Parameter and State Estimation in Queues and Related Stochastic Models: A Bibliography

Azam Asanjarani, Yoni Nazarathy and Philip K. Pollett

The School of Mathematics and Physics, The University of Queensland,

Brisbane Australia

January 17, 2017

Abstract

This is an annotated bibliography on estimation and inference results for queues and related stochastic models. The purpose of this document is to collect and categorise works in the field, allowing for researchers and practitioners to explore the various types of results that exist.

This bibliography attempts to include all known works that satisfy both of these requirements:

- Works that deal with queueing models.

- Works that contain contributions related to methodology of parameter estimation, state estimation, hypothesis testing, confidence interval and/or actual data-sets of application areas.

Our attempt is to make this bibliography exhaustive, yet there are possibly some papers that we have missed. As it is updated continuously, additions and comments are welcomed.

The sections below categorise the works based on several categories. A single paper may appear in several categories simultaneously. The final section lists all works in chronological order along with short descriptions of the contributions.

This bibliography is maintained at

http://www.maths.uq.edu.au/~pkp/papers/Qest/Qest.html

and may be cited as such. We welcome additions and corrections.

There are additional works not mentioned in this bibliography that are mildly related. This includes methods for parameter estimation of point processes, methods for parameter estimation of stochastic matrix analytic models as well as inference, estimation and tomography of communication networks not directly modelled as queueing networks.

1   Overview of the Literature

According to our count, there are just under 300 references associated with parameter and/or state estimation of queues. The majority of these references are in the format of journal and/or conference articles. A few are surveys, textbooks, book chapters, Ph.D. theses, and significant related materials which are listed in the table below.

Surveys	[37] [36]
Textbooks	[43] [191] [201] [212] [34] [268]
Book chapters	[35] –Chapter 10, [115] –Section 6.7, [99] –Chapter 6, [253] –Chapter 10
Ph.D. theses	[113] [216] [136] [155]
Significant related materials	[38] (A book)

2   Classification by Model

Model	References
M/M/1	[64] [114] [74] [150] [4] [228] [188] [189] [169] [13][12] [186] [223] [238] [58] [244] [80] [54] [266] [185] [241] [218] [219] [55] [214]
M/M/2 (Heterogenous Servers)	[79] [91]
M/M/1/K	[188] [189] [5]

Model	References
M/M/1/∞ (FIFO)	[10]
M/M/c	[178] [122] [190] [254] [225] [190] [2] [267] [55]
M/M/c/N	[256] [194]
M/M/∞	[14] [182]
k-Par/M/1 (k-Par denotes a mixture of k Pareto distributions)	[216]
M/D/1	[203] [240]
M/E$_k$/1	[147] [272]
M/G/∞ (Random translation models in general)	[32] [46] [44] [215] [209] [40] [206] [182] [116] [183] [200] [260]
M/G/1	[74] [247] [130] [131] [141] [123] [102] [184] [81] [139] [41] [143] [19] [95] [50] [221] [61] [158] [159] [161] [198] [271] [53] [193] [55] [109]
M/GI/1	[1]
E$_k$/G/1	[95]
M/G/c	[76]

Model	References
M/G/c/C	[135]
G/G/1	[250] [222] [62] [55]
GI/M/1 (state-dependent arrival rate)	[73] [205] [145]
GI/M/1/n	[1]
GI/M/c	[265] [18] [137]
GI/G/1	[74] [31] [211] [27] [3] [144] [69] [28]
GI/G/2	[149]
GI/G/c	[166]
G/G/∞	[215]
GI/K/1 (service time has Matrix-Exponential distribution)	[16] [17] [231] [235]
GI/GI/s	[83]
GI/GI/∞	[83]
M/D/∞ (Deterministic service)	[112]
M/D/1	[178]
G/D/1	[127] [126]

Model	References
MAP/D/1	[49]
Erlang Loss System	[224]
Jackson Networks or More General Networks	[249] [22]
Systems Where Little's Law Holds	[107]

3   Classification by Sampling Regime

Sampling Regime	References
Full Observation	[213]
Observation at Discrete Points	[225] [190]
Probing	[56] [233] [5] [128] [129] [22] [65] [203] [154] [8]
Queue Inference Engine	[170] [103] [51] [81] [179] [98] [207]

4 Classification by Statistical Paradigm

Statistical Paradigm	References
Bayesian	[196] [220] [259] [188] [189] [51] [248] [234] [12] [13] [11] [14] [139] [223] [66] [151] [15] [238] [19] [68] [50] [18] [194] [58] [20] [157] [217] [246] [193] [218] [219] [214]
Maximum Entropy	[125]
Emphasis on the way of selecting sampling time	[31]
Non-parametric	[66] [40] [151] [67] [68] [156] [206] [187] [109]
Change point detection	[142]
Adaptive Control	[130] [260]
Sequential Inference	[146] [26] [251]
Perturbation analysis	[133]
Large Deviations	[106]

5 Classification by Application

Statistical Paradigm	References
Telephone Call Centres	[45] [110] [137] [101] [260] [21]
Manufacturing	[148] [50] [137] [54] [260]
Health Care	[193] [267]
Transportation	[50] [172] [174] [243] [269] [9]
Economics	[213]
ATM	[87] [81] [67] [69] [50]
Communication/Telecommunication Networks	[103] [22] [24] [192]
Network Traffic Modelling	[232] [180] [49] [175] [25] [239] [246]

6   Chronological order with brief descriptions

1955

Cox [70]: An overview paper of queueing theory outlining the philosophy of estimating parameters of input processes vs. performance processes.

