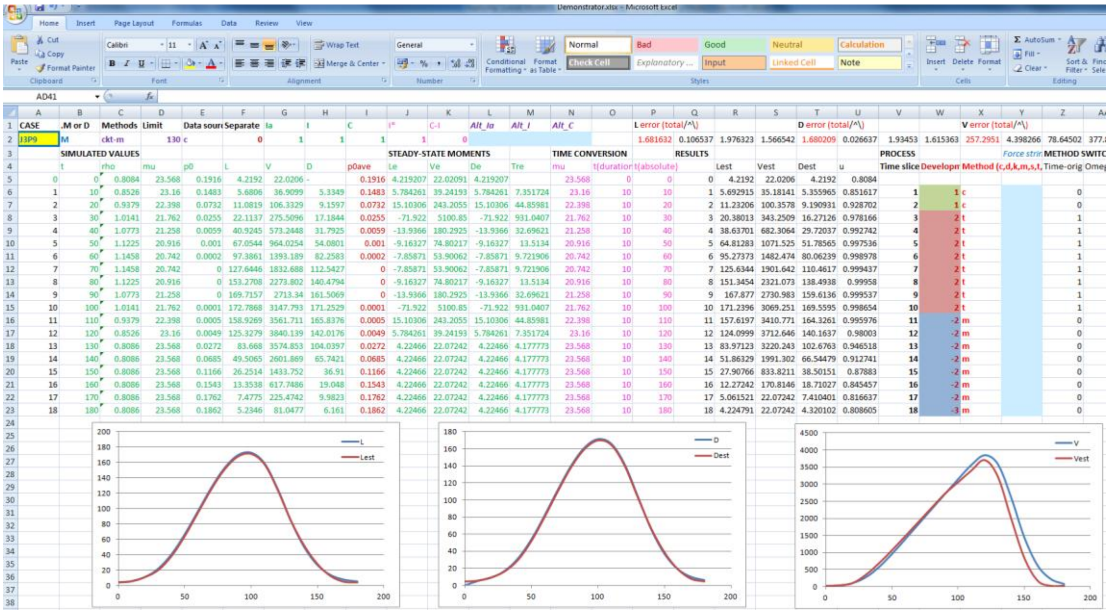
A sample of the ‘Case’ worksheet is shown in Figure 7.3.1. The green fields are data and steady-state values, and the black fields at the bottom are results. These are compared in the graphs of L, D and V respectively, below which errors are printed in red. The transposed and slightly rearranged ‘Demo’ version is illustrated in Figure 7.3.2.



A2 J3P9

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	CASE NAME	M or D	Methods	Limit	Data source	Separate	Ia	I	C	I*	C-I	Methods: sheared=s, corrected=ck, shifted=ktz, c			
2	J3P9	M	ckt-m	130	c	0	1	1	1	1	0	'D' should be used if service intervals are finite			
3	INPUT SUBSTITUTION AREA											Source of initial state: l=local, c=common, d=dat			
4	t														
5	rho														
6	mu														
7	p0														
8	L														
9	V														
10	D														
11															
12															
13															
14	ERROR (/^\total)	L:	1.68163209	0.10653684	1.97632314	1.56654202	D:	1.6802093	0.02663735	1.93453023	1.61536305	V:	257.295142	4.39826624	78.6450173
15	BENCHMARK VALUES														
16	J3P9	0	1	2	3	4	5	6	7	8	9	10	11	12	13
17	t	0	10	20	30	40	50	60	70	80	90	100	110	120	130
18	rho	0.8084	0.8526	0.9379	1.0141	1.0773	1.1225	1.1458	1.1458	1.1225	1.0773	1.0141	0.9379	0.8526	0.8086
19	mu	23.568	23.16	22.398	21.762	21.258	20.916	20.742	20.742	20.916	21.258	21.762	22.398	23.16	23.568
20	p0	0.1916	0.1483	0.0732	0.0255	0.0059	0.001	0.0002	0	0	0	0.0001	0.0005	0.0049	0.0272
21	L	4.2192	5.6806	11.0819	22.1137	40.9245	67.0544	97.3861	127.6446	153.2708	169.7157	172.7868	158.9269	125.3279	83.668
22	V	22.0206	36.9099	106.3329	275.5096	573.2448	964.0254	1393.1892	1832.6884	2273.8022	2713.3395	3147.7933	3561.7107	3840.1391	3574.8528
23	D	-	5.3349	9.1597	17.1844	31.7925	54.0801	82.2583	112.5427	140.4794	161.5069	171.2529	165.8376	142.0176	104.0397
24	p0ave	0.1916	0.1483	0.0732	0.0255	0.0059	0.001	0.0002	0	0	0	0.0001	0.0005	0.0049	0.0272
25	STEADY-STATE MOMENTS														
26	Le	4.21920668	5.78426052	15.1030596	-71.921986	-13.936611	-9.1632653	-7.8587106	-7.8587106	-9.1632653	-13.936611	-71.921986	15.1030596	5.78426052	4.2246604
27	Ve	22.0209117	39.2419302	243.205468	5100.85006	180.292505	74.8021658	53.9006211	53.9006211	74.8021658	180.292505	5100.85006	243.205468	39.2419302	22.0724159
28	De	4.21920668	5.78426052	15.1030596	-71.921986	-13.936611	-9.1632653	-7.8587106	-7.8587106	-9.1632653	-13.936611	-71.921986	15.1030596	5.78426052	4.2246604
29	Tre		7.35172377	44.8598051	931.040713	32.6962093	13.5134043	9.7219063	9.7219063	13.5134043	32.6962093	931.040713	44.8598051	7.35172377	4.17777278
30	TIME CONVERSION														
31	mu	23.568	23.16	22.398	21.762	21.258	20.916	20.742	20.742	20.916	21.258	21.762	22.398	23.16	23.568
32	t(duration)	0	10	10	10	10	10	10	10	10	10	10	10	10	10
33	t(absolute)	0	10	20	30	40	50	60	70	80	90	100	110	120	130
34															
35	RESULTS	0	1	2	3	4	5	6	7	8	9	10	11	12	13
36	L	4.2192	5.69291464	11.2320617	20.3801308	38.6370075	64.8128308	95.2737337	125.634434	151.345424	167.877042	171.239638	157.619732	124.099855	83.971231
37	V	22.0206	35.1814105	100.3578	343.250943	682.306446	1071.52459	1482.47363	1901.64153	2321.07303	2730.98266	3069.2513	3410.77077	3712.64612	3220.24341
38	D	4.2192	5.35596521	9.19093065	16.2712618	29.7203729	51.7856548	80.062394	110.461739	138.493812	159.613609	169.559462	164.326106	140.163743	102.676275

Figure 7.3.1 Part of Case worksheet for J3P9 (M/M/1) (one time-slice per column)

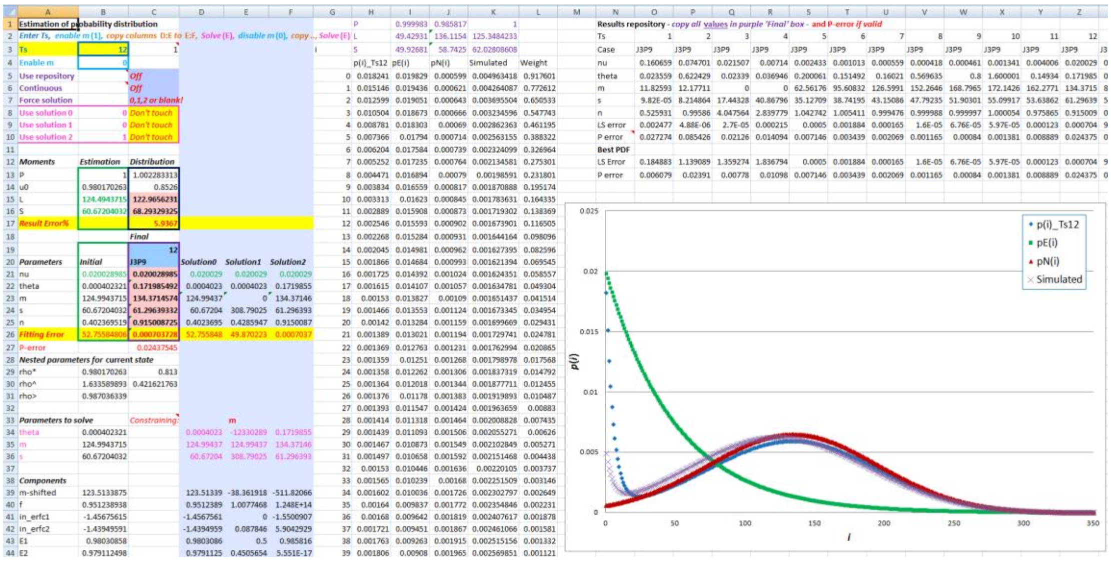


AD41

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	CASE	.M or D	Methods	Limit	Data sourc	Separate	Ia	I	C	I*	C-I	Alt_Ia	Alt_I	Alt_C		L error (total/^\)				D error (total/^\)				V error (total/^\)			
2	J3P9	M	ckt-m	130	c	0	1	1	1	1	0					1.681632	0.106537	1.976323	1.566542	1.680209	0.026637	1.93453	1.615363	257.2951	4.398266	78.64502	377.
3		SIMULATED VALUES								STEADY-STATE MOMENTS				TIME CONVERSION			RESULTS					PROCESS			Force strin	METHOD SWITC	
4		t	rho	mu	p0	L	V	D	p0ave	Le	Ve	De	Tre	mu	t(duration	t(absolute)		Lest	Vest	Dest	u	Time slice	Developm	Method (c,d,k,m,s,t,		Time-orig	Omeg
5	0	0	0.8084	23.568	0.1916	4.2192	22.0206	-	0.1916	4.219207	22.02091	4.219207		23.568	0	0	0	4.2192	22.0206	4.2192	0.8084						
6	1	10	0.8526	23.16	0.1483	5.6806	36.9099	5.3349	0.1483	5.784261	39.24193	5.784261	7.351724	23.16	10	10	1	5.692915	35.18141	5.355965	0.851617	1	1	c		0	
7	2	20	0.9379	22.398	0.0732	11.0819	106.3329	9.1597	0.0732	15.10306	243.2055	15.10306	44.85981	22.398	10	20	2	11.23206	100.3578	9.190931	0.928702	2	1	c		0	
8	3	30	1.0141	21.762	0.0255	22.1137	275.5096	17.1844	0.0255	-71.922	5100.85	-71.922	931.0407	21.762	10	30	3	20.38013	343.2509	16.27126	0.978166	3	2	t		1	
9	4	40	1.0773	21.258	0.0059	40.9245	573.2448	31.7925	0.0059	-13.9366	180.2925	-13.9366	32.69621	21.258	10	40	4	38.63701	682.3064	29.72037	0.992742	4	2	t		1	
10	5	50	1.1225	20.916	0.001	67.0544	964.0254	54.0801	0.001	-9.16327	74.80217	-9.16327	13.5134	20.916	10	50	5	64.81283	1071.525	51.78565	0.997536	5	2	t		1	
11	6	60	1.1458	20.742	0.0002	97.3861	1393.189	82.2583	0.0002	-7.85871	53.90062	-7.85871	9.721906	20.742	10	60	6	95.27373	1482.474	80.06239	0.998978	6	2	t		1	
12	7	70	1.1458	20.742	0	127.6446	1832.688	112.5427	0	-7.85871	53.90062	-7.85871	9.721906	20.742	10	70	7	125.6344	1901.642	110.4617	0.999437	7	2	t		1	
13	8	80	1.1225	20.916	0	153.2708	2273.802	140.4794	0	-9.16327	74.80217	-9.16327	13.5134	20.916	10	80	8	151.3454	2321.073	138.4938	0.99958	8	2	t		1	
14	9	90	1.0773	21.258	0	169.7157	2713.34	161.5069	0	-13.9366	180.2925	-13.9366	32.69621	21.258	10	90	9	167.877	2730.983	159.6136	0.999537	9	2	t		1	
15	10	100	1.0141	21.762	0.0001	172.7868	3147.793	171.2529	0.0001	-71.922	5100.85	-71.922	931.0407	21.762	10	100	10	171.2396	3069.251	169.5595	0.998654	10	2	t		1	
16	11	110	0.9379	22.398	0.0005	158.9269	3561.711	165.8376	0.0005	15.10306	243.2055	15.10306	44.85981	22.398	10	110	11	157.6197	3410.771	164.3261	0.995976	11	-2	m		0	
17	12	120	0.8526	23.16	0.0049	125.3279	3840.139	142.0176	0.0049	5.784261	39.24193	5.784261	7.351724	23.16	10	120	12	124.0999	3712.646	140.1637	0.98003	12	-2	m		0	
18	13	130	0.8086	23.568	0.0272	83.668	3574.853	104.0397	0.0272	4.22466	22.07242	4.22466	4.177773	23.568	10	130	13	83.97123	3220.243	102.6763	0.946518	13	-2	m		0	
19	14	140	0.8086	23.568	0.0685	49.5065	2601.869	65.7421	0.0685	4.22466	22.07242	4.22466	4.177773	23.568	10	140	14	51.86329	1991.302	66.54479	0.912741	14	-2	m		0	
20	15	150	0.8086	23.568	0.1166	26.2514	1433.752	36.91	0.1166	4.22466	22.07242	4.22466	4.177773	23.568	10	150	15	27.90766	833.8211	38.50151	0.87883	15	-2	m		0	
21	16	160	0.8086	23.568	0.1543	13.3538	617.7486	19.048	0.1543	4.22466	22.07242	4.22466	4.177773	23.568	10	160	16	12.27242	170.8146	18.71027	0.845457	16	-2	m		0	
22	17	170	0.8086	23.568	0.1762	7.4775	225.4742	9.9823	0.1762	4.22466	22.07242	4.22466	4.177773	23.568	10	170	17	5.061521	22.07242	7.410401	0.816637	17	-2	m		0	
23	18	180	0.8086	23.568	0.1862	5.2346	81.0477	6.161	0.1862	4.22466	22.07242	4.22466	4.177773	23.568	10	180	18	4.224791	22.07242	4.320102	0.808605	18	-3	m		0	

Figure 7.3.2 Part of Demo worksheet for J3P9 (M/M/1) (one time-slice per row)

The ‘Distribution’ worksheet shown in Figure 7.3.3 is used to estimate the probability distribution for one time slice of the ‘Case’ worksheet, using Excel's ‘Solver’ method.



	A	B	C	D	E	F
1	Estimation of probability distribution					
2	Enter Ts, enable m (1), copy columns D:E to E:F, Solve (E), disable m (0), copy .., Solve (E)					
3	Ts	12	1			
4	Enable m	0				
5	Use repository		Off			
6	Continuous		Off			
7	Force solution		0,1,2 or blank!			
8	Use solution 0	0	Don't touch			
9	Use solution 1	0	Don't touch			
10	Use solution 2	1	Don't touch			
11						
12	Moments	Estimation	Distribution			
13	P	1	1.002283313			
14	u0	0.980170263	0.8526			
15	L	124.4943715	122.9656231			
16	S	60.67204032	68.29329325			
17	Result Error%		5.9367			
18			Final			
19			12			
20	Parameters	Initial	J3P9	Solution0	Solution1	Solution2
21	nu	0.020028985	0.020028985	0.020029	0.020029	0.020029
22	theta	0.000402321	0.171985492	0.0004023	0.0004023	0.1719855
23	m	124.9943715	134.3714574	124.99437	0	134.37146
24	s	60.67204032	61.29639332	60.67204	308.79025	61.296393
25	n	0.402369519	0.915008725	0.4023695	0.4285947	0.9150087
26	Fitting Error	52.75584806	0.000703728	52.755848	49.870223	0.0007037
27	P-error		0.02437545			
28	Nested parameters for current state					
29	rho*	0.980170263	0.813			
30	rho^	1.633589893	0.421621763			
31	rho>	0.987036339				
32						
33	Parameters to solve		Constraining:		m	
34	theta	0.000402321		0.0004023	-12330289	0.1719855
35	m	124.9943715		124.99437	124.99437	134.37146
36	s	60.67204032		60.67204	308.79025	61.296393
37						
38	Components					
39	m-shifted	123.5133875		123.51339	-38.361918	-511.82066
40	f	0.951238938		0.9512389	1.0077468	1.248E+14
41	in_erfc1	-1.45675615		-1.4567561	0	-1.5500907
42	in_erfc2	-1.43949591		-1.4394959	0.087846	5.9042929
43	E1	0.98030858		0.9803086	0.5	0.985816
44	E2	0.979112498		0.9791125	0.4505654	5.551E-17

	G	H	I	J	K	L
1		P	0.999983	0.985817	1	
2		L	49.42931	136.1154	125.3484233	
3		S	49.92681	58.7425	62.02808608	
4		p(i)_Ts12	pE(i)	pN(i)	Simulated	Weight
5	0	0.018241	0.019829	0.000599	0.004963418	0.917601
6	1	0.015146	0.019436	0.000621	0.004264087	0.772612
7	2	0.012599	0.019051	0.000643	0.003695504	0.650533
8	3	0.010504	0.018673	0.000666	0.003234596	0.547743
9	4	0.008781	0.018303	0.00069	0.002862363	0.461195
10	5	0.007366	0.01794	0.000714	0.002563155	0.388322
11	6	0.006204	0.017584	0.000739	0.002324099	0.326964
12	7	0.005252	0.017235	0.000764	0.002134581	0.275301
13	8	0.004471	0.016894	0.00079	0.00198591	0.231801
14	9	0.003834	0.016559	0.000817	0.001870888	0.195174
15	10	0.003313	0.01623	0.000845	0.001783631	0.164335
16	11	0.002889	0.015908	0.000873	0.001719302	0.138369
17	12	0.002546	0.015593	0.000902	0.001673901	0.116505
18	13	0.002268	0.015284	0.000931	0.001644164	0.098096
19	14	0.002045	0.014981	0.000962	0.001627395	0.082596
20	15	0.001866	0.014684	0.000993	0.001621394	0.069545
21	16	0.001725	0.014392	0.001024	0.001624351	0.058557
22	17	0.001615	0.014107	0.001057	0.001634781	0.049304
23	18	0.00153	0.013827	0.00109	0.001651437	0.041514
24	19	0.001466	0.013553	0.001124	0.001673345	0.034954
25	20	0.00142	0.013284	0.001159	0.001699669	0.029431
26	21	0.001389	0.013021	0.001194	0.001729741	0.024781
27	22	0.001369	0.012763	0.001231	0.001762994	0.020865
28	23	0.001359	0.01251	0.001268	0.001798978	0.017568
29	24	0.001358	0.012262	0.001306	0.001837319	0.014792
30	25	0.001364	0.012018	0.001344	0.001877711	0.012455
31	26	0.001376	0.01178	0.001383	0.001919893	0.010487
32	27	0.001393	0.011547	0.001424	0.001963659	0.00883
33	28	0.001414	0.011318	0.001464	0.002008828	0.007435
34	29	0.001439	0.011093	0.001506	0.002055271	0.00626
35	30	0.001467	0.010873	0.001549	0.002102849	0.005271
36	31	0.001497	0.010658	0.001592	0.002151468	0.004438
37	32	0.00153	0.010446	0.001636	0.00220105	0.003737
38	33	0.001565	0.010239	0.00168	0.002251509	0.003146
39	34	0.001602	0.010036	0.001726	0.002302797	0.002649
40	35	0.00164	0.009837	0.001772	0.002354846	0.002231
41	36	0.00168	0.009642	0.001819	0.002407617	0.001878
42	37	0.001721	0.009451	0.001867	0.002461066	0.001581
43	38	0.001763	0.009263	0.001915	0.002515156	0.001332
44	39	0.001806	0.00908	0.001965	0.002569851	0.001121

