Statistics and Probability Letters 82 (2012) 998–1004

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

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

Fitting birth-and-death queueing models to data

Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA

ARTICLE INFO

Article history:
Received 6 January 2012
Received in revised form 11 February 2012
Accepted 12 February 2012
Available online 17 February 2012

Keywords:
Birth-and-death processes
Empirical birth-and-death processes
Fitting birth-and-death processes to data
Conservation laws
Operational analysis

ABSTRACT

Given measurements of the number of customers in a queueing system over a finite time interval, it is natural to try to fit a stationary birth-and-death process model, because it is remarkably tractable, even when the birth and death rates depend on the state in an arbitrary way. Natural estimators of the birth (death) rate in each state are the observed number of transitions up (down) from that state divided by the total time spent in that state. It is tempting to validate the model by comparing the steady-state distribution of the model based on those estimated rates to the empirical steady-state distribution recording the proportion of time spent in each state. However, it is inappropriate to draw strong conclusions from a close fit to the same data, because these two distributions are necessarily intimately related, even if the model assumptions are not nearly satisfied. We elaborate by (i) establishing stochastic comparisons between these two fitted distributions using likelihood-ratio stochastic ordering and (ii) quantifying their difference.

1. Introduction

This paper is motivated by efforts to fit stochastic queueing models to data from system measurements in call centers and hospitals, for example, as collected in Armony et al. (2011) and Brown et al. (2005). Since the number of customers in a queueing system typically increases or decreases by one at each transition, it is natural to consider fitting a stationary birth-and-death (BD) process to the observed segment of the sample path. Clearly, there also are many other applications where a BD model might be fitted to data.

Since data over a finite time interval will invariably involve only finitely many transitions, it is natural to fit a finite-state stationary BD process to the data. Thus the typical state space for the BD process is $\{0, 1, \ldots, m\}$. With these $m + 1$ states, there are $2m$ parameters, the m birth rates λ_j in states $0 \le j \le m - 1$, and the m death rates μ_j in states $1 \le j \le m$. If the model fit is genuinely good, then the BD model can be very helpful because it is remarkably tractable; see Section 2.

There already is quite an extensive statistical theory for estimating the parameters of queueing models and stationary BD processes; see, for example, Bhat et al. (1997), Billingsley (1961), Israel et al. (2001), Keiding (1975), Ross et al. (2007), and Wolff (1965). We will be considering the natural estimator of the birth and death rates considered in Wolff (1965); i.e., the birth (death) rate in each state is estimated by the observed number of transitions up (down) from that state divided by the total time spent in that state.

It is tempting to validate the BD model fit in that way by comparing the steady-state distribution of the model based on those estimated rates to the empirical steady-state distribution recording the proportion of time spent in each state. However, it is inappropriate to draw strong positive conclusions from a close fit to the same data, because these two distributions are necessarily intimately related, without any model assumptions being made. For example, the actual system could be highly non-stationary. Nevertheless, under minor regularity conditions, these two fitted distributions are asymptotically identical as the sample size increases; this is a special case of more general empirical global balance equations

E-mail address: ww2040@columbia.edu.

0167-7152/$ – see front matter
doi:10.1016/j.spl.2012.02.010

in the sample-path analysis of queues in Ch. 4 of El-Taha and Stidham Jr. (1999). In the queueing literature, the close relation between these two empirical fitted distributions seems to go back to the operational analysis of Buzen (1976), Buzen (1978), Buzen and Denning (1980), and Denning and Buzen (1978); see Sections 4.6 and 4.7 of El-Taha and Stidham Jr. (1999) for discussion and references.

Given that the two empirical fitted distributions are necessarily nearly identical, with virtually no conditions at all, it is evident that the relation has no direct implication about either (i) what is an appropriate stochastic model, or (ii) the system performance at other times. These important issues require further properties that must be checked empirically. For example, in justifying the applied relevance of operational analysis, in Section 2 of Denning and Buzen (1978) the authors discuss this issue and mention a variety of invariance assumptions to justify predictions at other times. Otherwise, this relation serves only as a (useful) consistency check on the data processing.