1957

Benes [32]: Transient M/M/∞ full observation over a fixed interval.
Clarke [64]: M/M/1 MLE with full observation. The first paper. Sampling until "busy time" reaches a pre-assigned value yields closed form MLEs.

1961

Billingsley [39]: Book on inference of Markov chains.

1965

Cox [71]: An overview paper on parameter estimation, separate analysis of input and service mechanism and problems connected with the sampling of queueing process.
Kovalenko [167]:
Wolff [264]: Large sample theory for birth-death queues.

1966

Lilliefors [173]: Confidence intervals for standard performance measurements based on parameter error.

1967

Greenberg [114]: Different ways of determining for how long to observe a stationary M/M/1 queue (e.g. fixed number of arrivals, fixed total observation time, etc...

1968

Daley [73]: The serial correlation coefficients of queue sizes in a stationary GI/M/1 queue are studied.
Daley [74]: The serial correlation coefficients of a (stationary) sequence of waiting times in a stationary M/M/1, M/G/1 and GI/G/l queueing system are studied.

1970

Brown [46]: Estimating the G in M/G/∞ with arrival and departure times without known what customers they related to.
Ross [226]: Discusses identifying the distributions of GI/G/k uniquely based on observation of the queueing process.

1971

Pakes [205]: The serial correlation coefficients of waiting times in the stationary GI/M/1 queue is studied. (completing [73] work)

1972

Goyal and Harris[111]: MLE for queues with Poisson arrivals with state dependent general service times when queue sizes are observed at departure points.
Jenkins [150]: Compares the asymptotic variance of two estimators for M/M/1.
Muddapur [196]: Adds a prior distribution to Clarke's [64] approach.

1973

Harris [124]: An overview paper presented the statistical analysis of queueing systems with emphasis on the estimation of input and service parameters and/or distributions.

Neal Kuczura [199]: Presents accurate approximation and asymptotic approximations (by using renewal theory) for the variance of any differentiable functions of different traffic measurements.

Reynolds [220]: Bayesian approach for estimation of birth death parameters.

1974

Aigner [4]: Compares properties of various estimators for M/M/1 with cross-sectional data.

Brillinger [44]: Estimates parameter for a linear time invariant model that generalizes the G/G/∞ queue.

1975

Keiding [160]: Analyses asymptotic properties of the MLE for a birth-and-death process with linear rates (both birth and death).

1979

Thiagarajan and Harris [247]: Exponential goodness of fit test for the service times of M/G/1 based on observations of the queue lengths and/or waiting times.

1980

Dave and Shah [79]: MLE of an M/M/2 queue with heterogenous servers.

1981

Basawa and Prabhu [30]: Estimates for non-parametric and parametric models of single server queues over a horizon up to the nth departure epoch. Also the m.l.e's of the mean inter-arrival time and mean service time in an M/M/1 observed over a fixed time-interval.

Walrand [252]: Proposes an elementary justification of the filtering formulas for a Markov chain and an analysis of the arrival and departure processes at a ./M/1 queue in a quasi-reversible network.

Grassmann [112]: This paper shows that in the M/D/∞ queueing system with service time S, the optimal way to estimate the expected number in the system is by sampling the system at time $0, S, 2S, \cdots, kS$.

1982

Halfin[118]: Finds the minimum-variance linear estimator for the expected value of a stationary stochastic process, observed over a finite time interval, whose covariance function is a sum of decaying exponentials.

Schruben and Kulkarni [228]: Studies the interface between stochastic models and actual systems for the special case of M/M/1 queue.

1983

Hernàndez-Lerma and Marcus [130]: Adaptive control of an M/G/1 queueing system, where the control chooses the service rate as to minimize costs.

1984

Eschenbach [93]: Briefly describes methods and results in the statistical analysis of queueing systems.

Edelman and McKellar [91]: Comments on [79].

Hernàndez-Lerma and Marcus [131]: Adaptive control of priority assignment in a multi-class queue.

Machihara [178]: The carried traffic estimate errors for delay system models are analyzed with emphasis on the analysis of the effect of the holding time distribution on the estimate errors.

Warfield and Foers [259]: A Bayesian method for analysing teletraffic measurement data is discussed.