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
1	Results repository - copy all values in purple 'Final' box - and P-error if valid													
2	Ts	1	2	3	4	5	6	7	8	9	10	11	12	
3	Case	J3P9	J3P9	J3P9	J3P9	J3P9	J3P9	J3P9	J3P9	J3P9	J3P9	J3P9	J3P9	J3
4	nu	0.160659	0.074701	0.021507	0.00714	0.002433	0.001013	0.000559	0.000418	0.000461	0.001341	0.004006	0.020029	0
5	theta	0.023559	0.622429	0.02339	0.036946	0.200061	0.151492	0.16021	0.569635	0.8	1.600001	0.14934	0.171985	0
6	m	11.82593	12.17711	0	0	62.56176	95.60832	126.5991	152.2646	168.7965	172.1426	162.2771	134.3715	8
7	s	9.82E-05	8.214864	17.44328	40.86796	35.12709	38.74195	43.15086	47.79235	51.90301	55.09917	53.63862	61.29639	5
8	n	0.525931	0.99586	4.047564	2.839779	1.042742	1.005411	0.999476	0.999988	0.999997	1.000054	0.975865	0.915009	0
9	LS error	0.002477	4.88E-06	2.7E-05	0.000215	0.0005	0.001884	0.000165	1.6E-05	6.76E-05	5.97E-05	0.000123	0.000704	9
10	P error	0.027274	0.085426	0.02126	0.014094	0.007146	0.003439	0.002069	0.001165	0.00084	0.001381	0.008889	0.024375	0
11	Best PDF													
12	LS Error	0.184883	1.139089	1.359274	1.836794	0.0005	0.001884	0.000165	1.6E-05	6.76E-05	5.97E-05	0.000123	0.000704	9
13	P error	0.006079	0.02391	0.00778	0.01098	0.007146	0.003439	0.002069	0.001165	0.00084	0.001381	0.008889	0.024375	0

Figure 7.3.3 Distribution worksheet example for J3P9 (M/M/1, Ts 12)

The method described earlier in Chapter 5 is applied sequentially, as follows. The required case is first set up in the ‘Case’ worksheet, whose inputs can either be drawn from pre-defined data in the M/M/1 or M/D/1 database worksheet, or set up manually in input fields (not recommended). Switching to the ‘Distribution’ worksheet:

1. Set required time slice (Ts) number in blue-bordered cell at top left
2. Set ‘Enable m’ field to 1 so that all three parameters θ, *m* and *s* will be solved for
3. Copy columns D:E to E:F - this initialises the solution column E to the starting values
4. Run Solver - this estimates optimum parameters in column E
5. Set ‘Enable m’ field to 0 so only parameters θ and *s* will be solved, with *m* forced to 0
6. Copy columns D:E to E:F - this copies the first solution to column F and re-initialises
7. Run Solver again - this estimates new parameters in column E
8. The optimum solution is selected automatically from either column E or F
9. Optionally copy result *values* from the purple-bordered box to the ‘Repository’ area, enabling parameters to be collected and for each time slice and restored
10. An option exists to compare with a simulated distribution stored in another worksheet.

The target and variable cell addresses for Solver are embedded in the worksheet and do not need to be changed. The solution values used are forced to be non-negative. If Solver arrives at a negative value it is replaced by the corresponding initial value, but if Solver moves back into the feasible region then the variable is reconnected. The constraint rules in Solver are not used.

As explained in Chapter 5, the option to set θ to zero, so forcing a pure exponential probability distribution, is not provided as it is considered unnecessary. If the calculated mean and variance or s.d. are consistent with an exponential/geometric distribution then the method is expected to return a very small value of θ, so a third run of Solver will not add anything. Of the two solutions the one with the minimum error is selected by default, but the choice can be forced by setting the appropriate 'Use solution' switch.

The probability distribution corresponding to the solution in the larger purple-bordered box is plotted, with the exponential component shown in **green**, the (truncated but unnormalised) Normal component in **maroon**, and the combined estimated distribution in **blue**. A simulated distribution, if present, is indicated by crosses. The mean and standard deviation of the estimated distribution are given in the smaller purple-bordered box for comparison with analytically estimated values in the green-bordered box to its left, along with % RMS error giving equal weight to mean and standard deviation.

Further options exist:

- To select 'Continuous' mode that suppresses the *h* correction (not recommended)
- To select 'Repository' that displays the distribution using stored data for the selected time slice instead of the Case data
- To force the choice of solution for purposes of comparison or checking.

The Repository will need to be copied and saved before a different case is analysed. Alternatively the entire worksheet can be copied, bearing in mind that Excel graphs retain links to the original worksheet, so these will need to be edited using <right-click>Select Data.

7.4. SUMMARY OF BENCHMARKING AND TEST PROGRAMS

A number of compiled Fortran 95 programs have been written to simulate queue development. **Series** and **Qsim** are described in Chapter 2 and pseudo-code of their principal algorithms given in Appendices A and B. These and other programs are summarised in the Table below.

Table 7.4.1 Summary of simulation and other programs

Program	Purpose	Command inputs	Input files	Output
Series	Evaluates Morse's M/M/1 series formula	options, resolution, L_0, ρ, μ, t	n/a	.tab=dimension (max i), p_0, L, D, V .out=probabilities{p_i} [by time slice]
Qsim	Markov simulation of M/M/1, M/D/1, M/D/1[G] and various Erlang processes	data name, [ρ], process, step size, initial queue, with optional modifiers	.dat= {ρ, μ, t} with various options by sign	.sum={ p_0, L, D, V } .dis=probabilities {p_i} [by time slice]
LodefV	Evaluation of Gaussian peaks and synthesis of time-sliced data for other programs	'*Jx*P*y*' junction ID, peak length, number of time slices	n/a	.lod=L, V .sin= data for Series .dat=data for Qsim
LaneMultiRandom	Random event simulation of queuing on one or more lanes with variable degree of selectivity of shortest queue	[randomise_seed], case_code, lanes, ρ, μ, selectivity, number_of_trials, [number of run-up trials not counted]	n/a	Multi<case>.txt= detailed record of arrival and service events and queue sizes, Multi<case>.tot= summary statistics for each lane including correlations
LaneMultiStates	Synthesises recurrence relations for up to six lanes, then applies them in Markov simulation to estimate probability distributions	ρ, lanes, step size, maximum total queue, number of cycles, modifier for selective arrivals	n/a	Table of recurrence relation coefficients for first few distinct recurrence relations. Ensemble and lane probability distributions. Joint probability matrix for two lanes.
LaneTurningRandom	Random event simulation of queuing on two lanes with turning movements	turnsin.txt=output filename, proportion of ahead movement using right lane, exit capacities, total demand, turning proportions, random seed, number of cycles, FIFO option[65]		Queue size probability distributions: total, left-lane, right-lane, left-turning, left-lane ahead, right-lane ahead, right-turning

[65] The FIFO option means that rather than just storing the number of units in each queue, the program records their actual positions dynamically in a large array, at the cost of some loss of speed.

CHAPTER 8: SUMMARY AND CONCLUSIONS

8.1. ACHIEVEMENTS

8.1.1 Extension and demonstration of tools for estimating time-dependent queues

Main results

By deriving and applying a new deterministic formula for queue variance, the tools available for macroscopic estimation of time-dependent queues have been extended, improving their accuracy and range of application, and enabling estimation of variance and reliability. Methods of fitting probability distributions to equilibrium and calculated time-dependent non-equilibrium moments have been developed, providing detail particularly for ‘tails’.

Guiding principles

An objective throughout has been to develop or enhance methods that are computationally efficient, robust, and as general as possible, by exploiting the underlying structures of queuing systems rather than seeking statistical matches to simulation benchmarks.

Validation and demonstration

The methods show good agreement when tested against Markov simulations of M/M/1 (priority-like) and M/D/1 (signal-like stochastic/overflow) processes based on recurrence relations. The M/M/1 Markov simulation method is itself validated against calculation based on the exact series formula as given by Morse, and random event simulation. Application to a range of oversaturated peak and other cases is demonstrated. A spreadsheet demonstration program enables mean queue size, delay and variance to be evaluated from piecewise time-dependent demand/capacity profiles, and includes a procedure for estimating instantaneous queue size probability distributions from the time-dependent moments.

8.1.2 Main innovations

New deterministic and equilibrium queue variance formulae have been derived. These respectively partner the deterministic and Pollaczek-Khinchin (P-K) equilibrium mean queue formulae, enabling the variance of a queue to be estimated at any point in time, and giving improved queue estimation accuracy through their role as asymptotic constraints.

The new variance formulae have been used with inherent structural features of queue processes to correct the sheared queue formula, a convenient and computationally efficient approximation that seamlessly combines time-dependent deterministic and equilibrium queue formulae.

A new approximation for queue decay based on exponential functions has been developed, that also relies on structural features of queuing processes, but avoids the quasi-static implication of the sheared method that appears inappropriate to the decay regime.

Mean queues with Erlang arrival/service have been parameterised in the Pollaczek-Khinchin formula, using the results of Markov simulations based on appropriate recurrence relations, confirming that these processes can be brought into the time-dependent approximation scheme.

New empirical formulae relating the mean and variance of a stochastic equilibrium queue at a signal to throughput capacity in the green period have been developed, based on an extended M/D/1 process defined by recurrence relations. These results are compatible with time-dependent modelling and provide moment expressions needed for estimating time-dependent queues and equilibrium probability distributions taking into account green capacity.

A new doubly-nested geometric approximation to discrete equilibrium probability distributions by fitting parameters to three queue moments has been developed and demonstrated. It is considered that three critical moments are necessary and sufficient to characterise a queue process for present practical purposes: the utilisation or its complement the probability that the queue is zero, mean queue size and variance.

A new method of fitting a dynamic queue size probability function to queue moments using a combination of exponential and Normal functions has been developed and demonstrated. The continuous distribution function can be discretised numerically or by integration.

A new method for estimating capacities and queues on shared lanes with turning movements has been described. The method based on a general approach, shows fair agreement in estimating the resulting time-dependent queues when compared against simulation.

8.1.3 Review of relevant past work and underlying issues

Relevant past work has been reviewed, and a number of issues related to the above methods and topics discussed. General discussion in the Introduction has been supported by reviewing and addressing specific issues in the appropriate sections. Particular issues include the generality of the deterministic variance formula, sources of inaccuracy in the sheared formula, alternative types of equilibrium distribution, the significance of initial probability distributions and the significance of utilisation or the probability of zero queue, the traffic and statistical parameters needed to characterise the subsequent evolution of a queue, methods applicable to multi-lane queues, and diffusion approximations to queue size probability distributions.

8.2. IMPACTS

8.2.1 More reliable traffic and assignment modelling

Microscopic simulation has gained in popularity in recent years, because of the apparent realism of its disaggregate approach, and its ease of visualisation. It has been criticised on the grounds that it delivers random samples rather than averages and uses internal models that are not directly verifiable. Macroscopic and mesoscopic, including agent-based and cell-transmission, methods make use of validated closed formulae, giving them the advantages of efficiency and providing average results in a single pass, potentially with uncertainty ranges.

8.2.2 Better treatment of variability and unreliability

Transportation traffic behaviour is variable and to a degree unpredictable. While microscopic simulation embodies randomness and variability at source, the large number of possible outcomes means it cannot explore the effects of variability efficiently and accurately. With reliability of transport becoming increasingly important as the scope for infrastructure provision decreases, an ability to evaluate the reliability of outputs as a function of the variability or uncertainty of inputs and processes becomes increasingly necessary. Although variability in queue and delay processes cannot embrace all possible types of uncertainty, such as ‘heavy tails’ or rare events, it may provide a base description to replace simple averages that could be misleading. Queuing theory has many applications apart from road traffic, including systems that do not conform to an AM/PM peak pattern but have systematic features in their arrival and service patterns, like airports and health care facilities. Robustness of queue estimation methods is essential where arrival and service profiles may be arbitrary. Estimation of variance means that critical points can be identified more reliably and potentially mitigated.

8.3. POTENTIAL FUTURE WORK

This work has aimed to provide practical tools for estimating traffic and similar queues in which variability is an integral part, supported by certain methodological extensions. Topics for future research in this area could include: extension at least approximately to more general arrival and service patterns including platooned arrivals; calculation of higher moments of distributions with possible application to ‘heavy tails’ or to further constrain queue development; other improvements to approximations to reduce the need for calibrated adjustments; application of the new methods to network assignment modelling. These may also lead to a fuller understanding of relevant mechanisms and increased confidence in predictions.

REFERENCES

AA (2009). http://www.theaa.co.uk. February 2009.

Abate J and Whitt W (1987). Transient behavior of the M/M/1 queue: starting at the origin. *Queueing Systems 2*, (1987), 41-65.

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

Addison J D and Heydecker B G (2006). Journey time variability on a congested link. *Proc. UTSG Conference 2006*, Dublin.

Allsop R E (1971). Delay-minimizing settings for fixed-time traffic signals at a single road junction. *IMA Journal of Applied Mathematics,* 8(2), 164-185.

Allsop R E / Hutchinson T P (1972). Delay at fixed-time traffic signals [Parts I and II respectively]. *Transportation Science* 6(**3**), 260-285, 286-305.

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.

Akçelik R (1998a). Traffic signals: capacity and timing analysis.*Report ARR 123*. Australian Road Research Board. [Most recent edition of report first published 1981]

Akçelik R (1998b). Roundabouts: capacity and performance analysis. *Report ARR 321*. Australian Road Research Board.

Akçelik (2001). *HCM 2000 back of queue model for signalised intersections*. Technical Note. Akçelik and Associates Pty Ltd, September 2001.

Akçelik R (2007). A review of gap-acceptance capacity models. *Proc. 29th Conference of Australian Institutes of Transport Research (CAITR 2007)*, University of South Australia, Adelaide, 5-7 December 2007.

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, dft_econappr_pdf_610439.

Asmussen S (1987). *Applied probability and queues*. Wiley.

Bar-Gera H (2002). Origin-based algorithm for the traffic assignment problem. *Transportation Science*, 36(**4**), 398-417.

Bar-Gera H and Boyce D (2003). Origin-based algorithms for combined travel forecasting models. *Transportation Research B*, 37(**2003**), 405–422

Beckman M (1952). A continuous model for transportation. *Econometrica*, **20**, 643-660.

Bell M G H (2009). Hyperstar: a multi-path Astar algorithm for risk-averse vehicle navigation. *Transportation Research*,43B, 2009, 97-107.

Bertini R L, Hasen S and Bogenberger K (2005). Empirical analysis of traffic sensor data surrounding a bottleneck on a German autobahn. *Transportation Research Record 1934*, Transportation Research Board, Washington DC.

Bin Han (1996). A new comprehensive sheared delay formula for traffic signal optimisation. *Transportation Research A*, 30(**2**), 155-171.

Binning J (1996). Visual PICADY/4 User Guide. *TRL Application Guide AG23*, Transport Research Laboratory. Crowthorne House.

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

Branston D (1978). A comparison of observed and estimated queue lengths at over-saturated traffic signals. *Traffic Engineering and Control* 19(**7**), 322-327.

Brilon W (2007). Time dependent delay at unsignalized intersections. *Proc. International Symposium on Transportation and Traffic Theory 2007*. Elsevier.

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.

Cantrell P E (1986). Computation of the transient M/M/1 queue cdf, pdf and mean with generalized Q-functions. *IEEE transactions on communications*, COM-94(**8**), August 1986.

Cantrell P E and Ojha A K (1987). Comparison of generalized Q-function algorithms. *Correspondence in IEEE Transactions on Information Theory*, IT-33(**4**), July 1987.

Cantrell P E and Beall G R (1988). Transient M/M/1 queue variance computation using generalized Q-functions. *IEEE Transactions on Communications*, 36(**6**), June 1988.

Carey M and Bowers M (2011). A review of properties of flow-density functions. *Transport Reviews*, 31(**1**), 49-73, Routledge.

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

CEDR (2009). *Traffic Incident Management: Final Report of Task 5*, February 2009. Conference of European Directors of Roads, http://www.cedr.fr

CEDR (2011). *Best practice in European Traffic Incident Management*.Final Report of Task 13 , May 2011. Conference of European Directors of Roads, http://www.cedr.fr

Cheng D X, Messer C J, Tian Z Z and Liu J (2003). Modification of Webster's Minimum Delay Cycle Length Equation Based on HCM 2000. [Texas Transportation Institute] *TRB 2003 Annual Meeting*, Washington DC. http://wolfweb.unr.edu/homepage/zongt/Publications_files/ChengDingXinTRB-03.pdf

Chow J Y J (2013). On observable chaotic maps for queueing analysis. *Proc. 92nd TRB Annual Meeting, January 2013*. Transportation Research Board, Washington DC.

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

COBA (2012). http://www.dft.gov.uk/publications/coba-11-user-manual/ [accessed 26/10/12]

Cronjé (1983a). Analysis of existing formulas for delay, overflow and stops. *Transportation Research Record 905*, Transportation Research Board, Washington DC.

Cronjé (1983b). Optimization model for isolated signalized traffic intersections. *Transportation Research Record 905*, 80-83, Transportation Research Board, Washington DC.

Daganzo C F, Cassidy M J and Bertini R L (1999). Possible explanations of phase transitions in highway traffic. *Transportation Research A*, 33(**1999**), 365-379.

DfT (2005). *Attitudes to congestion on motorways and other roads.* Department for Transport.

Dijkstra E W (1959). A note on two problems in connexion with graphs. *Numerische Mathematik 1*, 269-271.

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

Eddington Sir R (2006). *The Eddington Transport Study Main Report: Transport's role in sustaining the UK's productivity and competitiveness*. HMSO.