In this paper, after quickly reviewing basic BD theory in Section 2, and carefully specifying the fitting procedure in Section 3, we elaborate on the relation between these two fitted distributions based on data from a finite time interval in Section 4 by (i) establishing stochastic comparisons between these two fitted distributions using likelihood-ratio stochastic ordering and (ii) quantifying their difference. Afterwards, in Section 5, we briefly discuss additional data analysis steps to fully validate the fitted BD process. In Section 6, we discuss the associated statistical problem of estimating confidence intervals for the birth and death rates. Finally, in Section 7, we discuss remaining problems.

2. Review of birth-and-death process theory

Let $X \equiv \{X(t) : t \geq 0\}$ be a (finite-state stationary) BD process on the state space $\{0, 1, \ldots, m\}$ with strictly positive birth rates λ_i, $1 \leq i \leq m-1$, and death rates μ_i, $1 \leq i \leq m$. That means that X is a reversible irreducible continuous time Markov chain (CTMC) with all transitions up one or down one; for example, see Keilson (1979), Ch. 5 of Ross (1996), and Ch. 6 of Ross (2010). Thus, the characterizing $(m+1) \times (m+1)$ rate matrix (infinitesimal generator) Q of the CTMC has elements $Q_{i,i+1} \equiv \lambda_i$, $0 \leq i \leq m-1$, and $Q_{i,i-1} \equiv \mu_i$, $1 \leq i \leq m$, with all other off-diagonal elements 0 and all row sums 0. Equivalently, the successive holding times in state i are independent and identically distributed (i.i.d.) exponential random variables with mean $1/(\lambda_i + \mu_i)$, and the probability of an upward transition at each transition time from state i is $\lambda_i/(\lambda_i + \mu_i)$, independent of the holding time and all prior history.

An (irreducible finite-state) BD process X has a unique limiting steady-state probability distribution α, i.e.,

$$\alpha_j \equiv \lim_{t\to\infty} P(X(t) = j | X(0) = i) \quad \text{for all } i, \tag{1}$$

which is also the unique stationary distribution, i.e.,

$$\alpha_j = \sum_{i=0}^{m} \alpha_i P(X(t) = j | X(0) = i) \quad \text{for all } t > 0. \tag{2}$$

Because of the reversibility, the stationary distribution α can be expressed as the unique solution to the local balance equations

$$\alpha_i \lambda_i = \alpha_{i+1} \mu_{i+1}, \quad 0 \leq i \leq m-1, \tag{3}$$

such that $\sum_{i=0}^{m} \alpha_i = 1$. The local balance equations (3) can be solved recursively to give

$$\alpha_i = \frac{r_i}{\sum_{j=0}^{m} r_j}, \tag{4}$$

where $r_0 \equiv 1$ and

$$r_i \equiv \frac{\lambda_0 \times \cdots \times \lambda_{i-1}}{\mu_1 \times \cdots \times \mu_i}, \quad 1 \leq i \leq m; \tag{5}$$

for example, see Section 6.3 of Ross (2010). It is also not difficult to compute transient performance measures of finite-state BD processes, as shown by Keilson (1979).

The BD models considered in most applications have special structure. For example, many queueing applications involve the classical $M/M/s/r$ queueing model, which has a Poisson arrival process with rate λ (the first M), exponential service times with mean μ^{-1} (the second M), s homogeneous servers working in parallel and r extra waiting spaces. Thus the $M/M/s/r$ model has constant birth rates $\lambda_i \equiv \lambda$, $0 \leq i \leq s+r-1$, and simple death rates, $\mu_i \equiv \min\{i, s\}$, $1 \leq i \leq s+r$. However, it often is of interest to consider more general BD queueing models. For example, when there is balking (arrivals refusing to join when the line is too long), see Whitt (1999), for example, the arrival rate may be decreasing; when there is customer abandonment if they have not progressed rapidly enough, the death rate may be increasing more rapidly than above; when customers have non-exponential patience distributions, it can be effective to approximate by a BD model with a more general state-dependent death rate; see Whitt (2005). For complex real applications, it is natural to let the data dictate what the relevant birth and death rates are. For example, the number of working servers may not be a fixed deterministic quantity, but nevertheless a BD model could be useful.