Woodside, Stanford and Pagurek[265]: Presents optimal mean square predictors for queue lengths and delays in the stationary GI/M/m queue, based on a queue length measurement.

1985

Armero [10]: The posterior distribution of traffic intensity and the posterior predictive distribution of the waiting time and number of customers for a M/M/1/∞ FIFO queue are obtained given two independent samples of arrival and service times.

Warfield and Foers [258]: Bayesian analysis for traffic intensity in M/M/c/K type models and in retrial models.

1986

Subba Rao and Harishchandra [242]: Large normal approximation test based for the traffic intensity parameter in GI/G/s queues.

1987

Bhat and Rao [37]: A first major survey on statistical analysis of queueing systems.

Mcgrath, Gross and Singpurwalla [188]: Attempts to illustrate Bayesian approach through M/M/1 and M/M/1/K examples.

McGrath and Singpurwalla [189]: This is part II to [188] (without Gross). Here the the focus is on integrating the "Shannon measure of information"(cross-entropy) in the analysis.

Ramalhoto [215]: Discusses estimation of generalizations of GI/G/∞, i.e. random translations whose distribution is parameterized by a certain function, $h(\cdot)$.

1988

Basawa and Prabhu [31]: Estimation of GI/G/1 with exponential family densities. Full observation over $[0, T]$ where T is a stopping time. Several T's considered and asymptotic properties compared.

Chen, Harrison, Mandelbaum, Van Ackere, Wein, [52]: Empirical evaluation of a queueing network model for semiconductor wafer fabrication.

Harishchandra and Rao [123]: Inference for the M/E_k/1 queue.

Jain and Templeton [145]: Estimation of GI/M/1 (and GI/M/1/m with m known) parameters where the arrival rate is either λ or λ_1 depending on the queue level.

Nozari and Whitt [204]: Propose an indirect approach to estimate average production intervals (the length of time between starting and finishing work on each product) using work-in-process inventory measurements.

1989

Fendick and Whitt [96]: Proposes measurments and approximations to describe the variability of offered traffic to a queue (the variability of the arrival process together with the service requirements) and predictes the average workload in the queue (which assumed to have a single server, unlimited waiting space and a work-conserving service descipline).
Glynn and Whitt [107]: Using the little's $L = \lambda W$ and generalizations to infer L from W and the oppositive.
Hantler and Rosberg [122]: Parmaeter estimation of M/M/c queue with parameters in stochastic varying environment, first doing the constant invariant derivation and then using in conjunction with Kalman filter for time-varying case.
Jain and Templeton [146]: Sequential analysis view for M/E_k/1 queues.
Sengupta [231]: Present an algorithm for computing the steady-state distribution of the waiting time and queue length of the stable GI/K/l queue.

1990

Gaver and Jacobs [102]: Transient M/G/1 inference.
Gawlick [103]: Applies the (QIE) to ethernet data.
Larson [170]: This deals with "State Reconstruction" as opposed to "parameter inference" in what is called the "Queue Inference Engine" (QIE). This is the first of many papers on the idea of using transactional data to reconstruct an estimate of the queue length process.
Rubin and Robson [227]: Inference and estimation of number of arrivals for a queueing system with losses due to bulking and a server that works a fixed shift and stays to work after the shift. Small sample analysis as opposed to asymptotic properties.

1991

Hall and Larson [120]: Modifies (extends) the QIE [170] to finite queues and to a case where there is data about exceeding a certain level.
Jain and Templeton [147]: Confidence intervals for estimation for M/M/2 with heterogenous servers.
Jain [140]: Compares confidence intervals for ρ using several methods and sampling regimens in M/E_k/1 queues.
Larson [171]: An addendum to [170] reducing the computational complexity from $O(N^5)$ to $O(N^3)$.
Thiruvaiyaru, Basawa and Bhat[249]: Large sample theory for MLEs of Jackson networks.

1992

Asmussen [16]: proves that the stationary waiting time in a GI/PH/1 queue with phase-type service time is phase-type.
Asmussen and Bladt [17]: The Matrix-exponential distribution is introduced and some of its basic structural properties is given. Further, an algorithm for computing the waiting time distribution of a queue with matrix-exponential service times and general inter-arrival times is given. This algorithm is a slightly generalization of the algorithm for computing the waiting time distribution of $GI/PH/1$ queues.
Basawa and Bhat [26]: Presents sequential analysis methods for the traffic intensity of single server queues.

Bertsimas and Servi [33]: Improves on the $O(n^5)$ algorithm in [170] to $O(n^3)$. Also presents an on-line algorithm for estimating the queue length after each departure and includes time-varying Poisson generalizations.

Daley and Servi [75]: Continues the track of the QIE, using taboo probabilities.

Heyde [132]: Quasi-likelihood estimation methods for stationary processes and queueing examples.