Eliasson J (2006). Forecasting travel time variability. *Proc. European Transport Conference 2006*. PTRC

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

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/

Fosgerau M and Karlström A (2010). The value of reliability. *Transportation Research B,* 44(2010), 38-49. Elsevier.

Global Times (2010). *Highway jam enters its 9th day, spans 100km.* 23 Aug 2010. http://china.globaltimes.cn/society/2010-08/566070.html?loc=interstitialskip

Goodwin P (2010). 'Peak car'. Various articles in *Local Transport Today. Issues 548-554.*

Gordon A, Van Vuren T, Watling D, Polak J, Noland R, Porter S and Taylor N B (2001). Incorporating variable travel time effects into route choice models. *Proc. European Transport Conference, Homerton College, Cambridge, September 2001,* PTRC.

Greenshields B D (1935). A study of traffic capacity. *Proc. 14th Annual Meeting of Highway Research Board*.

Griffiths J D (1981). A mathematical model of a non-signalized pedestrian crossing. *Transportation Science*, 15(**3**), 222-232.

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.

Griffiths J D, Leonenko G M and Williams J E (2008). Time-dependent analysis of non-empty M/E*k*/1 queue. *Quality Technology and Quantitative Management*, 5(**3**), 309-320.

Gross D, Shortle J F, Thompson J M and Harris C M (2008). *Fundamentals of queueing theory*. Wiley.

Hart P E, Nilsson N J and Raphael B (1968). A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics SSC4*(**2**), 100–107.

Hawkes A G (1968). Gap-acceptance in road traffic. *J. Applied Probability*, 5(**1**), 84-92.

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

Heydecker B G (unpublished). *An extension of the Pollaczek-Khintchine delay formula*. Transport Studies Group, University College London.

Heydecker B G (1982). Vehicles, PCUs and TCUs in traffic signal calculations. *Traffic Engineering and Control*, 24(**3**), 111-114.

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.

Heydecker B G and Wu J (2001). Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference. *Advances in Engineering Software* 32(**2001**), 859-869.

Holland W and Griffiths J D (1999). A time-dependent approximation for the queue M/M(1,*s*)/*c. IMA J. of Mathematics Applied in Business & Industry*, 1999(**10**),213-223.

Kaparias I, Bell M G H and Belzner H (2008). A new measure of travel time reliability for in-vehicle navigation systems. *J. of Intelligent Transportation Systems*, 12(**4**), 202-211.

Kelly B (2012). A 'green-wave' reprieve. *Traffic Engineering and Control*, February 2012.

Kerner B S and Klenov S L (2003). Microscopic theory of spatio-temporal traffic patterns at highway bottlenecks. *Physical Review* E, **68**, 036130.

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

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, McDonald M and Hounsell N B (1985). Passenger car units in saturation flows: concept, definition, derivation. *Transportation Research*,19B(**1**), 39-61.

Kimber R M, McDonald M and Hounsell N B (1986). The prediction of saturation flows for road junctions controlled by traffic signals. *TRL Report RR67*. Transport Research Laboratory, Crowthorne House.

Kimber R M, Summersgill I and Burrow I J (1986). Delay processes at unsignalised junctions: the inter-relation between geometric and queueing delay. *Transportation Research*, 20B(**6**), 457-476.

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

Kleinrock L (1976). *Queueing systems: Volume 1I Computer applications*. Wiley

Knote T (2006). Kapazitat von Mischstromen in Nebenstrassenzufahrten von Kreuzungen und Einmündungen mit Vorfahrtbeschilderung. *Straßenverkehrstechnik,* 50(2), 75-80.

Kobayashi H (1974a). Application of the diffusion approximation to queueing networks – Part I: equilibrium queue distributions. *J Association for Computing Machinery*, 21(**2**), 316-328.

Kobayashi H (1974b). Application of the diffusion approximation to queueing networks – Part II: non-equilibrium queue distributions and applications to computer modeling. *J Association for Computing Machinery*, 21(**3**), 459-469.

Koenigsberg E (1991). Is queueing theory dead? *OMEGA International Journal of Management Science*, 19(**2/3**), 69-78. Pergamon.

Kouvatsos D D (1988). A maximum entropy analysis of the G/G/1 queue at equilibrium. *J. Operational Research Society*, 39(**2**), 183-200.

Kühne R and Lüdke A (2013). Traffic breakdowns and freeway capacity as extreme value statistics. *Transportation Research C*, 27(**2013**), 159-168. Elsevier.

Laio F (2004). Cramér-von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters. *Water Resources Research*, 40, W09308.

Lay M G (2011). Measuring traffic congestion. *Road and Transport Research,* 20(**2**), June 2011. Australian Road Research Board.

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

Lighthill Sir M J, Whitham G B(1955). On kinetic wave II: a theory of traffic flow oncrowded roads. *Proceedings of the Royal Society of London, Series A*, 229(**1178**), 317-345.

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

LTT (2012). In Passing, *in Local Transport Today Issue 593, 30 March – 12 April 2012*.

Maher M J (1992). SAM – a stochastic assignment model. *Proc. IMA Mathematics in Transport Conference.* Institute of Mathematics and its Applications.

Maher M J (1998). Algorithms for logit-based stochastic user equilibrium assignment. *Transportation Research – Methodological*, 32B(**8**), 539-549.

Mahmassani H S and Chang G-L (1987). On Boundedly-Rational User Equilibrium in Transportation Systems. *Transportation Science*, 21(**2**).

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

Meissl P (1963). Zufallsmodell einer Lichtsignalanlage mit mehrspurigem Stauraum. *Mathematik-Technik-Wirtschaft***,** Heft l/63, 1-4, and Heft 2/63,63-68, Wien.

Metz D (2009). Saturation of demand for travel. *Transport Reviews, 2009*, 1-16.

Metz D (2010). We have reached the limit of personal travel demand. *Local Transport Today. Issue 557,* 11 June 2010.

Miller A J (1961). A queueing model for road traffic flow. *Journal of the Royal Statistical Society*, B23(**1**), 64-90.

Miller A J (1969). *Some operating characteristics of fixed-time signals with random arrivals.* Institution of Highways and Traffic Research, University of New South Wales.

Mirchandani P B and Zou N (2007). Queuing models for analysis of adaptive signal control. *IEEE Transactions on Intelligent Transportation Systems*, 8(**1**), 50-59.

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.

NCTIM (2002). A road map to the future. *Proceedings of the National Conference on Traffic Incident Management*, Irvine CA, 2-4 March 2002.

Newell G F (1960). Queues for a fixed-cycle traffic light. *Annals of Mathematical Statistics*, 31, 589-597.

Newell G F (1968a). Queues with time-dependent arrival rates – Part I the transition through saturation. *Journal of Applied Probability*, 5(**2**), 436-451.

Newell G F (1968b). Queues with time-dependent arrival rates – Part II the maximum queue and return to equilibrium. *Journal of Applied Probability*, 5(**3**), 579-590.

Newell G F (1968c). Queues with time-dependent arrival rates – Part III a mild rush hour. *Journal of Applied Probability,* 5(**3**), 591-606.

Newell G F (1971,1982). *Applications of queuing theory*. Chapman and Hall.

Olszewski P S (1990). Modelling of queue probability distribution at traffic signals. *Proc. International Symposium of Traffic and Transportation Theory, 1990*.

Peterson M D, Bertsimas D J and Odoni A R (1995). Models and algorithms for queueing congestion at airports. *Management Science*, 41(**8**), 1279-1294.

Prashker J N (2008). Editorial in special issue of journal, *Transportation Research C,* 16 (2008), 275-276.

Rider K L (1976). A simple approximation to the average queue size in the time-dependent M/M/1 queue. *J Association of Computing Machinery*, 23(**2**), 361-367.

Rigobon R (2009). *Brownian motion and stochastic calculus introductory notes*. MIT, Fall 2009.

Robertson D I and Gower P (1977). User Guide to TRANSYT version 6. TRL report LR255, Transport Research Laboratory, Crowthorne House.

Rose C (1995). A statistical identity linking folded and censored distributions. *J Economic Dynamics and Control*, 19(**1995**), 1391-1403.

Sakasegawa H (1977). An approximation formula ... *Annals Inst. Stat. Math.* 29A(1), 67-75.

Sakasegawa H and Yamazaki G (1977). Inequalities and an approximation formula for the mean delay time in tandem queueing systems. *Annals Inst. Stat. Math.* 29A(1), 445-466.

SATURN (2012). https://saturnsoftware.co.uk/downloads/pdfs/Saturn_Brochure_300.pdf. [accessed 26/10/12]

Schrank D, Lomax T and Turner S (2010). *TTI's 2010 Urban Mobility Report.* Texas Transportation Institute.

Semmens M C (1985a). ARCADY2: an enhanced program to model capacities, queues and delays at roundabouts. TRL Report RR 35. Transport Research Laboratory, Crowthorne House.

Semmens M C (1985b). PICADY2: an enhanced program to model capacities, queues and delays at major/minor priority junctions. TRL Report RR 36. Transport Research Laboratory, Crowthorne House.

Sharma O P (1990). Markovian queues. Ellis Horwood.

Sheffi Y and Powell W B (1982). An algorithm for the equilibrium assignment problem with random link times. *Networks* 12(**2**), 191-207.

Slavin H (2012). Convergence of assignment methods. Presentation given at Modelling World Conference 2012, The Oval, London.

Soros G (1987). *The alchemy of finance*. John Wiley & Sons.

Spiess H. Conical volume-delay functions. *Transportation Science 24(**2**),* 153-158.

Tanner J C (1961). Two Papers on Applications of Stochastic Processes to Road Traffic Problems: Delays on a Two-Lane Road. *Journal of the Royal Statistical Society*, B23(**1**), 38-63.

Tanner J C (1962). A theoretical analysis of delays at an uncontrolled intersection. *Biometrika*, 49 (**1** and **2**), 163-169.

Taylor N B (1989). Speeding up quickest route assignment in CONTRAM using an heuristic algorithm.*Traffic Engineering and Control, February 1989*.

Taylor N B (1990). CONTRAM 5: An enhanced traffic assignment model. *TRL Report RR249*. Transport Research Laboratory, Crowthorne House.

Taylor N B (2003). The CONTRAM dynamic traffic assignment model. *Networks and Spatial Economics Journal – special issue on Dynamic Traffic Assignment.* 3: (2003) 297-322, Kluwer, September 2003.

Taylor N B (2005a). Variance and accuracy of the sheared queue model. *Proc. IMA Mathematics in Transport Conference*, University College London, 7-9 September 2005.

Taylor N B (2005b). The impact of Abnormal Loads on road traffic congestion. *Proc. European Transport Conference, Strasbourg, 3-5 October 2005*, AET.

Taylor N B (2007). A new approach to modelling variability in queues. *Proc. European Transport Conference, Leeuwenhorst, 17-19 October 2007.*

Taylor N B (2009). The management and impact of abnormal loads. *TRL Insight Report INS004*, Transport Research Laboratory, Crowthorne House.

Taylor N B (2011). An approach to time-dependent modelling of queues in multiple lanes with turning movements. *Proc. Universities Transport Studies Group (UTSG) Conference 2011,* Open University Milton Keynes.

Taylor N B (2012). A recipe for jam - can congestion be defined consistently? *Proc. Road Traffic and Control (RTIC) Conference, September 2012*, IET London.

Taylor N B (2013). The effect of green time on signal overflow queues. *Proc. Universities Transport Studies Group (UTSG) Conference, January 2013,* Oxford University.

Taylor N B and Heydecker B G (2013). The effect of green time on stochastic queues at traffic signals. *Transportation Planning and Technology - Special Issue based on UTSG 2013 Conference at Oxford University.* Published online on 18 October 2013.
http://dx.doi.org/10.1080/03081060.2013.844907

Taylor N B, Notley S, Bourne N and Skrobanski G (2008). Evidence for speed-flow relationships. *Proc. European Transport Conference, Leeuwenhorst, 6-8 October 2008*, AET.

Trabka E A and Marchand E W (1970). Mean and variance of the number of renewals of a Censored Poisson process. Biological Cybernetics, 7(**6**), 221-224, Springer.

TRLSoftware (2012). Cell Saturation flows and the noticable effects of switch from PDM to CTM. https://www.trlsoftware.co.uk/support/knowledgebase/articles/194 [accessed 26/10/12]

van Vliet (1977). D'Esopo: a forgotten tree-building algorithm. *Traffic Engineering and Control, July/August 1977.*

Wardrop J G (1952). Some theoretical aspects of road traffic research, *Proceedings, Institute of Civil Engineers*, PART II, Vol.1, 325-378.

Webster F V (1958). Traffic signal settings. *Road Research Technical Paper 39*. Road Research Laboratory, Langley.

Webster F V and Cobbe B M (1966). Traffic signals. *Road Research Technical Paper 56*. HMSO.

Whiting P D and Hillier J A (1960). A method of finding the shortest route through a road network. *Operational Research Quarterly,* 11(**1-2**), 37-40.

Whitt W (1982). Refining diffusion approximations for queues. *Operations Research Letters*, 1(**2**), 165-168.

Willmot G E (1986). Mixed compound Poisson distributions. *Astin Bulletin vol 16 S*, 59-79.

Wood S (2012). Traffic microsimulation – dispelling the myths, *Traffic Engineering and Control, October 2012*, 339-344.

Zhang K and Excell A (2013). Modelling a complex give-way situation – AIMSUN vs LINSIG. Road and Transport Research, 22(**2**), 16-26. Australian Road Research Board, June 2013.

ADDITIONAL BIBLIOGRAPHY

This section lists some other sources consulted but not referenced in the text.

Abate J and Whitt W (1992). Transient behavior of the M/G/1 workload process. *Operations Research*, 42(**4**), 750-763.

Akgüngör A P (2008). A new delay parameter dependent on variable analysis periods at signalized intersections. *Turkish Journal of Transport* 23(**1-2**), 31-36 and 91-94.

Akgüngör A P and A G R Bullen (2007). A new delay parameter for variable traffic flows and signalized intersections. *Turkish Journal of Engineering and Environmental Science* 31(**2007**), 61-70.

Atkinson J B (1995). The transient M/G/1/0 queue: some bounds and approximations for light traffic with application to reliability. *J. of Applied Mathematics and Stochastic Applications*, 8(**4**), 347-359.

Baykal-Gürsoy M, Xiao W and Ozbay K (2009). Modeling traffic flow interrupted by incidents. *European J. of Operational Research*, **195**(2009), 127-138.

Celeux G, Lavergne C and Vernaz Y (2000). Assessing material aging from doubly censored data: Weibull distribution vs Poisson process. *Rapport de Recherche no 3857*, Unité de Recherche INTIA Rhône-Alpes.

Cooper R B (1981). *Introduction to queueing theory*. Elsevier.

Cox D R and Miller H D (1965). *The theory of stochastic processes*. Wiley.

Daigle J N and Magalhães M N (1991). Transient behaviour of M/M^{ij}/1 queues. *Queueing Systems* **8** (1991), 357-378.

Duda A (1983). *Transient diffusion approximations for some queueing systems*. Université de Paris-Sud.

van Eenige M J A (1996). Queueing systems with periodic service. Eindhoven University of Technology.

Garcia J-M, Brun O and Gauchard D (2002). Transient analytical solution of M/M/1/N queues. *J. Applied Probability*, **39**, 853-864.

Gaver D P (1968). Diffusion approximations and models for certain congestion problems. *J. of Applied Probability*, **5**, 607-623.

Gaver D P and Miller R G (1962). Limiting distributions for some storage problems. *Studies in applied probability and management science*, 110-126, Stanford.

Grassmann W (1976). Transient solutions to Markovian queues – An algorithm for finding them and determining their waiting-time distributions. *European J. of Operational Research***1** (1977), 396-402.

Lee H W, Yoon S H and Lee S S (1996). A continuous approximation for batch arrival queues with threshold. *Computers Operations Research*, 23(**3**), 299-308.

van Leeuwaarden J S H (2006). Delay analysis for a fixed-cycle traffic-light queue. *Transportation Science 40(**2**)*, May 2006, 189-199.

Li H, Bovy P H L and Bliemer M C J (2008). Departure time distribution in the stochastic bottleneck model. *Int. J. of ITS Research*, 6(**2**), 79-86

Lipsky L (1992). *Queueing theory – a linear algebraic approach*. Macmillan

Loynes R M (1961). *A continuous-time treatment of certain queues and infinite dams.* Statistical Laboratory, University of Cambridge.

Miller R G Jr (1963). Continuous time stochastic storage processes with random linear inputs and outputs. *J. Math. Mech.* 12, 275-291.

Mullowney P and James A (2007). The role of variance in capped-rate stochastic growth models with external mortality. *J. of Theoretical Biology*, **244**(2007), 228-238.

Odoni A R and Roth E (1983). An empirical investigation of the transient behavior of stationary queueing systems. *Operations Research* 31(**3**), 432-455.

Selvin S (1974). Maximum likelihood estimation in the truncated or censored Poisson distribution. *J. of the American Statistical Association*, 69(**345**), Theory and Methods Section.

Stern T E (1979). Approximation of queue dynamics and their application to adaptive routing in computer communication networks. *IEEE Transactions on Communications, COM-27(**9**),* September 1979.

Viti F and van Zuylen H J (2009). The dynamics and the uncertainty of queues at fixed and actuated controls: a probabilistic approach. *J. Intelligent Transp. System*s, 13(**1**), 39-51.

Yeo G F (1961). *Single server queues with the modified service mechanisms*. Australian National University, Canberra.