3. Fitting the BD model to queueing system data

Consider a queueing system in which arrivals and departures occur one at a time. Let $X(s)$ be the number of customers in the system at time s. We now consider fitting a BD model to data collected over an interval $[0, t]$.

Let $\bar{\lambda}_i(t)$ and $\bar{\mu}_i(t)$ be natural direct estimates of the birth rates and death rates based on sample averages over the time interval $[0, t]$. Similarly, let $\bar{\alpha}_i(t)$ be natural direct estimates of the stationary distribution based on sample averages over the time interval $[0, t]$.

In particular, let $A_i(t)$ be the number of arrivals during the interval $[0, t]$ when the system is in state i; let $D_i(t)$ be the number of departures during the interval $[0, t]$ when the system is in state i; and let $T_i(t)$ be the total time during the interval $[0, t]$ in which the system is in state i; i.e.,

$$T_i(t) \equiv \int_0^t 1_{\{X(s)=i\}}\, ds, \quad t \geq 0, \tag{6}$$

where 1_A is the indicator function of the set A, equal to 1 on A and equal to 0 otherwise. Then let

$$\bar{\lambda}_i(t) \equiv \frac{A_i(t)}{T_i(t)}, \qquad \bar{\mu}_i(t) \equiv \frac{D_i(t)}{T_i(t)} \quad \text{and} \quad \bar{\alpha}_i(t) \equiv \frac{T_i(t)}{t}, \quad t \geq 0. \tag{7}$$

In general, this estimation procedure need not produce an irreducible BD process, because there can be initial and final transient states. However, there is a largest subset of states that is the state space of an irreducible BD process, with all other states being transient. Necessarily, $\bar{\lambda}_i(t) > 0$ for $a_1 \leq i \leq a_2$ with $\bar{\lambda}_i(t) = 0$ otherwise, and $\bar{\mu}_i(t) > 0$ for $d_1 \leq i \leq d_2$ with $\bar{\lambda}_i(t) = 0$ otherwise, for some constants a_1, a_2, d_1, and d_2. There are three possibilities for these "intervals of positive rates": (i) $a_1 = d_1 - 1$ and $a_2 = d_2 - 1$, (ii) $a_1 \leq d_1 - 1$ and $a_2 \geq d_2 - 1$, with at least one of these two inequalities being strict, or (iii) $a_1 \geq d_1 - 1$ and $a_2 \leq d_2 - 1$, with at least one of these two inequalities being strict. In case (i), the process is irreducible and the state space is $\{a_1, \ldots, a_2+1\} = \{d_1-1, \ldots, d_2\}$; in case (ii), there are transient states, so the BD process is reducible, with the state space of the irreducible BD process being $\{d_1 - 1, \ldots, d_2\}$, while all other states are transient, being visited by initial or final births, but never by deaths; in case (iii), again there are transient states, so the BD process is reducible, with the state space of the irreducible BD process being $\{a_1, \ldots, a_2 + 1\}$, while all other states are transient, being visited by initial or final deaths, but never by births. In all three cases, there is a unique stationary distribution, which places 0 probability on each transient state, if there are any.

For simplicity, henceforth we assume that the irreducible case (i) prevails with $a_1 = 0 < a_2 + 1 = d_2 = m$. From Section 2, we see that, under the simplifying assumption of irreducibility, this estimated BD process has the unique stationary probability distribution

$$\bar{\alpha}_i^e(t) \equiv \frac{\bar{r}_i(t)}{\sum_{j=1}^{m} \bar{r}_j(t)}, \quad 0 \leq i \leq m, \tag{8}$$

where $\bar{r}_0(t) \equiv 1$ and

$$\bar{r}_i(t) \equiv \frac{\bar{\lambda}_0(t) \times \cdots \times \bar{\lambda}_{i-1}(t)}{\bar{\mu}_1(t) \times \cdots \times \bar{\mu}_i(t)}, \quad 0 < i \leq m. \tag{9}$$