Jain [141]: Derives the relative efficiency of a parameter for the M/G/1 queueing system based on reduced and full likelihood functions. In addition, Monte Carlo simulations were carried out to study the finite sample properties for estimating the parameters of a M/G/1 queueing system.

Kumar [169]: Studies the bias in the means of average idle time and average queue length estimates, over the interval [0, t], in a transient M/M/1 queue.

Singpurwalla [234]: A discussion paper about [248]. The same issue fo QUESTA also has a rejoinder for the discussion.

Thiruvaiyaru and Basawa [248]: Discusses empirical Bayes estimation for variations of M/M/1 queues and Jackson networks.

1993

Daley and Servi [76]: Discusses a fairly general Markov chain setting for describing a stochastic process at intermediate time points r in $r \in (0, n)$ conditional on certain known behaviour of the process both on the interval and at the endpoints 0 and n.

Glynn, Melamed and Whitt [105]: Constructs confidence intervals for estimators and perform statistical tests by establishing a joint central limit theorem for customer and time averages by applying a martingle central limit theorem in a Markov framework.

1994

Armero and Bayarri [13]: Presents a Bayesian approach to predict several quantities in an M/M/1 queue in equilibrium.

Armero and Bayarri, M.J. [12]: Bayesian "prediction" in M/M/1 queues is considered. The meaning is Bayesian inference for steady state quantities such as the distribution of queue lengths.

Armero [11]: Another Bayesian inference paper.

Chandra and Lee [51]: Presents Bayesian methods for inferring customer behavior from transactional data in telecommunications systems.

Chen, Walrand and Messerschmitt [56]: Perhaps the first "probing" paper. Deals with arrivals a deterministic service time queue and estimates the Poisson arrival rates based on probe delays.

Jang and Liu [148]: Presents a new queueing formula applicable in manufacturing which uses variables easier to estimate than the variance such as the number of machine idle periods.

Jones and Larson [153]: Develops an efficient algorithm for event probabilities of order statistics and uses it for the queue inference engine ([170]).

Pitts [211]: Analysis of non-parametric estimation of the GI/G/1 queue input distributions based on observation of the waiting time.

1995

Duffield, Lewis, O'Connell, Russell, and Toomey [87]: Estimates directly the thermodynamic entropy of the data-stream at an input-port. From this, the quality-of-service parameters can be calculated rapidly.

Jain [142]: Change point detection in an M/M/1 queue.

Muthu and Sampathkumar [197]: The maximum likelihood estimates of the parameters involved in a finite capacity priority queueing model are obtained. The precision of the maximum likelihood estimates is studied using likelihood theory for Markov processes.

Masuda [186]: Provides sufficient conditions under which the intuition (based on partial observations) can be justified, and investigates related properties of queueing systems.

1996

Basawa, Bhat and Lund [27]: MLE for GI/G/1 based on waiting time data.

Dimitrijevic [81]: Considers the problem of inferring the queue length of an M/G/1 queue using transactional data of a busy period.

Manjunath and Molle [184]: Introduces a new off-line estimation algorithm for the waiting times of departing customers in an M/G/1 queue with FCFS service by decoupling the arrival time constraints from the customer departure times.

Sohn [236]: Simple M/M/1 Bayesian parameter estimation.

Sohn [237]: Bayesian estimation of M/M/1 using several competing methods.

1997

Armero and Bayarri [14]: Bayesian inference of $M/M/\infty$.

Basawa, Lund, and Bhat [29]: Extends [27] using estimating functions.

Bhat, Miller and Rao [36]: A survey paper, a decade after the previous Survey by Bhat and Rao, [37].

Daley and Servi [77]: Computes the distributions and moments of waiting times of customers within a busy period in a FCFS queuing system with a Poisson arrival process by exploiting an embedded Markov chain.

Glynn and Torres [106]: Deals with estimation of the tail properties of the workload process in both the M/M/1 queue and queues with more complex arrivals such as MMPP.

Ho and Cassandras [133]: A survey on perturbation theory.

Pickands and Stine [209]: Discrete time $M/G/\infty$ queue.

Toyoizumi [250]: Waiting time inference in G/G/1 queues in a non-parametric manner using "Sengupta's invariant relationship".

1998

Daley and Servi [78]: Computes the distributions and moments of waiting times of customers within a busy period in a FCFS queuing system with a Poisson arrival process by exploiting an embedded Markov chain.

Ganesh, Green, O'Connell and Pitts [100]: Appears like a "visionary" paper on the use of non-parametric Bayesian methods in network management.

Insua, Wiper and Ruggeri [139]: Bayesian inference for M/G/1 queues with either Erlang or hyper-exponential service distributions.

Mandelbaum and Zeltyn [179]: Queuing inference estimation in networks.

Rodrigues and Leite [223]: A Bayesian inference about the traffic intensity in an M/M/1 queue, without worrying about nuisance parameters.