APPENDIX A – DERIVATION OF MEAN AND VARIANCE FORMULAE

A.1 TIME DERIVATIVE AND DETERMINISTIC MEAN OF THE M/M/1 QUEUE

The dynamic recurrence relations for the M/M/1 queue, which can be formulated on an infinitesimal service time period, are:

$$\frac{dp_0}{dt} = \mu p_1 - \lambda p_0$$

$$\frac{dp_i}{dt} = \mu p_{i+1} - (\mu + \lambda) p_i + \lambda p_{i-1} \qquad \text{(A.1)}$$

The rate of change of the first moment is:

$$\sum_{i=1}^{\infty} i \frac{dp_i}{dt} = \mu \sum_{i=1}^{\infty} i p_{i+1} - (\mu+\lambda) \sum_{i=1}^{\infty} i p_i + \lambda \sum_{i=1}^{\infty} i p_{i-1}$$

$$= \mu \sum_{i=1}^{\infty} (i+1) p_{i+1} - \mu \sum_{i=1}^{\infty} p_{i+1} - (\mu+\lambda) \sum_{i=1}^{\infty} i p_i + \lambda \sum_{i=1}^{\infty} (i-1) p_{i-1} + \lambda \sum_{i=1}^{\infty} p_{i-1}$$

$$\dot{L} = \mu L - \mu p_1 - \mu(1 - p_0 - p_1) - (\mu+\lambda) L + \lambda L + \lambda \qquad \text{(A.2)}$$

$$\dot{L} = (\rho - 1 + p_0)\mu = (\rho - u)\mu \qquad \text{where} \quad \rho = \frac{\lambda}{\mu},\ u = 1 - p_0 \qquad \text{(A.3)}$$

Integrating:

$$L = L_0 + (\rho - x)\mu t \qquad \text{where} \quad x \equiv \frac{1}{t}\int_0^t u(y)\,dy \qquad \text{(A.4)}$$

A.2 DETERMINISTIC VARIANCE OF THE M/M/1 QUEUE

The rate of change of the second moment is:

$$\sum_{i=1}^{\infty} i^2 \frac{dp_i}{dt} = \mu \sum_{i=1}^{\infty} i^2 p_{i+1} - (\mu+\lambda) \sum_{i=1}^{\infty} i^2 p_i + \lambda \sum_{i=1}^{\infty} i^2 p_{i-1}$$

$$= \mu \sum_{i=1}^{\infty} (i+1)^2 p_{i+1} - 2\mu \sum_{i=1}^{\infty} (i+1) p_{i+1} + \mu \sum_{i=1}^{\infty} p_{i+1} - (\mu+\lambda) \sum_{i=1}^{\infty} i^2 p_i$$
$$+ \lambda \sum_{i=1}^{\infty} (i-1)^2 p_{i-1} + 2\lambda \sum_{i=1}^{\infty} (i-1) p_{i-1} + \lambda \sum_{i=1}^{\infty} p_{i-1} \qquad \text{(A.5)}$$

$$\dot{M}_2 = \mu(M_2 - p_1) - 2\mu(L - p_1) + \mu(1 - p_1 - p_0) - (\mu + \lambda)M_2 + \lambda M_2 + 2\lambda L + \lambda$$

$$\begin{aligned} \dot{M}_2 &= 2(\lambda - \mu)L + \mu + \lambda - \mu p_0 = (-2(1-\rho)L + 1 + \rho - p_0)\mu \\ &= (2\rho - 2(1-\rho)L)\mu - \dot{L} \end{aligned} \qquad \text{(A.6)}$$

Integrating:

$$M_2 = 2\rho\mu t - 2(1-\rho)\mu\int_0^t L(y)dy - L + \text{constant} \qquad \text{(A.7)}$$

Hence, on introducing the boundary conditions at t=0, and t=∞ where D=L_e:

$$V = V_0 + L_0(L_0 + 1) + 2(1-\rho)(L_e - D)\mu t - L(L+1) \qquad \text{where}$$

$$D \equiv \frac{1}{t}\int_0^t L(y)dy \ , \qquad L_e = \frac{\rho}{1-\rho} \qquad \text{(A.8)}$$

A.3 TIME DERIVATIVE AND DETERMINISTIC MOMENTS OF THE M/D/1 QUEUE

If notional states i = -1,0 are admitted, so that the real p_0 = $p_{(0)}$ + $p_{(-1)}$, then the recurrence relations of the M/D/1 queue, which is formulated on a finite service time interval, are:

$$p_i(\mu t + 1) = \sum_{j=0}^{i+1} \frac{\rho^j e^{-\rho}}{j!} p_{i+1-j}(\mu t) \qquad \text{(A.9)}$$

If the average rate of change is then taken to be given by the finite difference:

$$\frac{dp_i}{dt} \approx \mu\Delta p_i = \mu(p_i(\mu t + 1) - p_i(\mu t)) \qquad \text{(A.10)}$$

only derivatives of p_i for i > 0 contribute to the mean, but a singleton p_0 term arises from the sum of all the p_0 terms in the differential relations, so:

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

$$\begin{aligned} &= \sum_{i=0}^{\infty}\left[\left(\rho e^{\rho} + (i-1)e^{\rho} - ie^{\rho}\right)p_i\right] + p_0 \\ &= (\rho-1)e^{\rho}\sum_{i=0}^{\infty}\left[p_i\right] + p_0 \qquad \text{hence} \end{aligned} \tag{A.11}$$

$$\dot{L} = \left(\rho - 1 + e^{-\rho}p_0\right)\mu = (\rho - u)\mu \tag{A.12}$$

At equilibrium each expression must be zero, as for M/M/1, so:

$$p_0 = e^{\rho}(1-\rho) \tag{A.13}$$

But as explained in the main dissertation this is the p_0 at the *end* of the service period, not the *average* $\bar{p}_0$ in the service period that determines the utilisation. The second moment is:

$$\begin{aligned} \frac{e^{\rho}}{\mu}\dot{M}_2 &= \sum_{i=0}^{\infty}\left[\left(\sum_{j=0}^{\infty}(i+j-1)^2\frac{\rho^j}{j!} - i^2e^{\rho}\right)p_i\right] - p_0 \\ &= \sum_{i=0}^{\infty}\left[\left(\sum_{j=0}^{\infty}\left(j(j-1) + j(2i-1) + (i-1)^2\right)\frac{\rho^j}{j!} - i^2e^{\rho}\right)p_i\right] - p_0 \\ &= \sum_{i=0}^{\infty}\left[\left(\rho^2e^{\rho} + \left(2(i-1)+1\right)\rho e^{\rho} + (i-1)^2e^{\rho} - i^2e^{\rho}\right)p_i\right] - p_0 \end{aligned} \tag{A.14}$$

$$\begin{aligned} \dot{M}_2 &= \left(\rho^2 + 2\rho L - \rho - 2L + 1 - e^{-\rho}p_0\right)\mu \\ &= \left(\rho^2 - 2(1-\rho)L\right)\mu - \dot{L} \end{aligned} \tag{A.15}$$

This is the same as the M/M/1 formula (A.6) except for the first term. Integrating, again:

$$\begin{aligned} &V = V_0 + L_0(L_0+1) + 2(1-\rho)(L_e - D)\mu t - L(L+1) \qquad \text{where} \\ &D \equiv \frac{1}{t}\int_0^t L(y)dy \ , \qquad L_e = \frac{\rho^2}{2(1-\rho)} \end{aligned} \tag{A.16}$$

L_e is of course the equilibrium mean, but in principle we ought to prove this!

A.4 EQUILIBRIUM RECURRENCE RELATIONS OF THE M/D/1 QUEUE

Formulae for L_e and V_e can be found by evaluating the second and third moments of equilibrium probabilities respectively. Deriving them from a generating function may be more efficient, but in the main dissertation this results in a unit-in-service component. Exhaustive evaluation confirms the link between the recurrence relation formulation and the generally accepted result *without* unit-in-service.When the LHS and RHS of the time-dependent recurrence relations are equated, the following steady-state relations are obtained:

$$p_1 = \left(e^{\rho} - \rho - 1\right)p_0$$

$$p_i = \left(e^{\rho} - \rho\right)p_{i-1} - \sum_{j=2}^{i} \frac{\rho^j}{j!} p_{i-j} \quad (i > 1) \qquad \text{(A.17)}$$

When p_0 is added to p_1 a pattern appears:

$$\begin{aligned} p_0 + p_1 &= e^{\rho} p_0 - \rho p_0 \\ p_2 &= e^{\rho} p_1 - \rho p_1 - \frac{\rho^2}{2!} p_0 \\ p_3 &= e^{\rho} p_2 - \rho p_2 - \frac{\rho^2}{2!} p_1 - \frac{\rho^3}{3!} p_0 \end{aligned} \qquad \text{(A.18)}$$

Etc.

Some general results can now be stated,where M_k is the kth moment. Noting that p_1 appears as itself in any expression for a moment, if for some X:

$$M_k = X + p_1 \qquad \text{(A.19)}$$

Then

$$M_k + p_0 = X + \left(e^{\rho} - \rho\right)p_0 \qquad \text{(A.20)}$$

And for any power k:

$$\sum_{i=1}^{\infty}\sum_{j=1}^{i} (i-j)^k \frac{\rho_j}{j!} p_{i-j} = \left(e^{\rho} - 1\right)M_k$$

$$\sum_{i=1}^{\infty}\sum_{j=1}^{i} j(i-j)^k \frac{\rho_j}{j!} p_{i-j} = \rho e^{\rho} M_k \qquad \text{(A.21)}$$

$$\sum_{i=1}^{\infty}\sum_{j=1}^{i} j(j-1)(i-j)^k \frac{\rho_j}{j!} p_{i-j} = \rho^2 e^{\rho} M_k$$

Etc.

A.5 EQUILIBRIUM MEAN OF THE M/D/1 QUEUE

Taking the first moment simply recovers p_0, which is already available from (A.13):

$$p_0 = e^{\rho}(1-\rho) \tag{A.22}$$

Taking the second moment, expanding multipliers to match the probability indices:

$$M_2 + p_0 = \sum_{i=0}^{\infty} i^2 p_i + p_0 = \left(e^{\rho} - \rho\right)\sum_{i=1}^{\infty} i^2 p_{i-1} - \sum_{i=1}^{\infty}\sum_{j=2}^{i} i^2 \frac{\rho_j}{j!} p_{i-j}$$

$$= e^{\rho}\sum_{i=1}^{\infty} i^2 p_{i-1} - \sum_{i=1}^{\infty}\sum_{j=1}^{i} i^2 \frac{\rho_j}{j!} p_{i-j}$$

$$\begin{aligned} &= e^{\rho}\sum_{i=1}^{\infty}\left((i-1)^2 + 2(i-1) + 1\right)p_{i-1} \\ &- \sum_{i=1}^{\infty}\sum_{j=1}^{i}\left((i-j)^2 + 2j(i-j) + j^2\right)\frac{\rho_j}{j!} p_{i-j} \end{aligned} \tag{A.23}$$

$$M_2 + p_0 = e^{\rho}\left(M_2 + 2L_e + 1\right) - \left[\left(e^{\rho} - 1\right)M_2 + 2\rho e^{\rho} L_e + \rho^2 e^{\rho} + \rho e^{\rho}\right] \tag{A.24}$$

The M_2terms cancel, and after dividing through by e^{ρ}, and substituting for p_0 the M/D/1 mean, *without* unit-in-service, is left:

$$2(1-\rho)L_e = \rho^2 \qquad \text{or} \qquad L_e = \frac{\rho^2}{2(1-\rho)} \tag{A.25}$$

A.6 EQUILIBRIUM VARIANCE OF THE M/D/1 QUEUE

Taking the third moment:

$$M_3 + p_0 = \sum_{i=0}^{\infty} i^3 p_i + p_0 = \left(e^{\rho} - \rho\right)\sum_{i=1}^{\infty} i^3 p_{i-1} - \sum_{i=1}^{\infty}\sum_{j=2}^{i} i^3 \frac{\rho_j}{j!} p_{i-j}$$

$$= e^{\rho}\sum_{i=1}^{\infty} i^3 p_{i-1} - \sum_{i=1}^{\infty}\sum_{j=1}^{i} i^3 \frac{\rho_j}{j!} p_{i-j}$$

$$\begin{aligned}&= e^{\rho}\sum_{i=1}^{\infty}\left((i-1)^3+3(i-1)^2+3(i-1)+1\right)p_{i-1}\\&-\sum_{i=1}^{\infty}\sum_{j=1}^{i}\left((i-j)^3+3j(i-j)^2+3j^2(i-j)+j^3\right)\frac{\rho_j}{j!}p_{i-j}\end{aligned} \tag{A.26}$$

So:

$$\begin{aligned}&M_3+p_0=e^{\rho}\left(M_3+3M_2+3L_e+1\right)\\&-\left[\left(e^{\rho}-1\right)M_3+3\rho e^{\rho}M_2+3\rho^2e^{\rho}L_e+3\rho e^{\rho}L_e+\rho^3e^{\rho}+3\rho^2e^{\rho}+\rho e^{\rho}\right]\end{aligned} \tag{A.27}$$

The M_3 terms cancel, and using previous results for p_0 and L_e and dividing through by e^{ρ}:

$$3M_2+\frac{3\rho^2}{2(1-\rho)}+1-3\rho M_2-\frac{3\rho^4}{2(1-\rho)}-\frac{3\rho^3}{2(1-\rho)}-\rho^3-3\rho^2-\rho=1-\rho \tag{A.28}$$

Many terms cancel leaving:

$$M_2=\frac{\rho^4-\rho^3+3\rho^2}{6(1-\rho)^2} \tag{A.29}$$

Hence, the M/D/1 variance without unit-in-service is:

$$V_e=M_2-L_e^2=\frac{\rho^2\left(6-2\rho-\rho^2\right)}{12(1-\rho)^2} \tag{A.30}$$

A.7 SPECULATION ON THE GENERALITY AND MEANING OF THE VARIANCE RESULT

The deterministic variance formula derived for both M/M/1 and M/D/1 seems as universal as the deterministic mean formula, but in principle other queue processes could lead to variations that also satisfy the boundary conditions, for example through modification of the (1-ρ) factor or the form of the polynomial expression in *L*. In main Chapter 5, it is shown that the exact form of the deterministic variance formula, equation (5.4.13), as derived from the diffusion equation, depends on the step size between fractionally discrete states, although this is a relatively minor adjustment, and arises from a mathematical artifice, not a different physical process. The derivation nevertheless supports the generality of the deterministic variance formula.

Since moments are linear in probabilities, if a process whose mean is described by the P-K function with some statistical parameters could be represented by a weighted linear combination of M/M/1 and M/D/1 having the form of equations (4.3.2-3), as in equation (A.31) with the weighting factor α after eliminating the common factor ρ/(1-ρ), it could be inferred that the variance result also applies to it.

$$\alpha + \tfrac{1}{2}(1-\alpha)\rho \equiv \left(I + \tfrac{1}{2}(I_a - 1)\right) + (C - I)\rho) \qquad \text{(A.31)}$$

If this holds *for all* ρ, so that corresponding terms can be compared, while its value is independent of *I*, eliminating α between the resulting equations shows that *C* depends on *I*:

$$C = \frac{2I - I_a + 3}{4} \qquad \text{and} \qquad \alpha = 2(C-1) + I_a \qquad \text{(A.32)}$$

If *I* is taken to represent the choice between working with finite or infinitesimal service intervals, it can be only 0 or 1, which restricts *C* to two possible relationships with I_a, so unless the restriction on the values of *I* can be relaxed the arrival and service processes can no longer be independent. But it is not clear what relaxing this restriction would mean in terms of defining an actual process capable of producing the target distribution. So this approach seems artificial.

A more fundamental argument considers that, while the M/M/1 and M/D/1 calculations both arrive at the same deterministic formulae, as they must if the fornulae are general, there is no explicit account of randomness in the variance apart from the presence of L_e as the limiting value of *D*. Therefore, whatever random processes are involved, their effects apart from L_e must somehow be eliminated. In both M/M/1 and M/D/1 cases, evaluating the *second* moment of either the time-dependent or the equilibrium probability distribution yields the equilibrium *mean*, with the deterministic variance as a bonus in the first case. Therefore the deterministic variance calculation leaves the equilibrium variance indeterminate, just as the deterministic mean formula leaves the equilibrium mean indeterminate.

An (almost philosophical) question is then: if a queue process is such that analysis from first principles turns out to be intractable, its equilibrium moments being only emergent, as they could in principle be since an infinite number of infinitely long trials is required to determine them physically, how can it then lead to the deterministic formulae unless these are already guaranteed by a deeper set of principles? For *any* queue process, this is true of the deterministic mean queue formula since it represents conservation of queuing units. Conjecturing that this applies to higher moments (or sums of moments) may be a step too far, but it would be gratifying to find a physical quantity that is conserved by the variance.

The mean of a distribution is also the point on the state axis with the minimum probability-weighted sum of squared differences (PWSSD) from all possible states. As a queue develops from its initial to a final state this property applies along the whole path of the mean. The principle of least action says that any system follows the path of least action out of all possible paths. If the PWSSD is identified with action, then the path of the mean can be identified with the path of least action in the 'field' defined by the state probabilities, and the variance is a measure of this minimum action.This may be somewhat egregious because the probability distribution at each point depends on that at an earlier point, so the 'field' is not independent of the thing that 'moves' through it, but this could also be true of other systems.

In physics, action is the integral of the Lagrangian, which is the difference between kinetic and potential energy. The simpler form of the variance equation, equation (2.3.29), describes the time development of the sum of the first two raw moments. If the path of the mean is treated as determined then this sum could be analogous to action. However, what is the 'Lagrangian' in this case, and the connection with 'energy'? The hypothesised general differential formulae for the second moment in equations (A.6) and (A.15) generalise to:

$$\dot{M}_2 = \left(-\dot{L}\right) - 2(1-\rho)\mu\left(L - L_e\right) \qquad \text{(A.33)}$$

Speculatively, the first term on the RHS could be identified in some sense with 'kinetic energy', as the queue changes (gains 'energy' as it falls, loses as it rises), and the second term with 'potential energy', as the queue approaches or diverges from its equilibrium state which represents 'zero potential'. In (A.33), both kinds of 'energy' can have either sign. This could be a topic for future research.