Equivalently, $\bar{\alpha}^e(t)$ is the unique probability vector satisfying the local balance equation associated with the estimated birth and death rates; i.e.,

$$\bar{\alpha}_i^e(t)\bar{\lambda}_i(t) = \bar{\alpha}_{i+1}^e(t)\bar{\mu}_{i+1}(t) \quad \text{for all } i,\ 0 \leq i < m. \tag{10}$$

4. The relation between the two fitted stationary distributions

We now establish more precise connections between the distribution $\bar{\alpha}^e(t)$ based on formula (8) using the estimated birth and death rates and the direct empirical distribution $\bar{\alpha}(t)$ in (7) for finite values of t. We emphasize that these two probability distributions will not have the same support, and thus of course they could not be equal, if the irreducibility condition above is not satisfied. We thus assume irreducibility below. In practice, the irreducibility can always be achieved by removing the initial and/or final transient from the sample path if either (or both) is (are) there.

For the stochastic comparison, we use the notion of *likelihood ratio stochastic ordering* (LR) for probability mass functions (pmfs); for example, see Section 9.4 of Ross (1996). Let X_1 and X_2 be two random variables, each taking values in the non-negative integers, with pmfs $p_i(k) \equiv P(X_i = k)$. We say that X_1 is stochastically less than or equal to X_2 in the LR ordering, and write $X_1 \leq_{LR} X_2$ or $p_1 \leq_{LR} p_2$, if

$$\frac{p_1(k+1)}{p_1(k)} \leq \frac{p_2(k+1)}{p_2(k)} \quad \text{for all integers } k, \tag{11}$$

where at least one is positive. It is well known that LR ordering implies ordinary stochastic order. We say that X_1 is stochastically less than or equal to X_2, and write $X_1 \leq_{st} X_2$ or $p_1 \leq_{st} p_2$, if

$$\sum_{j=k}^{\infty} p_1(j) \leq \sum_{j=k}^{\infty} p_2(j) \quad \text{for all } k. \tag{12}$$

Equivalently, $X_1 \leq_{st} X_2$ if $E[f(X_1)] \leq E[f(X_2)]$ for all non-negative non-decreasing real-valued functions f on $\mathbb{R}$.

Theorem 1 (*Stochastic Comparison of the Two Stationary Distributions*)**.** *Consider a sample path segment over an interval* $[0, t]$ *of a stochastic process with only finitely many transitions, all of which are* ± 1*. Suppose that a BD process is fitted to this data, as in* (7)*, and suppose that it is irreducible with state space* $\{0, \ldots, m\}$*. For* $i_0 \equiv X(0)$ *and* $i_t \equiv X(t)$*, there are three mutually exclusive and exhaustive alternatives:*

$$\begin{aligned} &\text{If } i_0 = i_t, \quad \text{then } \bar{\alpha}(t) = \bar{\alpha}^e(t); \\ &\text{if } i_0 < i_t, \quad \text{then } \bar{\alpha}(t) \geq_{LR} \bar{\alpha}^e(t); \\ &\text{if } i_0 > i_t, \quad \text{then } \bar{\alpha}(t) \leq_{LR} \bar{\alpha}^e(t). \end{aligned} \tag{13}$$

Moreover, the difference, $\bar{\Delta}_i(t) \equiv \bar{\alpha}_i^e(t) - \bar{\alpha}_i(t)$*, can be quantified by solving the finite recursion (for* $0 \leq i \leq m$*)*

$$\bar{\Delta}_i(t)\bar{\lambda}_i(t) = \bar{\Delta}_{i+1}(t)\bar{\mu}_{i+1}(t) + \frac{\bar{e}_i(t)}{t} \quad \text{with } \sum_{i=0}^{m} \bar{\Delta}_i(t) = 0, \tag{14}$$

where $\bar{e}_i(t) \equiv 1_{\{i_0 \geq i > i_t\}} - 1_{\{i_0 \leq i < i_t\}}$*, so that* $\bar{e}_i(t) = 0$ *for all but* $|i_t - i_0|$ *values of* i*.*