Sharma and Mazumdar [233]: Proposes several schemes that the call acceptance controller, at entering node of an ATM network, can use to estimate the traffic of the users on the various routes in the network by sending a probing stream.

Wiper [263]: Perhaps complements [139] with analysis of G/M/c queues with the G being Erlang or hyper-exponential renewal processes.

1999

Acharya [3]: Analyses the rate of convergence of the distribution of MLEs in GI/G/1 queues with assumptions on the distributions as being from exponential families.

Conti [66]: Baysian inference for a Geo/G/1 Discrete time queue.

Bingham and Pitts [40]: Non-parameteric estimation in M/G/∞ queues.

Bingham and Pitts [41]: Estimates the arrival rate of an M/G/1 queue given observations of the busy and idle periods of this queue.

Jones [151]: Analyses queues in the presence of balking, using only the service start and stop data utilized in Larson's Queue Inference Engine.

Rodrigo and Vazquez [222]: Analyses a general G/G/1 retrial queueing systems from a statistical viewpoint.

Sharma [232]: Using the measurement tools available in Internet, suggests and compares different estimators to estimate aggregate traffic intensities at various nodes in the network.

2000

Armero and Conesa [15]: Statistical analysis of bulk arrival queues from a Bayesian point of view.

Duffield [86]: Analyses the impact of measurement error within the framework of Large Deviation theory.

Glynn and Zeevi [108]: Estimates tail probabilities in queues.

Jain [143]: Sequential analysis.

Jain and Rao [144]: Investigates the problems of statistical inference for the GI/G/1 queueing system.

2001

Alouf, Nain and Towsley [5]: Probing estimation for M/M/1/K queues using moment based estimators based on a variety of computable performance measures.

Huang and Brill [135]: Deriving the minimum variance unbiased estimator (MVUE) and the maximum likelihood estimator (MLE) of the stationary probability function of the number of customers in a collection of independent M/G/c/c subsystems.

Jang, Suh and Liu [149]: Presents a new GI/G/2 queueing formula which uses a slightly different set of data easier to obtain than the variance of inter-arrival time.

Paschalidis and Vassilaras [208]: Buffer overflow probabilities in queues with correlated arrival and service processes using large deviations.

2002

Conti [67]: Non-parametric statistical analysis of a discrete-time queueing system is considered and estimation of performance measures of the system is studied.

Conti and De Giovanni [69]: Considers performance evaluation of a discrete-time GI/G/1 queueing model with focus on the equilibrium distribution of the waiting time.

Sohn [238]: Even though the title has "Robust"', this paper appears to be a standard M/M/1 Baysian inference paper using the input data.

Zhang, Xia, Squillante and Mills [270]: A general approach to infer the per-class service times at different servers in multi-class queueing models.

2003

2004

Ausín, Wiper and Lillo [19]: Bayesian inference of M/G/1 using phase type representations of the G.

Conti [68]: A Bayesian non-parametric approach to the analysis of discrete-time queueing models.

Fearnhead [95]: Using forward-backward algorithm to do inference for M/G/1 and Er/G/1 queues.

Hall and Park [119]: An M/G/∞ non-parametric paper.

Wang, Chen, and Ke [256]: Maximum likelihood estimates and confidence intervals of an M/M/R/N queue with balking and heterogeneous servers.

2005

Bladt and Sørensen [42]: Likelihood inference for discretely observed Markov jump processes with finite state space.

Brown, Gans, Mandelbaum, Sakov, Shen, Zeltyn and Zhao [45]: Major paper dealing with telephone call centre data analysis.

Hei, Bensaou and Tsang [128]: Probing focusing on the inter-departure SCV of the probing stream in tandem finite buffer queues.

Mandjes and van de Meent [180]: Propose an approach to accurately determine burstiness of a network link on small time-scales (for instance 10 ms), by sampling the buffer occupancy (for instance) every second.

Prieger [213]: Shows that the MLE based on the complete inter-arrival and service times (IST) dominates the MLE based on the number of units in service (NIS), in terms of ease of implementation, bias, and variance.

Ross and Shanthikumar [224]: Estimating effective capacity in Erlang loss systems under competition.

Neuts [202]: Reflections on statistical methods for complex stochastic systems.

2006

Castellanos, Morales, Mayoral, Fried and Armero [50]: Develops a Bayesian analysis of queueing systems in applications of the machine interference problem, like job-shop type systems, telecommunication traffic, semiconductor manufacturing or transport.

Chick [57]: A survey chapter on subjective probability and the Bayesian approach, specifically in Monte-Carlo simulation, yet gives some insight into queueing inference.

Chu and Ke [60]: Construction of confidence intervals of mean response time for an M/G/1 FCFS queueing system.

Doucet, Montesano Jasra [85]: Presents a trans-dimensional Sequential Monte Carlo method for online Bayesian inference in partially observed point processes.