APPENDIX B – M/M/1 PROBABILITY DISTRIBUTION ALGORITHM BASED ON SERIES FORMULA

```
Private Sub Series_Distribution(T As Double)
' Dim T As Double             ' Absolute or relative time

' Calculate the queue size probability distribution as given by Morse (1958)

' Public NP as Integer       ' Dimension of arrays = maximum resolution
' Public NQ as Integer       ' Half of NP, for one Sin & Cos quadrant
' Public NR As Integer       ' Resolution of calculation
' Public NN As Integer       ' One less than NR – i.e. maximum queue size
' Public N0 As Integer           ' Actual maximum queue size in distribution
' Public calcs As Long          ' Number of major operations performed
' Public Pe(NP) As Double       ' Equilibrium distribution
' Public Pm(NP) As Double       ' Base distribution
' Public Pn(NP) As Double       ' Calculated distribution
' Public srho(-NP To NP) As Double  ' Powers of sqrt(rho)
' Public Const Pi As Double = 3.14159265358979
' Public Const STOPON As Boolean = True
' Public Const STOPP As Double = 0.0000001    ' Minimum value of Pn

Dim i As Integer
Dim m As Integer
Dim n As Integer
Dim addexp As Double
Dim sumexp(NP) As Double
Dim sumP As Double
Dim Pmin As Double
Dim Pmax As Double
Dim force_zero As Boolean

force_zero = False
Pmin = 0
Pmax = 0
sumP = 0



' Initialise with equilibrium distribution
For n = 0 To NN
    Pn(n) = Pe(n)
Next n
```

```
' Add time-dependent terms
For n = 0 To NN
    N0 = n
    If force_zero = True Then
        Pn(n) = 0
        Exit For
    Else
        For m = 0 To NN
            If Pm(m) > 0 Then
               sumexp(n) = 0
               For i = 1 To NN
                   addexp = S(i, m) * S(i, n) * Exp(-X(i) * T) / X(i)
sumexp(n) = sumexp(n) + addexp
               Next i
               Pn(n) = Pn(n) + 2 * Pm(m) * srho(n - m) * sumexp(n) / NR
End If
        Next m
        sumP = sumP + Pn(n)
        If Pn(n) >= Pmax Then
            Pmax = Pn(n)
        Else
            Pmin = STOPP
        End If
        If STOPON And (n > 1 And Pn(n) < Pmin) Then force_zero = True
    End If
Next n

' Normalise probabilities
For n = 0 To N0
    Pn(n) = Pn(n) / sumP
Next n

End Sub
```

```
Private Function Sine(i As Long, NS As Integer) As Double

' Interpolate value of Sine in any quadrant, where i is relative to NS, representing
Pi radians
' N should be a power of two and not exceed 2 NQ

Dim nquadrant As Integer
Dim j As Long
Dim k As Integer
Dim l As Long
Dim qn As Integer

' Determine quadrant within one cycle and equivalent point within quadrant

' Following tricks avoid overflow
If NS >= NQ Then
qn = NS / NQ
    l = (2 * i) / qn
Else
qn = (2 * NQ) / NS
    l = i * qn
End If
k = Int(l / NQ)
j = l Mod NQ
k = k Mod 4

Select Case k
Case 0
    Sine = SI(j)
Case 1
Sine = SI(NQ - j)
Case 2
    Sine = -SI(j)
Case 3
    Sine = -SI(NQ - j)
End Select

End Function
```

```
Private Function Cosine(i As Long, NS As Integer) As Double

'  Interpolate  value  of  Cosine  in  any  quadrant,  where  i  is  relative  to  NS,
representing Pi radians. N should be a power of two and not exceed 2 NQ

Dim ic As Long

ic = i + NS / 2
Cosine = Sine(ic, NS)

End Function



Private Sub Fill_Sines()

' Fill in Sines at (public) NQ points between 0 and Pi/2
' NQ should ideally be a power of 2

Dim Piover2 As Double
Dim i As Integer

Piover2 = Pi / 2
SI(0) = 0
SI(NQ) = 1
For i = 1 To NQ - 1
    SI(i) = Sin((Piover2 * i) / NQ)
Next i
calcs = calcs + NQ - 1

End Sub
```

APPENDIX C – MARKOV PROBABILITY DISTRIBUTION ALGORITHM AND EXAMPLES

```
Sub Markov_Distribution(rho As Double, T As Double, step As Double, NN As Integer, r
As Integer, m as Integer, G As Integer, Pm As Double)

‘ Dim rho As Double                     ' Demand intensity
‘ Dim mu As Double                      ' Capacity rate parameter
' Dim T As Double                      ' Absolute or relative time
' Dim step as Double                   ' Iterative time step
' Dim NN As Integer                    ' Maximum queue size
‘ Dim r As Integer                      ‘ Erlang arrivals parameter
‘ Dim m As Integer                      ‘ Erlang service parameter
‘ Dim G As Integer                      ‘ Process / Green period throughput
‘ Dim Pm(0:NN) As Double                ‘ Initial distribution/state proxy

Dim exprho, tt, s, sumP As Double
Dim i, j As Integer
Dim Pn(-100:NN), dP(-100:NN), Pp(0:NN), Poisson(0:NN) As Double

‘ G should be >= 1 for M/D/1 and zero otherwise
‘ Bulk parameter is negative of Erlang parameter in reversed position
erho = rho * Max(-r,1) / Max(-m,1)

' Poisson terms for M / D / 1
exprho = Exp(-G*rho)
Poisson(0) = 1
For j = 1 To NR
    Poisson(j) = (G*rho)^j * exprho / j
Next j

tt = 0
' Initial distribution
For i = -G To NN
    Pn(i) = Pm(i)
Next i
Pn(NN + 1) = 0



‘ Each time step

While tt <= T

    N0 = 0
```

```
‘ Each queue state in the distribution
    For i = -G To NN



‘ Process to be simulated
       If G = 0 Then
‘ Calculation of new probabilities at each time step



' Mm / Mr / 1 or Er / Em / 1 (only one of r and m may be ≠ 1)

            If i < NN

              dP(i) = Max(m,1) * Pn(1+Abs(r)) – Max(r,1) * erho * Pn(i)
              If i = 0 Then
                   For j = 1 to –(r+1)
                    dP(i) = dP(i) + Pn(j)
                   Next j
                Endif
                If i >= Max(r,1) Then
..               dP(i) = dP(i) – Max(m,1) * Pn(i)
                If i >= Abs(m) Then
..               dP(i) = dP(i) + Max(r,1) * erho * Pn(i-Abs(m))

            Else

              dP(i) = - Max(m,1) * Pn(i) + Max(r,1) * erho * Pn(i - m)

            End If
           dP(i) = dP(i) * step * mu

        Else



' M / D / 1

            s = 0
            If i = 0 Then s = Pn(0)
            For j = 0 To i + G
                s = s + Poisson(j) * Pn(i + G - j)
            Next j
            dP(i) = (s * exprho - Pn(i)) * step * mu
        End If

        N0 = i
        If Pn(i) + dP(i) < 0.0000000001 Then Exit For

    Next i

‘ Note that range of loops can be limited if higher probabilities are zero
‘ For M / D / 1 processes notional negative states are absorbed into p0

    For i = 1 to -G
       dp(0) = dp(0) + dp(i)
    Next i

‘ Changes are accumulated to absolute probabilities that are normalized

    sumP = 0
    For i = 0 To N0
Pn(i) = Pn(i) + dP(i)
        sumP = sumP + Pn(i)
Next i
    If sumP > 0 Then
        For i = 0 To N0
            Pn(i) = Pn(i) / sumP
        Next i
    End If


‘ For Erlang processes convert ‘stages’ probabilities to ‘customers’

   If r > 1 or m > 1 Then
       For i = 0 to NN
          Pp(i) = Pn(i)
          Pn(i) = 0
       Next i
   Endif

   If r > 1 Then

       For i = 0 to NN/r + 1
          For j = i*r to (i+1)*r-1
              Pn(i) = Pn(i) + Pp(j)
          Next j
       Next i
```

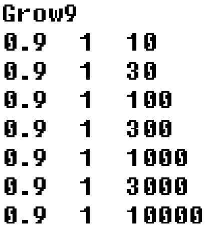
```
Else if m > 1 Then

       For i = 0 to NN/m + 1
          For j = Max((i-1)*m+1,0) to i*m
              Pn(i) = Pn(i) + Pp(j)
          Next j
       Next i

   Endif

‘ Finishing calculations for this time step

    tt = tt + step

Wend

End Sub
```

Following a listing of Qsim parameters and options, and a typical data file, examples of the output of Qsim are given below, for growth of a queue starting from zero with ρ=0.9 measured at several points, and various processes. The first example (M/M/1) recalculates each time point from zero, the next and subsequent calculate in time slices starting with the probability distribution from the previous time slice. Results of the two methods are very close. The ‘mean’ and ‘variance’ are calculated from the probability distribution, ‘delay(s)’ is calculated by averaging the mean queue during the calculation step, ‘delay(t)’ relates to the time from zero, and ‘delay(v)’ is back-calculated from the variance using the variance formula. Mean and variance of (average) p_0 are practically relevant only to M/D/1[G>1].

```
Program Qsim version 15/05/12, run on 12/08/12
General command form: Qsim<data> [/output] [=<rho>] <process><step><init_q>
<data_file> default extension is .DAT
<process> = E.E.[][.G] where 0 reads as infinity so:
 1.1 is M/M/1, 1.0 is M/D/1, 1.0.G is M/D/1[G]
 E values not both >1. Numerical parameters in range 0-99
<step_size> is 1 by default, but may need to be reduced
<initial_queue> exact value used once, then reset to zero
 if <0 then pure state, if >=0 then equilibriated
 if fractional "m.s" then Normal p.d: mean=m, sd=m*(.s)
 Step size = seconds between evaluations of recurrence
 relations, and may need to be reduced.
 .DAT format: Title_line / {rho mu t} .. :
 rho(+continue previous distribution, -reset to zero queue)
 mu (+as given, -multiplied by relaxation time)
 t (+absolute time, -slice length)
```

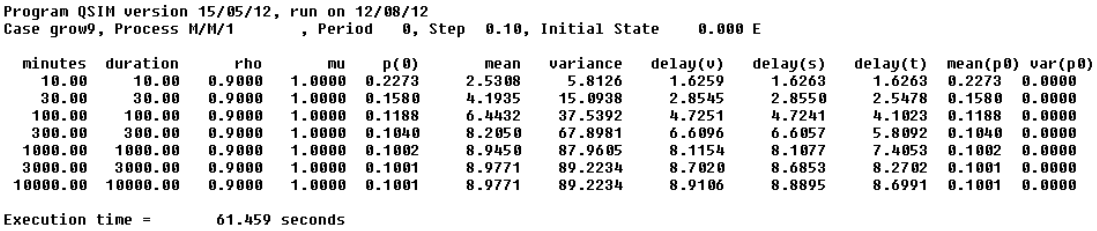
Figure C.1 Qsim condensed help listing

Qsim grow9.dat 1.1 0.1 0 [data file, M/M/1, step=0.1, initial queue=0]



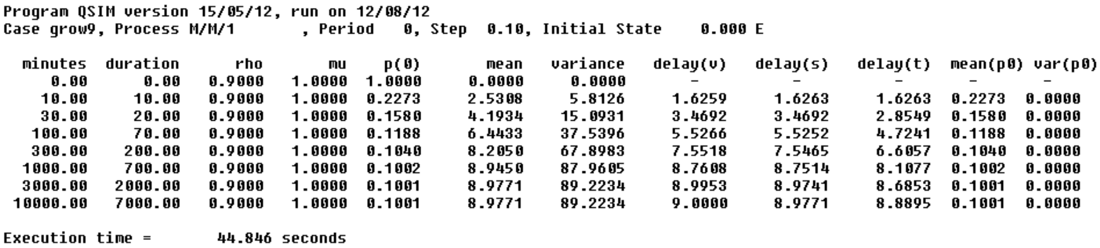
```
Grow9
0.9  1  10
0.9  1  30
0.9  1  100
0.9  1  300
0.9  1  1000
0.9  1  3000
0.9  1  10000
```

Figure C.2 Qsim command and data file for simple growth problem



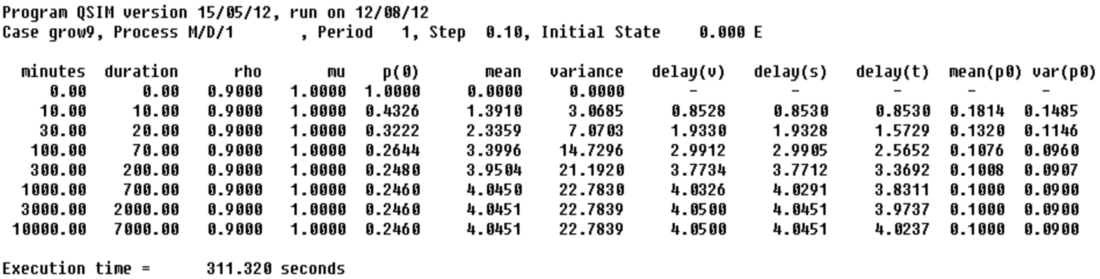
```
Program QSIM version 15/05/12, run on 12/08/12
Case grow9, Process M/M/1       , Period   0, Step  0.10, Initial State    0.000 E

  minutes  duration       rho        mu    p(0)       mean   variance   delay(v)   delay(s)   delay(t)  mean(p0) var(p0)
    10.00     10.00    0.9000    1.0000  0.2273     2.5308     5.8126     1.6259     1.6263     1.6263  0.2273  0.0000
    30.00     30.00    0.9000    1.0000  0.1580     4.1935    15.0938     2.8545     2.8550     2.5478  0.1580  0.0000
   100.00    100.00    0.9000    1.0000  0.1188     6.4432    37.5392     4.7251     4.7241     4.1023  0.1188  0.0000
   300.00    300.00    0.9000    1.0000  0.1040     8.2050    67.8981     6.6096     6.6057     5.8092  0.1040  0.0000
  1000.00   1000.00    0.9000    1.0000  0.1002     8.9450    87.9605     8.1154     8.1077     7.4053  0.1002  0.0000
  3000.00   3000.00    0.9000    1.0000  0.1001     8.9771    89.2234     8.7020     8.6853     8.2702  0.1001  0.0000
 10000.00  10000.00    0.9000    1.0000  0.1001     8.9771    89.2234     8.9106     8.8895     8.6991  0.1001  0.0000

Execution time =       61.459 seconds
```

Figure C.3 Markov calculation of M/M/1 growth from zero over increasing time periods



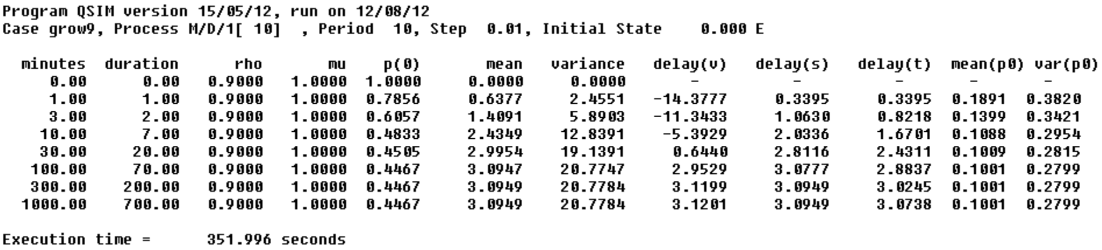
```
Program QSIM version 15/05/12, run on 12/08/12
Case grow9, Process M/M/1       , Period   0, Step  0.10, Initial State    0.000 E

  minutes  duration      rho       mu    p(0)       mean   variance   delay(v)   delay(s)   delay(t)  mean(p0) var(p0)
     0.00      0.00   0.9000   1.0000  1.0000     0.0000     0.0000         -          -          -        -       -
    10.00     10.00   0.9000   1.0000  0.2273     2.5308     5.8126     1.6259     1.6263     1.6263   0.2273  0.0000
    30.00     20.00   0.9000   1.0000  0.1580     4.1934    15.0931     3.4692     3.4692     2.8549   0.1580  0.0000
   100.00     70.00   0.9000   1.0000  0.1188     6.4433    37.5396     5.5266     5.5252     4.7241   0.1188  0.0000
   300.00    200.00   0.9000   1.0000  0.1040     8.2050    67.8983     7.5518     7.5465     6.6057   0.1040  0.0000
  1000.00    700.00   0.9000   1.0000  0.1002     8.9450    87.9605     8.7608     8.7514     8.1077   0.1002  0.0000
  3000.00   2000.00   0.9000   1.0000  0.1001     8.9771    89.2234     8.9953     8.9741     8.6853   0.1001  0.0000
 10000.00   7000.00   0.9000   1.0000  0.1001     8.9771    89.2234     9.0000     8.9771     8.8895   0.1001  0.0000

Execution time =       44.846 seconds
```

Figure C.4 M/M/1 growth in successive time slices of increasing duration, compare C.3



```
Program QSIM version 15/05/12, run on 12/08/12
Case grow9, Process M/D/1       , Period   1, Step  0.10, Initial State    0.000 E

  minutes  duration      rho       mu    p(0)       mean   variance   delay(v)   delay(s)   delay(t)  mean(p0) var(p0)
     0.00      0.00   0.9000   1.0000  1.0000     0.0000     0.0000         -          -          -        -       -
    10.00     10.00   0.9000   1.0000  0.4326     1.3910     3.0685     0.8528     0.8530     0.8530   0.1814  0.1485
    30.00     20.00   0.9000   1.0000  0.3222     2.3359     7.0703     1.9330     1.9328     1.5729   0.1320  0.1146
   100.00     70.00   0.9000   1.0000  0.2644     3.3996    14.7296     2.9912     2.9905     2.5652   0.1076  0.0960
   300.00    200.00   0.9000   1.0000  0.2480     3.9504    21.1920     3.7734     3.7712     3.3692   0.1008  0.0907
  1000.00    700.00   0.9000   1.0000  0.2460     4.0450    22.7830     4.0326     4.0291     3.8311   0.1000  0.0900
  3000.00   2000.00   0.9000   1.0000  0.2460     4.0451    22.7839     4.0500     4.0451     3.9737   0.1000  0.0900
 10000.00   7000.00   0.9000   1.0000  0.2460     4.0451    22.7839     4.0500     4.0451     4.0237   0.1000  0.0900

Execution time =      311.320 seconds
```

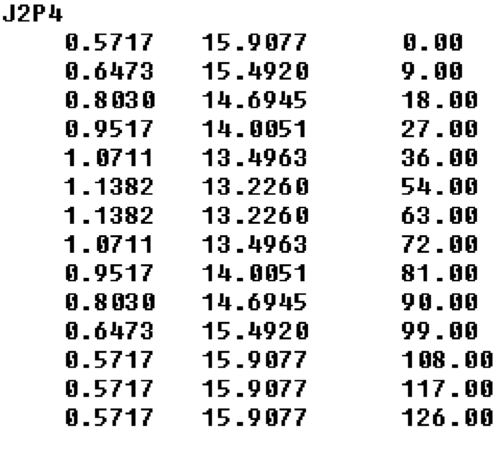
Figure C.5 Markov calculation of M/D/1 queue growth, note larger p_0, smaller L, same $p_{0(\text{ave})}$



```
Program QSIM version 15/05/12, run on 12/08/12
Case grow9, Process M/D/1[ 10]  , Period  10, Step  0.01, Initial State    0.000 E

  minutes  duration      rho       mu    p(0)       mean   variance   delay(v)   delay(s)   delay(t)  mean(p0) var(p0)
     0.00      0.00   0.9000   1.0000  1.0000     0.0000     0.0000         -          -          -        -       -
     1.00      1.00   0.9000   1.0000  0.7856     0.6377     2.4551   -14.3777     0.3395     0.3395   0.1891  0.3820
     3.00      2.00   0.9000   1.0000  0.6057     1.4091     5.8903   -11.3433     1.0630     0.8218   0.1399  0.3421
    10.00      7.00   0.9000   1.0000  0.4833     2.4349    12.8391    -5.3929     2.0336     1.6701   0.1088  0.2954
    30.00     20.00   0.9000   1.0000  0.4505     2.9954    19.1391     0.6440     2.8116     2.4311   0.1009  0.2815
   100.00     70.00   0.9000   1.0000  0.4467     3.0947    20.7747     2.9529     3.0777     2.8837   0.1001  0.2799
   300.00    200.00   0.9000   1.0000  0.4467     3.0949    20.7784     3.1199     3.0949     3.0245   0.1001  0.2799
  1000.00    700.00   0.9000   1.0000  0.4467     3.0949    20.7784     3.1201     3.0949     3.0738   0.1001  0.2799

Execution time =      351.996 seconds
```

Figure C.6 Markov calculation of M/D/1 [G=10] growth. Time points and step size must be reduced by factor of G to compensate for lengthened service intervals.