Proof. By the definitions in (7),

$$\bar{\alpha}_i(t)\bar{\lambda}_i(t) = \left(\frac{T_i(t)}{t}\right)\left(\frac{A_i(t)}{T_i(t)}\right) = \frac{A_i(t)}{t} \tag{15}$$

and

$$\bar{\alpha}_{i+1}(t)\bar{\mu}_{i+1}(t) = \left(\frac{T_{i+1}(t)}{t}\right)\left(\frac{D_{i+1}(t)}{T_{i+1}(t)}\right) = \frac{D_{i+1}(t)}{t}. \tag{16}$$

However, since all births in state i take the system to state $i+1$, while all deaths in state $i+1$ take the system to state i,

$$|A_i(t) - D_{i+1}(t)| \leq 1 \quad \text{for all } i. \tag{17}$$

We can say more if we look at the initial state i_0 and the ending state i_t. First, if $i_0 = i_t$, then $A_i(t) = D_{i+1}(t)$ for all i. Combining this with (15) and (16), we see that $\bar{\alpha}(t)$ satisfies the local balance equation (10). Hence, in this case, with $i_0 = i_t$, we must have $\bar{\alpha}(t) = \alpha^e(t)$, as claimed in (13).

Next, if $i_0 < i_t$, then

$$A_i(t) = D_{i+1}(t) + 1 \quad \text{for } i_0 \leq i < i_t; \quad \text{else } A_i(t) = D_{i+1}(t). \tag{18}$$

As a consequence, instead of the local balance equations in (10), in this case, the probability vector $\bar{\alpha}(t)$ satisfies the associated system of inequalities

$$\bar{\alpha}_i(t)\bar{\lambda}_i(t) \leq \bar{\alpha}_{i+1}(t)\bar{\mu}_{i+1}(t) \quad \text{for all } i. \tag{19}$$

However, we can immediately rewrite (19) as

$$\frac{\bar{\alpha}_{i+1}(t)}{\bar{\alpha}_i(t)} \geq \frac{\bar{\lambda}_i(t)}{\bar{\mu}_{i+1}(t)} = \frac{\bar{r}_{i+1}(t)}{\bar{r}_i(t)} = \frac{\alpha_{i+1}^e(t)}{\alpha_i^e(t)}, \tag{20}$$

so that $\bar{\alpha}(t) \geq_{LR} \alpha^e(t)$, as claimed in (13).

By similar reasoning, if $i_0 > i_t$, then

$$A_i(t) = D_{i+1}(t) - 1 \quad \text{for } i_0 \leq i < i_t, \quad \text{else } A_i(t) = D_{i+1}(t), \tag{21}$$

so that $\bar{\alpha}(t) \leq_{LR} \alpha^e$.

In general, we have (10) and

$$\bar{\alpha}_i(t)\bar{\lambda}_i(t) = \bar{\alpha}_{i+1}(t)\bar{\mu}_{i+1}(t) + e_i/t \quad \text{for all } i. \tag{22}$$

Subtracting these equations directly yields (14). □

From the three conditions in (13), we see that the two distributions are always identical with appropriate initial and terminal conditions. The minor differences more generally are only due to "edge effects", just as in the related conservation law $L = \lambda W$.

If we add additional regularity conditions, then we can also show that the difference due to these edge effects is asymptotically negligible as $t \to \infty$. We can also bound the rate of convergence. For this purpose, we consider the estimation as a function of the interval endpoint t. See Chapter 4 of El-Taha and Stidham Jr. (1999) for related asymptotic results. We avoid problems caused by dividing by small t by restricting the setting to $t \geq 1$. We deduce the following bound on the rate of convergence from Theorem 1; we omit the proof.

Corollary 4.1 (*Bound on Rate of Convergence*). *If* $m(t) < m < \infty$, $0 < a_1 \leq \bar{\lambda}_i(t) \leq a_2 < \infty$ *and* $0 < b_1 \leq \bar{\mu}_i(t) \leq b_2 < \infty$ *for all* $t \geq 1$, *then*

$$|\bar{\Delta}_i(t)| \leq K/t \quad \text{for all } t \geq 1, \tag{23}$$

where K *is a function of* m, a_1, a_2, b_1, *and* b_2.