Hansen and Pitts [121]: Non-parameteric estimation of the service time distribution and the traffic intensity in M/G/1 queues based on observations of the workload.

Hei, Bensaou and Tsang [129]: Similar to [128] but here the focus is on inter-departure SCV of the two consecutive probing packets.

Ke and Chu [156]: Proposes a consistent and asymptotically normal estimator of intensity for a queueing system with distribution-free inter-arrival and service times.

Liu, Heo, Sha and Zhu [176]: Proposes a queueing-model-based adaptive control approach for controlling the performance of computing systems.

Liu, Wynter, Xia and Zhang [177]: presents an approach for solving the problem of calibration of model parameters in the queueing network framework using inference techniques.

Rodrigo [221]: Analyse the M/G/1 retrial queue from a statistical viewpoint.

Wang, Ke, Wang and Ho [254]: Studies MLE and confidence intervals of an M/M/R queue with heterogeneous servers under steady-state conditions.

2007

Ausín, Lillo and Wiper [18]: Considers the problem of designing a GI/M/c queueing system.

Chu and Ke [61]: Proposes a consistent and asymptotically normal estimator of the mean response time for a G/M/1 queueing system, which is based on empirical Laplace function.

Chu and Ke [62]: Estimation and confidence interval of mean response time for a G/G/1 queueing system using data-based recursion relation and bootstrap methods.

Morales, Castellanos, Mayoral, Fried and Armero [194]: Exploits Bayesian criteria for designing an M/M/c//r queueing system with spares.

Park [206]: The use of auxiliary functions in non-parametric inference for the M/G/∞ queueing model is considered.

Ross, Taimre and Pollett [225]: Estimation of rates in M/M/c queues using observations at discrete queues and MLE estimates of an approximate Orenstein Ullenbeck (OU) process.

2008

Ausín, Wiper and Lillo [20]: Bayesian inference for the transient behaviour and duration of a busy period in a single server queueing system with general, unknown distributions for the inter-arrival and service times is investigated.

Basawa, Bhat and Zhou [28]: Parameter estimation based on the differences of two positive exponential family random variables is studied.

Casale, Cremonesi and Turrin [48]: Proposed service time estimation techniques based on robust and constrained optimization.

Casale, Zhang and Smirni [49]: Several contributions to the Markovian traffic analysis.

Choudhury and Borthakur [58]: Bayesian-based techniques for analysis of the M/M/1 queueing model.

Dey [80]: Bayes' estimators of the traffic intensity and various queue characteristics in an M/M/1 queue under quadratic error loss function have been derived.

Ke, Ko and Sheu [159]: Proposes an estimator for the expected busy period of a controllable M/G/1 queueing system in which the server applies a bicriterion policy during his idle period.

Ke, Ko and Chiou [158]: Presents a sensitivity investigation of the expected busy period for a controllable M/G/1 queueing system by means of a factorial design statistical analysis.

Kim and Park [166]: Introduces methods of queue inference which can find the internal behaviours of queueing systems with only external observations, arrival and departure time.

Sutton and Jordan [244]: Analysing queueing networks from the probabilistic modelling perspective, applying inference methods from graphical models that afford significantly more modelling flexibility.

Ramirez, Lillo and Wiper [216]: Considers a mixture of two-parameter Pareto distributions as a model for heavy-tailed data and use a Bayesian approach based on the birth-death Markov chain Monte Carlo algorithm to fit this model.

2009

Baccelli, Kauffmann and Veitch [22]: Points out the importance of inverse problems in queueing theory, which aim to deduce unknown parameters of the system based on partially observed trajectories.

Baccelli, Kauffmann and Veitch [23]: Evaluates the algorithm proposed in [22].

Comert and Setin [65] Presents a real-time estimation of queue lengths from the location information of probe vehicles in a queue at an isolated and under-saturated intersection.

Chu and Ke [63]: Constructs confidence intervals of intensity for a queueing system, which are based on four different bootstrap methods.

Duffy and Meyn [89]: Conjectures and presents support for this: a consistent sequence of non-parametric estimators can be constructed that satisfies a large deviation principle.

Gorst-Rasmussen Hansen [110]: Proposes a framework based on empirical process techniques for inference about waiting time and patience distributions in multi-server queues with abandonment.

Heckmuller and Wolfinger [127]: Proposes methods to estimate the parameters of arrival processes to G/D/1 queueing systems only based on observed departures from the system.

Ibrahim and Whitt [137]: Studies the performance of alternative real-time delay estimators based on recent customer delay experience.

Kiessler and Lund [161]: A note that considers traffic intensity estimation in the classical M/G/1 queue.

Ke and Chu [157]: Proposes a consistent and asymptotically normal estimator of intensity for a queueing system with distribution-free inter-arrival and service times.

Kraft, Pacheco-Sanchez, Casale and Dawson [168]: Proposes a linear regression method and a maximum likelihood technique for estimating the service demands of requests based on measurement of their response times instead of their CPU utilization.