Note even larger p_0, smaller L, but again same $p_{0(\text{ave})}$ (delay back-calculated from variance is no longer valid)



```
J2P4
     0.5717    15.9077       0.00
     0.6473    15.4920       9.00
     0.8030    14.6945      18.00
     0.9517    14.0051      27.00
     1.0711    13.4963      36.00
     1.1382    13.2260      54.00
     1.1382    13.2260      63.00
     1.0711    13.4963      72.00
     0.9517    14.0051      81.00
     0.8030    14.6945      90.00
     0.6473    15.4920      99.00
     0.5717    15.9077     108.00
     0.5717    15.9077     117.00
     0.5717    15.9077     126.00
```

Figure C.7 Data for a peak case J2P4, (ρ, μ, t),where the first time slice is equilibrated

In the following three figures, the distributions settle as time progresses. The last two figures have not been smoothed to avoid undershoot.

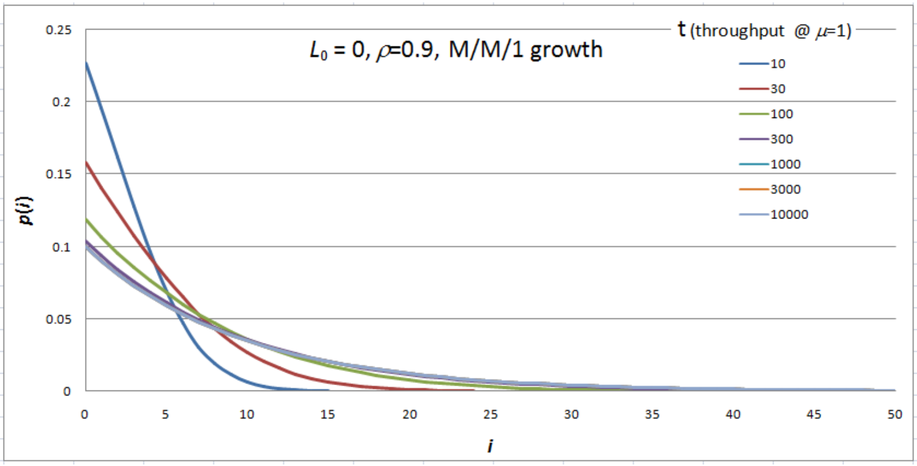


Figure C.8 Development of Markov-simulated distributions for M/M/1 queue growth

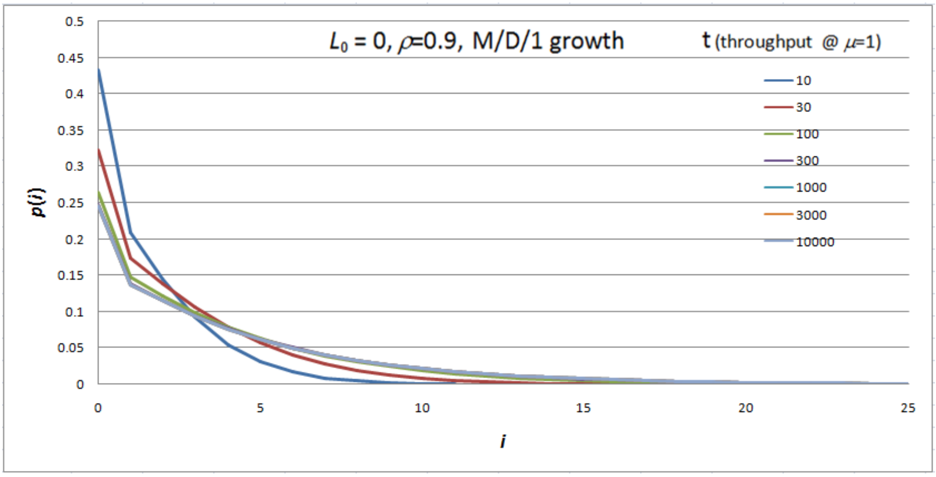


Figure C.9 Development of Markov-simulated distributions for M/D/1 queue growth

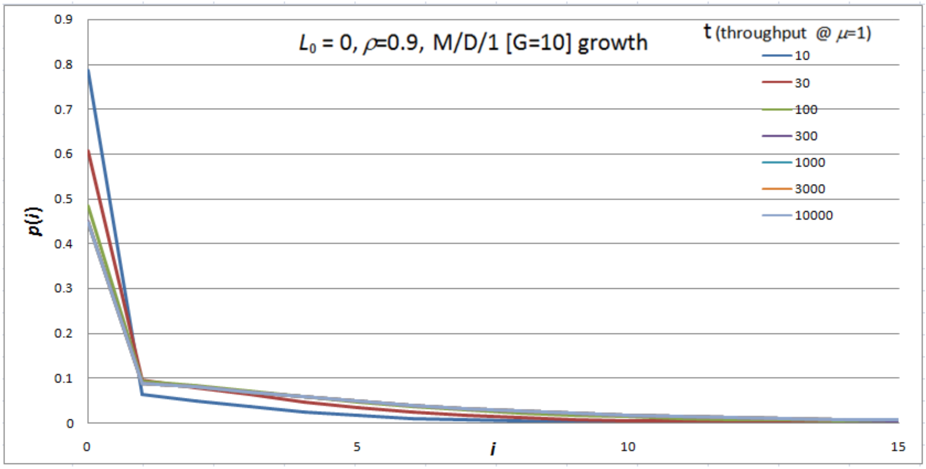


Figure C.10 Development of Markov-simulated distributions for M/D/1 [G=10] growth

APPENDIX D – HEIDEMANN'S SIGNAL QUEUE FORMULA

This Appendix compares a result obtained by Heidemann (1994) for the queue size at a signal with a one-lane approach that draws on work of other authors including Meissl (1963), with that of Webster and Cobbe (1966). Heidemann's formula is time-dependent and claimed to be exact for any green period length, whereas as explained Chapter 2 of the main text Webster's stochastic term applies only to the nominal case where each service period can serve only one customer (M/D/1). However, it is asserted that Webster's empirical correction term, which depends on the service period throughput capacity *G*, compensates for this. The objective here is to show that when the M/D/1[*G*] mean is inserted into Heidemann's result under equilibrium, the match to Webster's result is fair. Because the notation used by Heidemann differs from that used in this Dissertation and one symbol overlaps, a glossary is provided. Equations given by Heidemann (1994) are recognised by an additional "(H.-)" marker.

Table D.1 Glossary of variables

Quantity represented	**Heidemann's symbol**	**Local symbol**
Service interval	T	$1/s$
Saturation flow	$1/T$	s
Cycle time	T_{Cy}	c
Red period	μT	r
Green period	νT	g
Red period input capacity	μ	rs
Green period throughput capacity	ν	$gs=\mu c=G$
Green proportion	$\nu/(\mu+\nu)$	$\Lambda=g/c$
Probability Generating Function	$P(y)$	$P(y)$
Mean arrival rate	$f=P'(1)/T$	λ
Capacity	$C=\mu/((\mu+\nu)T)$	$\mu=\Lambda s$
Degree of saturation	f/C	$x=\lambda/\mu$
Mean arrivals in one service interval	$P'(1)$	$\lambda/s=\Lambda x$
Ave. overflow queue at start of red	$R'(1)$	$L_{\nu[G]}$

The Probability Generating Function of the queue size is:

$$P(y)=\sum_{i=0}^{\infty} p_i y^i \qquad \text{where} \qquad \text{(D.1)}$$

p_i = probability of i arrivals in a time interval that equals one service time

The queue size given by Meissl (1963) is quoted by Heidemann as his equation (17):

$$L = \frac{1}{1-P'(1)} . \left(\frac{R'(1).\mu}{\mu+\nu} + \frac{P'(1).\mu.(\mu+1)}{2.(\mu+\nu)} + \frac{P''(1)}{2} \right) \qquad \text{(D.2=H.17)}$$

To evaluate this, Heidemann first determines the derivative of $R(1)$ as his equation (25):

$$R'(1) = \sum_{k=1}^{\nu-1} \frac{\exp(P'(1).(x_k-1)) - P'(1).x_k}{\exp(P'(1).(x_k-1)) - x_k} - \frac{\nu.(\nu-1) - [P'(1).(\mu+\nu)]^2}{2.[\nu - P'(1).(\mu+\nu)]} \qquad \text{(D.3=H.25)}$$

where the $\{x_k\}$ are the (ν-1) zeros lying within the complex unit circle of the function:

$$\phi(x) = x^{\nu} - \exp[P'(1).(\mu+\nu).(x-1)] \qquad \text{(D.4=H.~)}$$

It is assumed that arrivals are Poisson distributed (exponentially distributed intervals):

$$p_i = \frac{(\Lambda x)}{i!} e^{-\Lambda x} \qquad \text{(D.5)}$$

so, as $P'(1)$ is the derivative of $P(y)$ evaluated at y=1:

$$P'(1) = \Lambda x, \qquad P''(1) = (\Lambda x)^2 \qquad \text{(D.6)}$$

Using equations (D.5) and writing $L_{V[G]}$ for the stochastic ('overflow') queue $R'(1)$, Heidemann's expression for the mean steady-state queue (D.2) translates to:

$$L = \frac{\left[2(1-\Lambda)L_{V[G]} + x\mu c(1-\Lambda)\left(1-\Lambda\left(1-\frac{1}{G}\right)\right) + (\Lambda x)^2 \right]}{2(1-\Lambda x)} \qquad \text{(D.7)}$$

where $L_{V[G]}$ is the stochastic queue component, and the Webster-Cobbe phase component can be discerned in the middle of the expression. While G and μc are actually identical, both are used in (D.7) in order to retain the original form of each term. To evaluate the stochastic term it is necessary to translate equation (D.3), in which:

$$P'(1).(\mu+\nu) \to \Lambda x.cs = \Lambda x.G\frac{c}{g} = Gx \qquad \text{(D.8)}$$

so $$R'(1) \to L_{V[G]} = \sum_{j=1}^{G-1} \frac{\exp(\Lambda x(y_j - 1)) - \Lambda x y_j}{\exp(\Lambda x(y_j - 1)) - y_j} + \frac{Gx^2 - (G-1)}{2(1-x)} \qquad \text{(D.9)}$$

where the $\{y_j\}$ are the roots of: $$y^G = \exp(Gx(y-1)) \qquad \text{(D.10)}$$

If G=1, the sum and the final element of (D.9) vanish, leaving the usual M/D/1 mean steady-state queue formula. In general L_V depends on G as one would expect, but in a way that does not appear convenient to evaluate. The middle term in (D.7) is recognisable as the phase queue in Webster's formula, equation (3.7.1) in main text, with a modification that vanishes when the absolute green period capacity is very large. However, a peculiarity of (D.7) is that the stochastic component is modified by a factor that does not depend explicitly on G.

The problem is to decide which modifications should belong to individual terms and which should be corrections to the whole formula. L_V may be taken as given, but there is no obvious reason why the phase term, being derivable deterministically, should differ from Webster's. The modifications to it and the $L_{V[G]}$ term and the last term of (D.7) may therefore be identified with Webster's correction term. Equation (D.7) can now be rearranged into the traditional Webster form, with a modified correction term involving the stochastic queue:

$$L = L_P + L_{V[G]} + L_H \quad \text{where} \quad L_H = \frac{\Lambda x - \Lambda(1-x)(2L_{V[G]} + \Lambda x)}{2(1-\Lambda x)} \qquad \text{(D.11)}$$

Figure D.1 shows fair agreement between (D.11) and Webster's formula, using the approximate formulae for $L_{V[G]}$ developed in Chapter 3. The cases are defined in Table D.2.

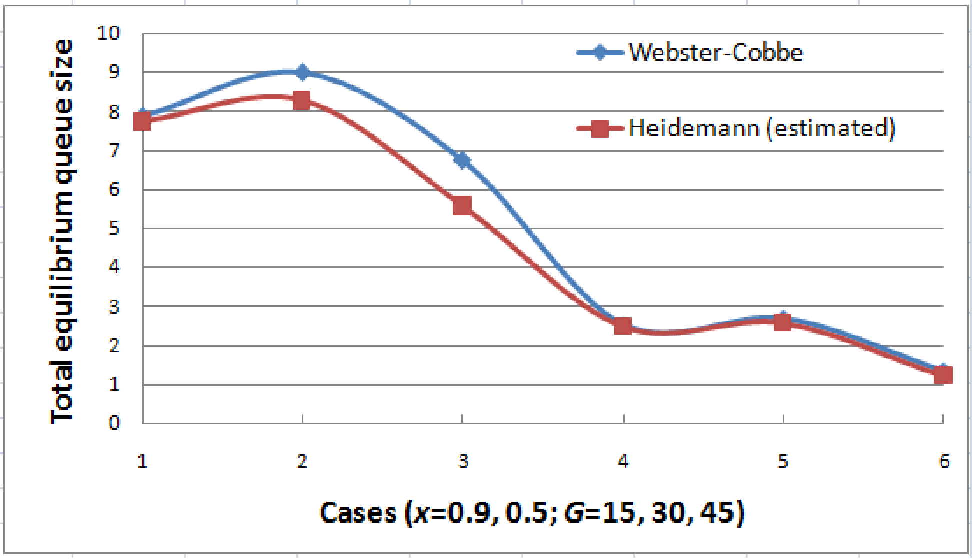


Figure D.1 Comparison with Webster-Cobbe queue for cases based on Heidemann's Figs. 3 and 4, using equation (D.11) and estimated formula for M/D/1[G] mean queue

Table D.2 Parameters used to generate cases in Figure D.1

Parameter/Case	**1**	**2**	**3**	**4**	**5**	**6**
G	15	30	45	15	30	45
x	0.9	0.9	0.9	0.5	0.5	0.5
Λ	0.25	0.5	0.75	0.25	0.5	0.75

The slight rise in the graphs in Figure D.1 at *G*=30 can be traced to the artificial manoeuvre of keeping *x* constant in each group of cases.

A feature of equation (D.7) is that the phase and stochastic components appear to get ‘mixed up’, but a different interpretation may be possible. In view of the derivation, it is tempting to suggest that the whole middle term of (D.7) be taken as the phase queue, reducing to the Webster-Cobbe form when the green capacity is large. Calling this modified term L^*_P, the correction term takes a more suggestive form:

$$L = L^*_P + L_{V[G]} + L^*_H \qquad \text{where} \quad L^*_H = \frac{(\Lambda x)^2}{2(1-\Lambda x)} - \frac{(1-x)\Lambda L_{V[G]}}{(1-\Lambda x)} \qquad \text{(D.12)}$$

In equation (D.12), the M/D/1 form has reappeared in the component L^*_H but with the average degree of saturation *x* replaced by the true saturation level in the green period, during which the actual capacity is not μ but *s*. However, the physical meaning of the second term in L^*_H is obscure, and the separation into three components is itself artificial.

Local references also in main references

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

Meissl P (1963). Zufallsmodell einer Lichtsignalanlage mit mehrspurigem Stauraum. *Mathematik-Technik-Wirtschaft***,** Heft l/63, 1-4, and Heft 2/63, 63-68, Wien.

Webster F V and Cobbe B M (1966). Traffic signals. *Road Research Technical Paper 56*. HMSO.

APPENDIX E – COMPARISON OF SOME ALTERNATIVE TIME-DEPENDENT QUEUE/DELAY FORMULATIONS

In the following, the notation used in the main body of the dissertation is used, so formulae of other authors have been ‘translated’ where necessary.

E.1 CATLING’S FORMULATION OF SHEARING

Catling (1977) approaches the problem from the point of view of delay to individuals. Delay is related to queue size by Little’s formula (E.1), where *x* is average degree of saturation, and equilibrium delay (E.2) is a case of the Pollaczek-Khinchin equilibrium mean formula:

$$d = \frac{L}{\mu x} \qquad \text{`} \tag{E.1}$$

$$d_e = I + \frac{C\rho}{1-\rho} \tag{E.2}$$

The time-dependent ‘coordinate transformed’ function of Doherty (1977) is then given as a quadratic solution, which when converted back into the quadratic and rearranged into the quasi-static pattern has the form:

$$(d-I)\left(1+\frac{2(\mu(d-I)+C)}{(1-\rho)\mu t}\right) = d_e - I \tag{E.3}$$

This clearly gives the right result as *t*→∞. It appears to blow up at *t*=0, but in fact the solution requires *d*=*I* at this point, so the first bracket shrinking to zero and the second bracket tending to infinity combine to produce the finite RHS. The degree of saturation ‘*x*’ in the original has been interpreted as ρ here because that is understood to be the intention. The formulation does not tackle or resolve the ambiguity between the demand intensity ρ and the degree of saturation of service *x*. This is germane if converting from delay to queue size, because Little’s formula involves the degree of saturation of service, not the demand intensity.

Another important difference from the sheared queue as presented in main Chapter 3 is that equations (E.2-3) include delay incurred during the run-out period. While individual delay is a true out-turn value its relationship to queue size, which is statically measurable at any instant, depends on the behaviour of capacity. The same queue size will result in different individual delays if the capacity changes before the queue is fully discharged.

However, assume that the capacity remains constant. Catling points out that the queue function has to be evaluated by equating the average individual delay to the integral under its envelope as shown in Figure E.1, leading to a somewhat intractable expression.

Catling proposes an alternative by shearing the queue itself, getting a result identical to L_s in equations (4.2.8), but with L_0=0 and I=0. He then goes on to derive expressions for the delay to individual arrivals when the arrival rate varies. Whether capacity is allowed to vary is unclear since the traffic pattern is described in terms of degrees of saturation, interpreted here as demand intensities.