5. Validating the stationary BD model

5.1. Time stationarity

In many queueing applications, such as call centers and hospitals, there tends to be systematic variation in arrival rates and performance measures over time. Thus, it is important to check that a stationary model is really appropriate for the time interval under consideration. This is a vital step if any stationary stochastic model is used.

Given data over a time interval $[0, t]$, stationarity can be checked by considering subintervals $[t_1, t_2]$ with $0 \leq t_1 < t_2 \leq t$. First, let $A_i(t_1, t_2)$ be the number of arrivals during the interval $[t_1, t_2]$ when the system is in state i, and similarly for the other processes. Then, as in (7), let $\bar{\lambda}_i(t_1, t_2) \equiv A_i(t_1, t_2)/T_i(t_1, t_2)$. Let f be an arbitrary real-valued function on the state space. We can extend the statistics $\bar{\lambda}_i(t)$, $\bar{\mu}_i(t)$, $\bar{\alpha}_i(t)$ and $\bar{\alpha}_i^e(t)$ to associated statistics as functions of the triple (f, t_1, t_2), for example, by letting

$$\bar{\lambda}_f(t_1, t_2) \equiv \sum_{i=0}^{\infty} f(i)\bar{\lambda}_i(t_1, t_2). \tag{24}$$

It is good to check that, for various functions f, the statistics $\bar{\lambda}_f(t_1, t_2)$, $\bar{\alpha}_f(t_1, t_2)$ and $\alpha_f^e(t_1, t_2)$ are approximately constant, independent of the subinterval $[t_1, t_2]$. For example, if $f(i) = 1$ for all i, then $\bar{\lambda}_f$ is the estimated total arrival rate; if $f(i) = i^k$, then $\bar{\alpha}_f$ is the kth moment of the estimated steady-state distribution. A simple approach is to choose a few representative functions f, fix $t_1 = 0$, and plot as a function of t_2, $0 \leq t_2 \leq t$.

Similarly, if prediction is contemplated for another time, then evidence should be sought that these estimated rates are still relevant. In the spirit of Section 2 of Denning and Buzen (1978), those are concrete invariance properties to check.

5.2. The BD model assumptions

Given that the system is consistent with a stationary stochastic process for which all transitions are ± 1, it remains to check the Markov property. A manageable way to check that is to use the fact that, within that setting, a BD process can be characterized by having the times spent in each state and the transition at the transition epoch be random variables that are mutually independent and independent of the system history. For a practical test, let $X_k^{(i)}$ be the time spent in state i after arriving in state i from elsewhere, let $J_k^{(i)}$ be $+1$ if the transition at the end of that interval is up one and let $J_k^{(i)} = -1$ otherwise, and let $Y_k^{(i)}$ be the length of time spent away from state i immediately after the interval $X_k^{(i)}$. The sequence of vectors $\{(X_k^{(i)}, J_k^{(i)}, Y_k^{(i)}) : k \geq 1\}$ should be i.i.d. random vectors with $X_k^{(i)}$ being independent of both $J_k^{(i)}$ and $Y_k^{(i)}$; for example,

$$P(X_k^{(i)} > t, J_k^i = 1) = P(X_k^{(i)} > t)P(J_k^i = 1) = e^{-(\lambda_i+\mu_i)t}\left(\frac{\lambda_i}{\lambda_i + \mu_i}\right). \tag{25}$$

With the data, we should check that the empirical distribution (histogram) of $X_k^{(i)}$ is approximately exponential; we should check that the covariances between $X_k^{(i)}$ and $X_{k+1}^{(i)}$, between $J_k^{(i)}$ and $J_{k+1}^{(i)}$, and between $X_k^{(i)} + Y_k^{(i)}$ and $X_{k+1}^{(i)} + Y_{k+1}^{(i)}$ are suitably negligible. We may also want to compare estimates of the asymptotic variance for functions of a BD process to the exact values for a BD process, for which Proposition 1 of Whitt (1992) can be used for comparison.