Liu, Wu, Ma, and Hu [175]: Presents an approach to estimate time-dependent queue length even when the signal links are congested.

Mandjes and Żuraniewski [182]: Develops queueing-based procedures to (statistically) detect overload in communication networks, in a setting in which each connection consumes roughly the same amount of bandwidth.

Mandjes and van De Meent [181]: The focus is on dimensioning as the approach for delivering performance requirements of network.

Nam, Kim and Sung [198]: Estimates the available bandwidth for an M/G/1 queueing system.

Novak and Watson [203]: Presents a technique to estimate the arrival rate from delay measurements, acquired using single-packet probing.

2010

Chen and Zhou [54]: Proposes a non-linear quantile regression model for the relationship

between stationary cycle time quantiles and corresponding throughput rates of a manufacturing system.

Duffy and Meyn [88]: Deals with large deviations showing that in broad generality, that estimates of the steady- state mean position of a reflected random walk have a high likelihood of over-estimation.

Frey and Kaplan [98]: Introduces an algorithms for queue inference problems involving periodic reporting data.

Gans, Liu, Mandelbaum, Shen, and Ye [101]: Studies operational heterogeneity of call center agents where the proxy for heterogeneity is agents' service times (call durations).

Heckmueller and Wolfinger [126]: Proposes methods to estimate the parameters of arrival processes to G/D/1 queueing systems only based on observed departures from the system.

Pin, Veitch and Kauffmann [210]: Focuses on a specific delay tomographic problem on a multicast diffusion tree, where end-to-end delays are observed at every leaf of the tree, and mean sojourn times are estimated for every node in the tree.

Ramirez-Cobo, Lillo, Wilson and Wiper [217]: Presents a method for carrying out Bayesian estimation for the double Pareto lognormal distribution.

Sutton and Jordan [245]: Presents a viewpoint that combines queueing networks and graphical models, allowing Markov chain Monte Carlo to be applied.

Xu, Zhang and Ding [266]: Discusses testing hypotheses and confidence regions with correct levels for the mean sojourn time of an M/M/1 queueing system.

Zhang and Xu [271]: Discuss constructing confidence intervals of performance measures for an M/G/1 queueing system.

Zuraniewski, Mandjes and Mellia [273]: Explores techniques for detecting unanticipated load changes with focus on large-deviations based techniques developed earlier in [182].

2011

Abramov [1]: Statistical bounds for certain output characteristics of the M/GI/1/n and GI/M/1/n loss queueing systems are derived on the basis of large samples of an input characteristic of these systems.

Amani, Kihl and Robertsson [6]: An applications paper to computer systems.

Ban, Hao, and Sun [25]: Studies how to estimate real time queue lengths at signalized intersections using intersection travel times collected from mobile traffic sensors.

Chen, Nan, Zhou [53]: Investigates the statistical process control application for monitoring queue length data in M/G/1 systems.

Feng, Dube, and Zhang [97]: Considers estimation problems in G/G/∞ queue under incomplete information. Specifically, where it is infeasible to track each individual job in the system and only aggregate statistics are known or observable.

Grübel and Wegener [116]: In M/G/∞ systems, considers the matching and exponentiality problems where the only observations are the order statistics associated with the departure times and the order in which the customers arrive and depart, respectively.

Ibrahim and Whitt [138]: Develops real-time delay predictors for many-server service systems with a time-varying arrival rate, a time-varying number of servers, and customer abandonment.

Mandjes and Zuraniewski [183]: M/G/∞ change point detection using large deviations.

Manoharan and Jose [185]: Considers an M/M/1 queueing system with the customer impatience in the form of random balking.

McCabe, Martin and Harris [187]: Presents an efficient probabilistic forecasts of integer-valued random variables that can be interpreted as a queue, stock, birth and death process or branching process.

Park, Kim and Willemain [207]: Proposes new approaches that can analyse the unobservable queues using external observations.

Sousa-Vieira [239]: Considers the suitability of the $M/G/\infty$ process for modelling the spatial and quality scalability extensions of the H.264 standard in video traffic modelling.

Srinivas, Rao and Kale [241]: Maximum likelihood and uniform minimum variance unbiased estimators of measures in the M/M/1 queue are obtained and compared.

Sutton and Jordan [246]: A Bayesian inference paper by computer systems researchers.

2012

Duffy and Meyn [90]: Large deviation asymptotics for busy periods for a queue.

Fabris-Rotelli and [94]: A historical and theoretical overview of G/M and M/G queueing processes.

Hu and Lee [134]: Consider a parameter estimation problem when the state process is a reflected fractional Brownian motion (RFBM) with a non-zero drift parameter and the observation is the associated local time process.

Jones [152]: Remarks on queue inference from departure data alone and the importance of the queue inference engine.