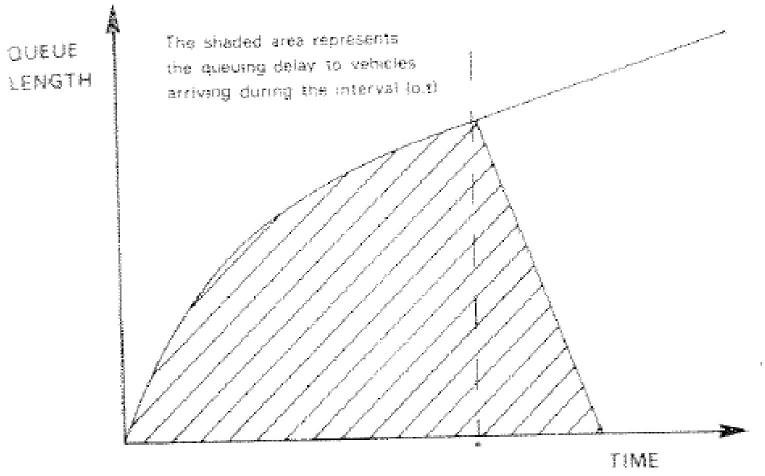

QUEUE
LENGTH
The shaded area represents
the queuing delay to vehicles
arriving during the interval (o,t)
TIME

Figure E.1 Original Figure 2 from Catling (1977), relating to Doherty's method

The analysis runs up against the same complication of run-out periods noted above. In computer implementations like ARCADY, PICADY and OSCADY for junctions and CONTRAM for networks (see main References), this complication is avoided by calculating average individual delay either by dividing total population delay by total throughput, or by following each individual (or 'packet' of notional individuals) through the queue discharge process. **However, it is apparent that Catling developed the sheared queue formula virtually simultaneously with and independently of Whiting, Kimber and Hollis.**

E.2 VOLUME-DELAY AND CONICAL FUNCTIONS

Volume/delay or volume/travel-time functions can take various forms of varying complexity, but one of the simplest is the US Bureau of Public Roads (BPR) formula for travel time when the demand intensity is ρ and t_0 is free-flow travel time:

$$t = t_0\left(1 + \beta\rho^{\alpha}\right) \tag{E.4}$$

In the formula, β is 1 in the case of BPR, but could have some other calibrated value (e.g. as in the case of the SATURN software). Volume/delay functions are not time-dependent and hence of limited interest in this context. However, where they are used it is recognised that the power can cause escalation of delay if demand exceeds capacity substantially. A way to moderate this is to use a conical function. This defines a hyperbolic function on axes derived from a conic section, as sketched in Figure E.2, which looks remarkably similar to shearing (see main Chapter 4). The hyperbolic function before transformation of the axes is 1/(1-ρ), which is not dissimilar from the M/M/1 equilibrium mean queue.

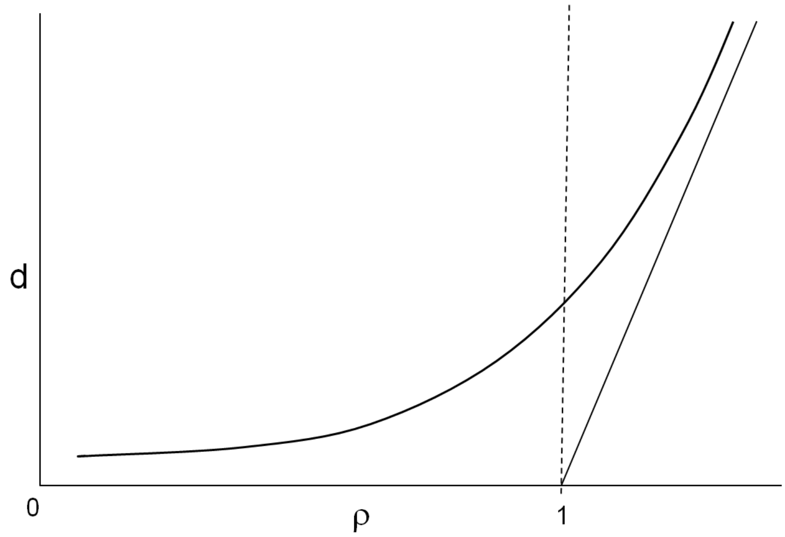


Figure E.2 Sketch of conical volume-delay function

The delay function is now asymptotic to a linear relationship when ρ>> 1, and its formula is a quadratic (Spiess 1990):

$$t = t_0\left[2 - \beta - \alpha(1-\rho) + \sqrt{\alpha^2(1-\rho)^2 + \beta^2}\right] \quad \text{where} \quad \beta = \frac{2\alpha - 1}{2\alpha - 2} \qquad \text{(E.5)}$$

Time-dependent shearing takes this a step further, not only by using a physically meaningful equilibrium queue function but by making the asymptote time-dependent to reflect the deterministic conservation constraint.

Equation (E.5) has only one parameter α, which determines both the slope of the asymptote and the sharpness of the curve. This may be an advantage when calibrating the function to empirical data or simulation outputs, so it can be used as a component of a smooth, differentiable objective function to be minimised for an entire traffic pattern. However, it does not have room for time-dependence.

E.3 THE HIGHWAY CAPACITY MANUAL DELAY FUNCTION

The US HCM/Canadian delay function (Cheng *et al* 2003, Akgüngör and Bullen 2007), equation (E.6), is time-dependent, and consistent with the M/D/1 equilibrium queue value (E.7) if C=0.5 and the upstream filtering factor ϕ is given its default value of 1:

$$d = \frac{t}{4}\left(-(1-\rho) + sign(1-\rho)\sqrt{(1-\rho)^2 + \frac{8C\phi\rho}{\mu t}} \right) \qquad \text{(E.6)}$$

$$L_e = \frac{C\rho^2}{1-\rho} \qquad \text{(since } d_e \equiv L_e\text{, assuming } \phi=1) \qquad \text{(E.7)}$$

It is not certain that this is the intention since C is described as an ‘incremental delay factor dependent on actuated control settings’. On the other hand, values > 0.5 make the equilibrium queue excessively large. Note that in (E.6) and similar equations here the factor *sign*(1-ρ) is applied to the (positive) square-root to give sensible asymptotic results when $t \to \infty$. This is not normally found in the references. However, in the case ρ>1 the formula predicts:

$$d \to \tfrac{1}{2}(\rho - 1)t \qquad \text{(E.8)}$$

which is wrong because the factor ½ should apply to average delay over the growth period, not to delay expected by the current arrival. As in Section E.1, no distinction is made between demand intensity and degree of saturation. There is a problem in how to translate (E.6) to queue length, as Little’s formula requires utilisation, which in general is not identical to ρ.
The Federal Highway Administration quotes the formula (FHWA):

$$L = \frac{\mu t}{4}\left(-(1-\rho) + sign(1-\rho)\sqrt{(1-\rho)^2 + Z} \right) \qquad \text{(E.9)}$$

where Z is a ‘composite factor’ for whose definition the reader is referred elsewhere. This appears to assume that utilisation equals 1, which is realistic only for ‘heavy traffic’. Akçelik (2001) quotes an interpretation that appears to be an extension of the HCM formula as implemented in his SIDRA software (see also next). Z in its simplest form can broadly be identified with the last term in (E.6), but this no longer gives the M/D/1 equilibrium queue. Either the term or the whole formula needs to be multiplied by ρ to give that result.

E.4 THE AUSTRALIAN FORMULATION

Akçelik (1998a) considers a signal queue, that can be decomposed into a red/green phase queue and a stochastic (random and oversaturation) queue, as described in main Chapter 2. He formulates the stochastic queue in such a way that it incorporates a term involving the absolute green period, addressed empirically in main Chapter 3. Akçelik's formula is:

$$L = \frac{\mu t}{4}\left(-(1-\rho) + sign(1-\rho)\sqrt{(1-\rho)^2 + \frac{12(\rho-\rho_0)}{\mu t}} \right) \text{(if } \rho > \rho_0\text{)} \qquad \text{(E.10)}$$

$$\rho_0 = 0.67 + \frac{G}{600} \qquad \text{where } G = \text{saturation flow * green time} \qquad \text{(E.11)}$$

Again '*x*' has been translated into ρ. This predicts no stochastic queue if ρ is less than a certain positive value. Rearranging (E.10) as (E.12) and expanding the square root to first order, it is easier to see that when $\rho<1$ and $t\to\infty$, the equilibrium queue is given by (E.13):

$$L = \frac{(1-\rho)\mu t}{4}\left(-1 + sign(1-\rho)\sqrt{1 + \frac{12(\rho-\rho_0)}{(1-\rho)^2 \mu t}} \right) \quad \text{(if } \rho > \rho_0\text{, else zero)} \qquad \text{(E.12)}$$

$$L_e = \frac{1.5(\rho-\rho_0)}{1-\rho} \qquad \text{(E.13)}$$

Although the approach is simpler than M/D/1[G], the empirical lower limit on demand intensity, equation (E.11), represents a drawback in our view. The results of the empirical M/D/1[G] Section in main Chapter 3 can be used to compare equilibrium queue predictions. As shown in Figure E.3, the results are significantly different for smaller values of ρ and *G*, although the queues there are so small this may not matter much in practice.

Akgüngör and Bullen (2007) summarise how Akçelik (1980) generates a smooth sheared delay function taking into account the average uniform (phase) queue acting as a lower bound, the situation where Bin Han (1996) finds an issue in producing a continuously differentiable combined function.

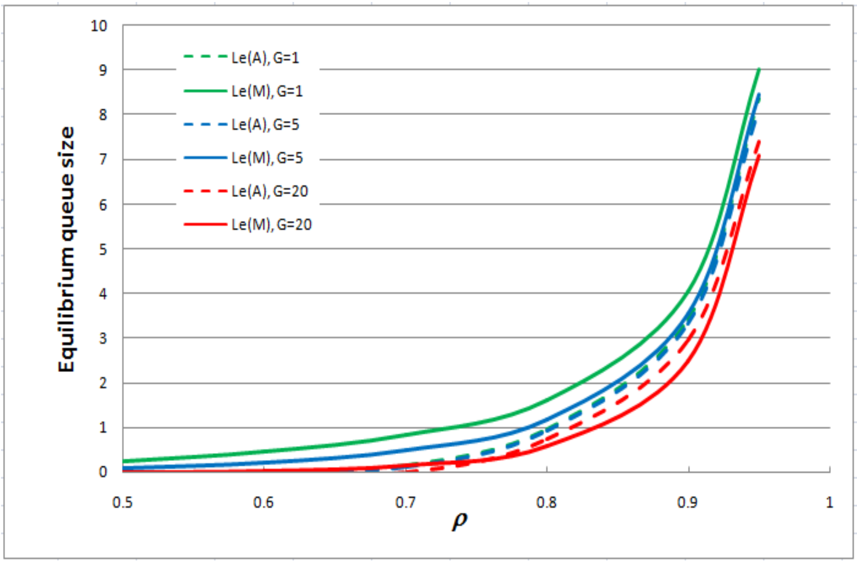


Figure E.3 Akçelik (A) and M/D/1[G] (M) equilibrium stochastic queues compared (graphs sink with increasing *G*, broken lines relate to Akçelik's formula)

The phase queue differs from an initial queue in that it is dependent on ρ<1. However, the phase delay is is a separate component added to the sheared delay, whose modified statistical parameter *C* in the Pollaczek-Khinchin formula takes account of the effect of green capacity. Akçelik and his collaborators estimate this parameter as follows:

$$C_G = 1.22G^{-0.22} \qquad \text{(E.14)}$$

It is unclear whether this is actually applied in the Australian queue formula, but if it is, this should more closely resemble the HCM formula, becoming:

$$L = \frac{\mu t}{4}\left((\rho-1)+\sqrt{(\rho-1)^2+\frac{8C_G(\rho-\rho_0)}{\mu t}}\right) \qquad \text{(if } \rho>\rho_0\text{, else zero)} \qquad \text{(E.15)}$$

If a statistical parameter is to be used it should either be 1.5 from (E.13), or C_G as above, since it compensates to some extent for ρ_0. A value *C*=0.5 would be much too small.

E.5 COMPARISON OF THE METHODS

Figure E.4 compares four different formulae for queue growth with ρ=0.9 and *G*=1, showing that the Australian method estimates a shorter queue. For smaller values of ρ the difference is much greater, as is to be expected because of the substantial value of ρ_0.

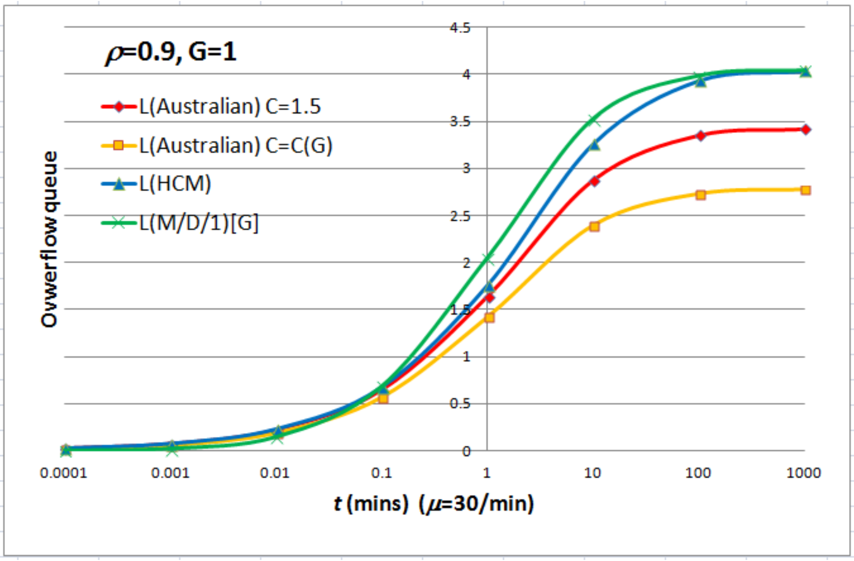


Figure E.4 Different methods compared for theoretical minimum green case

Figure E.5 compares the same methods for a high value of green capacity, showing greater differences between the methods. In this case M/D/1[G] lies between the two possible Australian curves (the upper one being the standard version). As far as it is possible to tell, the HCM method does not allow for green capacity, so estimates a much greater queue.

Reports like Akçelik (1998a) contain a large amount of empirical factors and adjustments to allow for things like junction geometry, turning movements and signal control algorithms. So interpreting narrowly-focused comparisons like the above is problematic, as is drawing any conclusion about which method is most realistic. This analysis is therefore intended mainly to draw attention to structural features and assumptions that justify adopting M/D/1[G] in the present theoretical line of research.

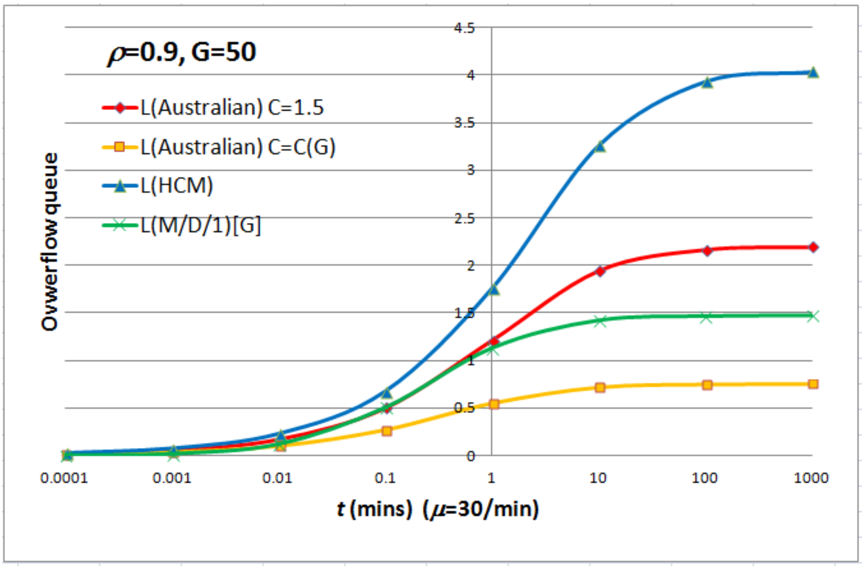


Figure E.5 Different methods compared for a case with a long green time

E.6 COMMON BEHAVIOUR AND UNDERLYING STRUCTURE

The most fundamental issue is the explicit quasi-static assumption. All the methods described satisfy the following relationship with *their* equilibrium queue value L_e, as can be seen by evaluating (E.16/17) using first-order Taylor expansion of the square root when t is large.

$$L = \frac{\mu t}{4}\left(-(1-\rho)+sign(1-\rho)\sqrt{(1-\rho)^2+\frac{8(1-\rho)L_e}{\mu t}}\right) \tag{E.16}$$

or

$$L = \frac{(1-\rho)\mu t}{4}\left(-1+sign(1-\rho)\sqrt{1+\frac{8L_e}{(1-\rho)\mu t}}\right) \tag{E.17}$$

This formula most certainly does *not* work for quasi-static sheared M/D/1, which has a far lower initial growth rate. Average rate of growth can be calculated by differentiating the deterministic queue formula, and thence average utilisation over the period concerned, according to equation (E.18), where $\bar{u}$ is used rather than x since it applies between pairs of calculation points, which in the calculations giving Figures E.3-4 advance logarithmically.

$$\frac{dL}{dt} = (\rho - \bar{u})\mu \tag{E.18}$$

Utilisations are graphed in Figure E.6, where it is evident that the negative initial utilisations of the Australian and HCM functions are unphysical, although they persist for a short time.

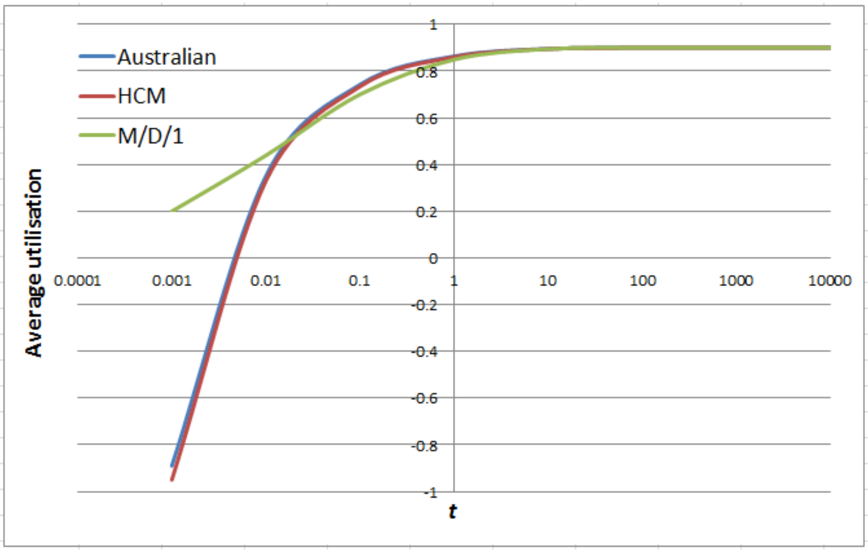


Figure E.6 Behaviour of utilisations of the different methods

The generic non-M/D/1 equation (E.17) is the solution of:

$$L + \frac{2L^2}{(1-\rho)\mu t} = L_e \tag{E.19}$$

This simple formula is problematic when *t* approaches zero. Assuming no initial queue, as above, and on the basis that average utilisation must satisfy the deterministic formula:

$$L = (\rho - x)\mu t \tag{E.20}$$

after some manipulation at an expression for *L*(*t*) in terms of *x*(*t*) without *t* is arrived at:

$$L = \frac{(1-\rho)L_e}{1+\rho-2x} \tag{E.21}$$

The L_e corresponding to the HCM formula is that for M/D/1, so if the method *were* quasi-static it should be possible to replace every instance of ρ by *x* (or some function of instantaneous utilisation) that can then be calculated from the *L* values. Figure E.7 shows how these are initially very different but converge once *x* gets close enough to ρ. It is possible to show that the initial rate of change of the queue at *t*=0 is infinite, and obviously both the instantaneous rate and its time average change very rapidly at first, so many methods in the main dissertation that depend on the initial utilisation or on p_0 could not be used.