6. Estimating confidence intervals

Assuming that the validation steps have been carried out, such that we think a BD model is appropriate, it is natural regard the sample averages $\bar{\lambda}_i(t)$, $\bar{\mu}_i(t)$ and $\bar{\alpha}_i(t)$ in (7) as finite-sample estimates of the true, but unknown, model parameters λ_i

and μ_i of a stationary BD model and the associated steady-state probabilities α_i. Following standard statistical practice, it is appropriate to evaluate the effectiveness of these estimates by also estimating the sample variance and confidence intervals. Given that the data come from a single observed sample path, it is natural to use the method of batch means as in simulation output analysis in discrete-event stochastic simulation; for example, see Section 3.3 of Bratley et al. (1987). With that in mind, we suggest two alternative estimation procedures.

The first estimation procedure is based on the observation that the method of batch means applies more naturally to simple time averages, as for $\bar{\alpha}_i(t)$ in (7). We can work directly with time averages if we estimate $\gamma_i \equiv \alpha_i\lambda_i$ and $\delta_i \equiv \alpha_i\mu_i$ by the time averages $\bar{\gamma}_i(t) \equiv t^{-1}A_i(t)$ and $\bar{\delta}_i(t) \equiv t^{-1}D_i(t)$. We then obtain the associated estimators $\bar{\lambda}_{\gamma,i}(t) \equiv \bar{\gamma}_i(t)/\bar{\alpha}_i(t)$ and $\bar{\mu}_{\delta,i}(t) \equiv \bar{\delta}_i(t)/\bar{\alpha}_i(t)$. These alternative estimators are also consistent, but dividing by the small unreliable values $\bar{\alpha}_i(t)$ can lead to poor efficiency (high variance).

The second procedure is based on looking only at the total time spent in state i. That can be done by concatenating the variables $X_k^{(i)}$ in Section 5.2. We also keep track of the transitions at each interval end point. Let $U_k^{(i)}$ and $V_k^{(i)}$ be the intervals between the $(k-1)$th and kth births and deaths, respectively. By the Poisson splitting theorem, Proposition 2.3.2 of Ross (1996), for a BD process, the associated two counting processes are independent Poisson processes. Thus, for a BD process, the sequences $\{U_k^{(i)} : k \geq 1\}$ and $\{V_k^{(i)} : k \geq 1\}$ are mutually independent sequences of i.i.d. exponential random variables with means λ_i^{-1} and μ_i^{-1}. We use this representation, but, to avoid assuming that the BD model is correct, we work with batch means. To illustrate, suppose that we have $n = mk$ observations of $U_k^{(i)}$. (Starting from a fixed interval $[0, t]$, the actual number n is random, which introduces bias, but we shall not consider that issue.) We then form the batch means

$$\bar{U}_{m,k}^{(i)} \equiv k^{-1}\sum_{j=m(k-1)}^{mk} U_j^{(i)}, \qquad \bar{U}_m^{(i,b)} \equiv m^{-1}\sum_{k=1}^{m}\bar{U}_{m,k}^{(i)} \equiv n^{-1}\sum_{j=1}^{n} U_j^{(i)}. \tag{26}$$

and the sample variance

$$\bar{\sigma}^2_{U_m^{(i,b)}} \equiv s^2_{U_m^{(i,b)}} \equiv (m-1)^{-1}\sum_{k=1}^{m}(\bar{U}_{m,k}^{(i)} - \bar{U}_m^{(i,b)})^2. \tag{27}$$

In great generality, even when the process is not a BD process, the batch means $\bar{U}_{m,k}^{(i)}$ should be approximately m i.i.d. normal random variables. If the variance $\sigma^2_{U^{(i,b)},m}$ were known, then the random variable $\bar{U}_m^{(i,b)}$ has a normal distribution with variances $\sigma^2_{U_m^{(i,b)}}/m$. Since the variance is in fact unknown, we act as if $\bar{U}_m^b$ has the Student-t distribution with $m-1$ degrees of freedom. Thus, for $m = 20$, a two-sided 95% confidence interval estimate for $E[U^{(i)}]$ based on the method of batch means applied to the $n = mk$ observations is

$$\bar{U}_m^{(i,b)} \pm 2.09\sqrt{\bar{\sigma}^2_{U_{20}^{(i,b)}}/19} = \bar{U}_m^{(i,b)} \pm 0.48\bar{\sigma}_{U_{20}^{(i,b)}}. \tag{28}$$

Given an estimated two-sided 95% confidence interval $[a_1, a_2]$ for $E[U^{(i)}]$ as in (6), the interval $[1/a_2, 1/a_1]$ provides an associated estimated two-sided 95% confidence interval for $\lambda_i = 1/E[U^{(i)}]$.

7. Remaining issues

Even though it is now common to have large data sets, there may not be enough data to fit a general model. The available data often must be reduced to obtain intervals in which the system can be regarded as approximately stationary. Even for a system that is approximately stationary over a long interval, having $2m$ parameters for $m+1$ states is likely to produce a model with too many parameters. If the actual state space is large, then the data for many states will be inadequate to obtain reliable rate estimates. The statistical analysis discussed in Section 6 should provide guidance. It often will be important to combine data over multiple subintervals and fit a more restrictive model for which the birth and death rates have structure. The references offer guidance; see, for example, Keiding (1975); Ross et al. (2007). It is natural to consider piecewise linear functions, in the spirit of the piecewise-linear estimates of the rate of a non-homogeneous Poisson process in Massey et al. (1996).

Acknowledgments

This research was supported by NSF grant CMMI 1066372. The author thanks a referee for a careful reading of the paper.

References

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y., Yom-Tov, G., 2011. Patient flow in hospitals: a data-based queueing-science perspective. Working Paper, NYU.
Bhat, U.N., Miller, G.K., Subba Rao, S., 1997. Statistical analysis of queueing systems. In: Dshalalow, J.H. (Ed.), Frontiers in Queueing Theory. CRC Press, Boca Raton, FL, pp. 351–394 (Chapter 13).
Billingsley, P., 1961. Statistical Inferences for Markov processes. University of Chicago Press, Chicago.
Bratley, P., Fox, B.L., Schrage, L.E., 1987. A Guide to Simulation, second ed. Springer, New York.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., Zhao, L., 2005. Statistical analysis of a telephone call center: a queueing-science perspective. J. Amer. Statist. Assoc. 100, 36–50.
Buzen, J.P., 1976. Fundamental operational laws of computer system performance. Acta Inform. 7, 167–182.
Buzen, J.P., 1978. Operational analysis: an alternative to stochastic modeling. In: Ferrari, D. (Ed.), Performance of Computer Installations. North-Holland, Amsterdam, pp. 175–194.
Buzen, J.P., Denning, P.J., 1980. Measuring and calculating queue length distributions. IEEE Trans. Comput. 18, 33–44.
Denning, P.J., Buzen, J.P., 1978. The operational analysis of queueing network models. Comput. Surv. 10, 225–261.
El-Taha, M., Stidham Jr., S., 1999. Sample-Path Analysis of Queueing Systems. Kluwer, Boston.
Israel, R.B., Rosenthal, J.S., Wei, J.Z., 2001. Fitting generators for Markov chains via empirical transition matrices, with applications to credit ratings. Math. Finance 11, 245–265.
Keiding, N., 1975. Maximum likelihood estimation in the birth-and-death process. Ann. Statist. 3, 363–372.
Keilson, J., 1979. Markov Chain Models—Rarity and Exponentiality. Springer, New York.
Massey, W.A., Parker, G.A., Whitt, W., 1996. Estimating the parameters of a nonhomogeneous Poisson process with linear rate. Telecommun. Syst. 5, 361–388.
Ross, S.M., 1996. Stochastic Processes, second ed. Wiley, New York.
Ross, S.M., 2010. Introduction to Probability Models, tenth ed. Elsevier, Amsterdam.
Ross, J.V., Taimre, T., Pollett, P.K., 2007. Estimation for queues from queue length data. Queueing Syst. 55, 131–138.
Whitt, W., 1992. Asymptotic formulas for Markov processes with applications to simulation. Oper. Res. 40, 279–291.
Whitt, W., 1999. Improving service by informing customers about anticipated delays. Manage. Sci. 45, 192–207.
Whitt, W., 2005. Engineering solution of a basic call-center model. Manage. Sci. 51, 221–235.
Wolff, R.W., 1965. Problems for statistical inference for birth and death queueing models. Oper. Res. 13, 343–357.