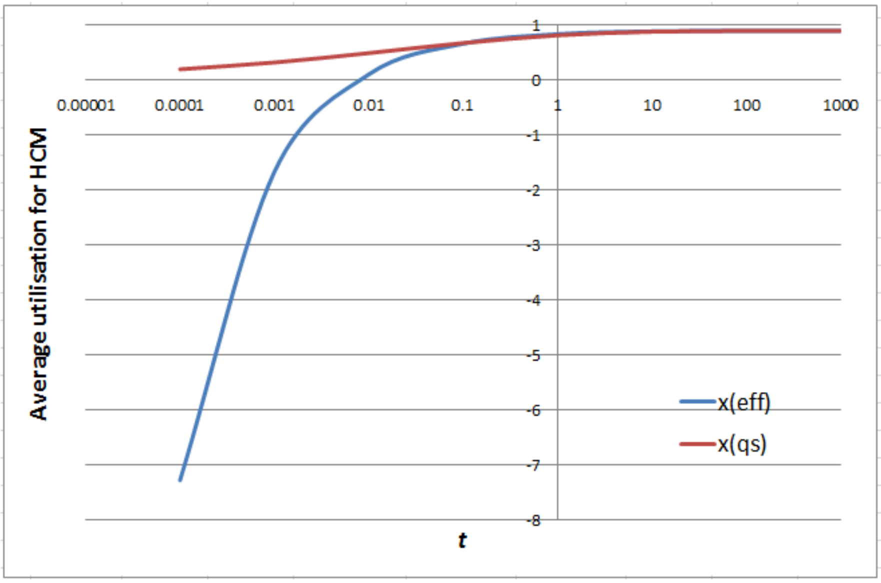


Figure E.7 Effective versus quasi-static utilisations for HCM queue function

E.7 ANALOGOUS FORMULATION OF M/D/1 SOLUTION

The quasi-static sheared M/D/1 formula is more complicated than the HCM and similar formulae. From the expressions in main Chapter 4, the equivalent of (E.19) is:

$$L + \frac{(\mu t - C)L^2}{2C\rho\mu t + ((1-\rho)\mu t)^2} = L_e\left(\frac{\rho\mu t}{2L_e + \rho\mu t}\right) \qquad \text{(E.22)}$$

When t is very large, $L \rightarrow L_e$ as it should. Unlike (E.19), when $t \rightarrow 0$, the RHS vanishes, and the LHS is consistent with L starting from zero and initially growing linearly.

E.8 POSSIBILITY OF EXPONENTIAL APPROXIMATIONS

The exponential formula developed in main Chapter 4, constrained by time constants derived from initial and asymptotic states, appears to approximate queue decay quite well, though not queue growth. Although a similar function is used to estimate the variance of a decaying queue, the mean queue and variance formulae are not formally consistent. A simplified version of the mean queue function, growing from zero with unvarying time constant, is:

$$L = L_e\left(1 - \exp\left(-\frac{\rho\mu t}{L_e}\right)\right) \qquad \text{(E.23)}$$

However, the time development of (E.23) does not match simulation well, as shown by Figure E.8, where it converges to the equilibrium value much too quickly.

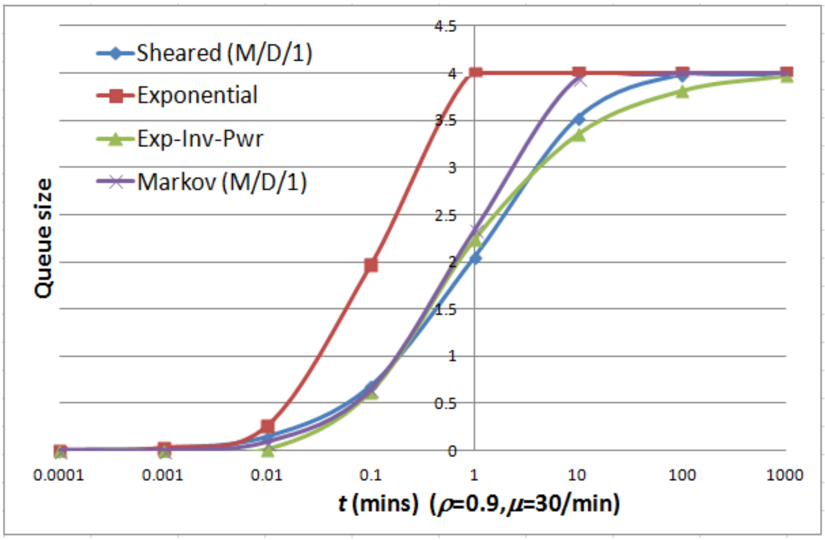


Figure E.8 Exponential alternatives to the M/D/1 function

Realistic behaviour seems to demand a more complex function. The exponential function of *inverse* time, (E.24) can fit the general shape of the M/D/1 function[66], as shown in Figure (E.8), but cannot reproduce the initial rate of change, and involves calibration constants:

$$L = L_e \exp\left(-(2t)^{-0.35}\right) \tag{E.24}$$

Returning to equation (E.23), the delay-per-unit-time is:

$$D = \frac{1}{t}\int_{y=0}^{t} L(y)\,dy = L_e\left(1 + \frac{L_e}{\rho\mu t}\left(\exp\left(-\frac{\rho\mu t}{L_e}\right) - 1\right)\right) \tag{E.25}$$

To get the value at t=0 it is necessary to take a limit, and only the second-degree term in the exponential survives, giving $D \rightarrow \frac{1}{2}\rho\mu t$ as $t \rightarrow 0$. Putting (E.25) into the variance equation:

$$W = 2(1-\rho)(L_e - D)\mu t = \frac{2(1-\rho)L_e^2}{\rho}\left(1 - \exp\left(-\frac{\rho\mu t}{L_e}\right)\right) \tag{E.26}$$

When $t \rightarrow \infty$, in the cases of M/M/1 and M/D/1 respectively, the limiting values are:

$$W_e = \frac{2\rho}{1-\rho} \qquad \text{the correct M/M/1 value being} \quad W_e = \frac{2\rho}{(1-\rho)^2} \tag{E.27}$$

$$W_e = \frac{\rho^3}{2(1-\rho)} \qquad \text{the correct M/D/1 value being} \quad W_e = \frac{\rho^2\left(6 - 4\rho + \rho^2\right)}{6(1-\rho)^2} \tag{E.28}$$

So this simple model is not consistent with variance. It is asserted that an extension of (E.23) with a weighted sum of two exponential terms, satisfying the initial rate of change, also cannot be made consistent with variance. Referring to the series formula given by Morse (1958), an infinite weighted sum of exponential functions represents M/M/1 exactly, so perhaps there is some finite number of terms greater than two that can make mean and variance consistent, though not necessarily higher moments. This is a question that might be investigated in further research.

[66] As noted elsewhere, the Markov simulation may not be completely reliable near equilibrium.

E.9 SUMMARY

The quasi-static shearing approach seems better able to match all the structural properties of queue growth than the simpler HCM formula or its more empirical Australian development, or ostensibly simpler exponential functions that would require calibration constants. Further investigation of the consequences of the quasi-static assumption could be of interest, though not materially affecting this work whose objective is to make the minimum adjustments to the sheared queue method necessary to allow variance also to be calculated. The same applies to further investigation of exponential approximations.

Local references also in main references

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

Akçelik R (1998a). Traffic signals: capacity and timing analysis. *Report ARR 123*. Australian Road Research Board. [Most recent edition of report first published 1981]

Akçelik (2001). *HCM 2000 back of queue model for signalised intersections*. Technical Note. Akçelik and Associates Pty Ltd, September 2001.

Akgüngör A P and A G R Bullen (2007). A new delay parameter for variable traffic flows and signalized intersections. *Turkish J. of Eng. and Env. Science* 31(**2007**), 61-70.

Bin Han (1996). A new comprehensive sheared delay formula for traffic signal optimisation. *Transportation Research A*, 30(**2**), 155-171.

Cheng DX, Messer C J, Tian Z Z and Liu J (2003). Modification of Webster's Minimum Delay Cycle Length Equation Based on HCM 2000. [Texas Transportation Institute] *TRB 2003 Annual Meeting*, Washington DC.
http://wolfweb.unr.edu/homepage/zongt/Publications_files/ChengDingXinTRB-03.pdf

FHWA. http://ops.fhwa.dot.gov/publications/fhwahop08054/sect4.htm,

Morse P M (1958). *Queues inventories and maintenance.* Wiley.

Spiess H. Conical volume-delay functions. *Transportation Science 24(**2**),* 153-158.

APPENDIX F - STANDARD ERROR OF EQUILBRIUM MEAN QUEUE ESTIMATE

To predict the accuracy of the mean in a simulation of an equilibrated queue process, as a function of the number of events simulated, we consider the relationship between the run length of sequences of events and the variance of the means of runs, for various run lengths. This analysis is required to confirm that simulation results with a given number of events are consistent with what would be expected, and there can be confidence in the simulation.

In a run of events of an equilibrated queue, the mean of run means is the same as the overall mean, and assuming the total number of events is large this should approach the expected mean. The variance of run means will start equal to the overall variance of queue sizes when run length is 1, should eventually decrease to zero as run length is increased. In a sequence of random observations, where there is no correlation between successive observations, the usual standard error formula should apply. However, successive queue sizes are not uncorrelated, and the relationship with run length is therefore likely to be quite different. In principle, we would expect the ‘standard error’ to be greater than that for uncorrelated observations, meaning that more events must be simulated to get an accurate estimate of the mean than would be the case for a random sequence.

In a simulation of a single-lane M/M/1 queue with ρ=0.9, mean queue size averaged over 900,000 arrival and service events (not just service events), numbered from 100,001 to 1M to allow an initial period of stabilisation, is L=8.834 which is not far from the expected value of 9.0. Variance V is 82.61 which is somewhat less than the expected value of 90.0 but at least comparable. In terms of standard deviation, σ=9.09 compared to expected 9.49.

Considering that N observations $\{x_i\}$ are divided into m runs of n successive observations, and assuming that observations in any run can be considered equivalent to those in any other, the variance between runs is estimated by:

$$\tilde{V}_n = \frac{1}{m}\sum_{i=1}^{m}\mu_i^2 - \mu^2 = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j - \mu^2 = \frac{V}{n} + \left(1-\frac{1}{n}\right)X_n^2 \qquad \text{(F.1)}$$

Here μ is the ensemble mean, $V \equiv \tilde{V}_1$ is the ensemble variance, and X_n^2 is the mean excess value of any product of two *different* elements x_i and x_j in a run after subtraction of μ^2.

Clearly, when n=1, this expression reduces to the ensemble variance, and for n>1 if the last term were absent it would give the standard error result.

Considering just adjacent queue measurements, the probabilities of the next observed queue size being greater or less than the current value are in the ratio ρ:1 (on the assumption that arrivals and service do not occur literally simultaneously, queue size cannot be unchanged). So if the initial queue size is i then the expected result of the next queue observation is:

$$j = \frac{\rho(i+1)+(i-1)}{\rho+1} = i - \left(\frac{1-\rho}{1+\rho}\right) \qquad \text{(F.2)}$$

Therefore the mean value of products of successive (n=2) observation pairs is given by:

$$\Pi_2 = \sum ijp_i = \sum\left(i^2 - \left(\frac{1-\rho}{1+\rho}\right)i\right)p_i = V_e + L_e^2 - \left(\frac{1-\rho}{1+\rho}\right)L_e \qquad \text{(F.3)}$$

For ρ=0.9, the theoretical value is 170.53. Using the simulated values of mean and variance, it is 160.18. The actual value from the simulation is 158.65. The final term in (F.3) makes only a small contribution. The important point is that the mean value of the product is greater by around the variance than would be expected if successive observations were uncorrelated, which would be just the square of the mean, i.e. 81.0 expected or 78.04 as simulated.

When n>2, the products in X_n^2 are not all of adjacent observations. Queue size observations become progressively less correlated the farther they are apart, but the degree of correlation should continue to depend on ρ. For example, if ρ≈1 then if x_i is large x_{i+k} is likely to be large also. For more modest values of ρ, and eventually for sufficiently large k, the second term in the product will tend towards the mean queue size, and it is not unreasonable to suppose that relaxation should be approximately exponential.

A simple function to predict the mean value of queue size observations k events away from an observation i, which satisfies the end constraints and has roughly the right initial rate of change, is:

$$j(k) = \mu + abs(i-\mu)\exp\left(\frac{-(1-\rho)k}{(1+\rho)abs(i-\mu)}\right) \qquad \text{(F.4)}$$

The mean value of j can be estimated as the integral of (F.4) broadly within the range $[0,n]$, divided by the range, though there may be an issue whether the range ought to be $[1,n]$ or $[0,n\text{-}1]$ which can perhaps be ignored if n is sufficiently large. The approximate result is given by (F.5) which has extremal values i at n=0 and μ at $n=\infty$, and gives the correct value (F.2) for n=1 when approximated to second order:

$$\bar{j} = \mu - sign(i-\mu)\frac{(1+\rho)(i-\mu)^2}{n(1-\rho)}\left[\exp\left(\frac{-(1-\rho)n}{(1+\rho)abs(i-\mu)}\right)-1\right] \qquad \text{(F.5)}$$

The evaluation of the product moment (F.3) as an integral of a continuous function looks awkward because of the exponential function of i in (F.5), which leads to a factor involving i^3 and an exponential term containing both i and its inverse. Alternatively, (F.5) might be converted to a discrete form. Either appears tedious, and in practice the desired numerical results can be obtained by direct calculation, the variance between runs being estimated by:

$$\tilde{V}_n = \Pi_n - L_e^2 \qquad \text{(F.6)}$$

Figure F.1 compares the simulated and estimated standard deviation between runs as a function of run length, showing that they match closely but are substantially different from the $\sigma/\sqrt{n}$ standard error for uncorrelated observations, which is also shown. This confirms the model and predicts σ error in the mean of 0.216 for n = 1M events. It is reasonable to expect that, if the simulation is accurate, the error in the simulated mean will be of similar order, the actual value being 0.162.

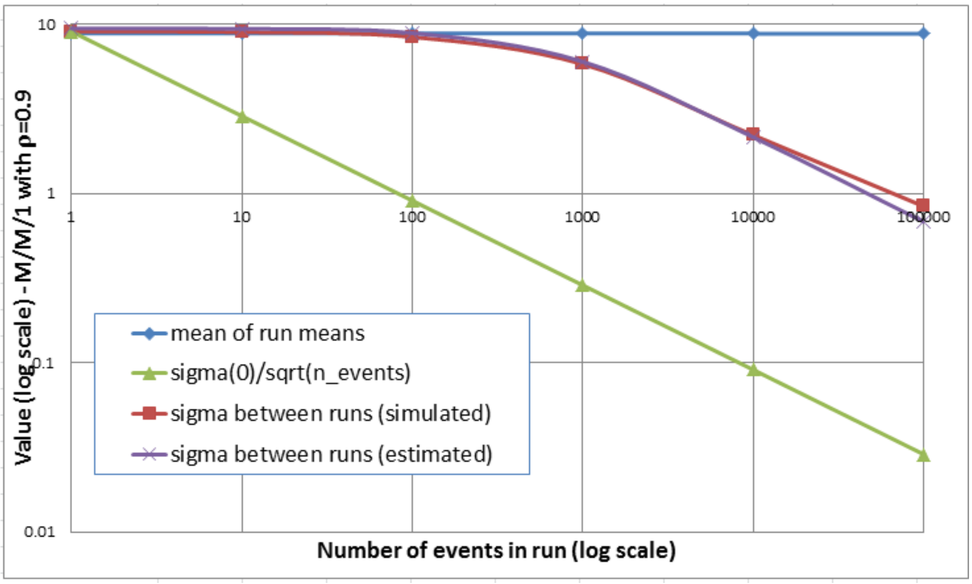


Figure F.1 Standard error of mean queue estimate as function of no. of events, M/M/1, ρ=0.9

A useful observation is that, for sufficiently large *n*, the expected error decreases as 1/√*n*, allowing it to be extrapolated to any greater range, although it is displaced from the uncorrelated case by several orders of magnitude.

For *n*=10,000, the number of events simulated in earlier experiments which were considered unsatisfactory, the predicted error in the mean is 2.16, implying a one-sigma range of 6.84-11.16, so the result is likely to be substantially different from expectation. This is evident in Table F.1, which may be compared with Table 6.2.1 in main Chapter 6. On the face of it this is a surprising result since the stochastic relaxation constant should be only around 1.9*379.7 ≅ 722 events (including both arrival and service events).

Table F.1 Queue sizes in earlier multi-lane simulations (ρ=0.9, μ=1.0, 10,000 events)

No. of Lanes	**Capacity μ and type of process**	**Total Queue**	**Individual Lane Queues**				**Correlation of Queue Sizes(R^2)**
			Lane 1	**Lane 2**	**Lane 3**	**Lane 4**	
1	1.00 random	7.317	7.317	-	-	-	n/a
4	0.25 indep.	41.633	10.772	8.823	10.377	11.661	-0.033
4	1.00 shared	6.743	1.467	1.796	1.673	1.806	0.237
4	1.00 shortest	6.743	1.644	1.687	1.693	1.719	0.844

In conclusion, the foregoing analysis has provided a means of estimating the expected accuracy of a simulated M/M/1 equilibrium queue as a function of the length of simulation, or conversely providing an indication of how many events need to be simulated to get a result of given accuracy, and has shown that this number can be several orders of magnitude greater than would be expected on the basis of the usual standard error rule.