Kauffmann [154]: Proposes a new approach, on the basis of existing TCP connections and reaching therefore a zero probing overhead based on the theory of inverse problems in bandwidth sharing networks.

Kim and Whitt [162]: Statistical analysis with Little's law.

Kim and Whitt [163]: Contains supplementary martial to [162].

Kim and Whitt [165]: Estimating waiting times with the time-varying Little's law.

Kim and Whitt [164]: Appendix to [165].

McVinish and Pollett [190]: Uses estimating equations to get estimators for M/M/c queues and related models. Performance is compared to [225].

Mohammadi and Salehi-Rad [193]: Exploits the Bayesian inference and prediction for an M/G/1 queuing model with optional second re-service.

Nelgabats, Nov and Weiss [200]: $M/G/\infty$ estimation.

Ren and Li [219]: Bayesian estimator of the traffic intensity in an M/M/1 queue is derived under a new weighted square error loss function.

Ren and Wang [218]: Similer to [219]. Bayesian estimators of the traffic intensity in an M/M/1 queue are derived under a precautionary loss function.

Whitt [261]: Fitting birth-and-death queueing model to data.

2013

Acharya, Rodrguez-Snchez and Villarreal-Rodrguez [2]: Presents the derivation of maximum likelihood estimates for the arrival rate and service rates in a stationary M/M/c queue with heterogeneous servers.

Chow [59]: Analysis of queueing model based on chaotic mapping.

Li, Chen, Li and Zhang [172]: Proposes a new algorithm based on the temporal–spatial queueing model to describe the fast travel-time variations using only the speed and headway time series that is measured at upstream and downstream detectors.

2014

Azriel, Feigin and Mandelbaum [21]: Proposes a new model called Erlang-S, where "S" stands for Servers where there is a pool of present servers, some of whom are available to serve customers from the queue while others are not, and the process of becoming available or unavailable is modelled explicitly.

Dinha, Andrewa and Nazarathy [82]: A conceptual and numerical contribution on design and control of speed-scaled systems in view of parameter uncertainty.

He, Li, Huang and Lei [125]: Considers the queuing system as a black box and derive a performance index for the queuing system by the principle of maximum entropy only on the assumption that the queue is stable.

Senderovich, Weidlich, Gal, and Mandelbaum [260]: Establish a queueing perspective in operational process mining and demonstrates the value of queue mining using the specific operational problem of online delay prediction.

Yom-Tov and Mandelbaum [267]: Analyses a queueing model, where customers can return to service several times during their sojourn within the system.

2015

Bakholdina1 and Gortsev [24]: Focused on the problem of optimal estimation of the states of the modulated semi-synchronous integrated flow of events.

Burkatovskaya, Kabanova, and Vorobeychikov [251]: CUSUM algorithms for parameter estimation in queueing systems where the arrival process is an Markov-modulated Poisson process.

Cahoy, Polito, and Phoha [47]: Statistical analysis of fractional M/M/1 queue and fractional birth-death processes; the point processes governed by difference differential equations containing fractional derivative operators.

Chen and Zhou [55]: Propose the cumulative sum (CUSUM) schemes to efficiently monitor the performance of typical queueing systems based on different sampling schemes.

Dong and Whitt [83]: Explores a stochastic grey-box modelling of queueing systems by fitting birth-and-death processes to data.

Dong and Whitt [84]: Using a birth-and-death process to estimate the steady-State distribution of a periodic queue.

Efrosinin, Winkler, and Martin [92]: Considers the problem of estimation and confidence interval construction of a Markovian controllable queueing system with unreliable server and constant retrial policy.

Goldenshluger [109]: Non-parametric estimation of service time distribution of the M/G/1 queue from incomplete data on the queue.

Gurvich, Huang and Mandelbaum [117]: Proposes a diffusion approximation for a many-server Erlang-A queue.

Liu, Wu, and Michalopoulos [174]: Improves queue size estimation by proposing different ramp queue estimation algorithms.

Mohajerzadeh, Yaghmaee, and Zahmatkesh [192]: Proposed a method to prolong the network lifetime and to estimate the target parameter efficiently in wireless sensor networks.

Senderovich, Leemans, Harel, Gal, Mandelbaum, and van der Aalst [229]: Explores the influence of available information in the log on the accuracy of the queue mining techniques.

Senderovich, Weidlich, Gal, Mandelbaum [230]: Queue mining for delay prediction in multi-class service processes.

Srinivas and Kale [240]: Compares the Maximum Likelihood (ML) and Uniformly Minimum Variance Unbiased (UMVU) estimation for the M/D/1 queueing systems.

Sutartoa and Joelianto [243]: Presents an overview of urban traffic flow from the perspective of system theory and stochastic control.

Wang and Casale [255]: Proposes maximum likelihood (ML) estimators for service demands in closed queueing networks with load-independent and load-dependent stations.

Wang, Prez, and Casale [257]: A software for parameter estimation.

Whitt [262]: Sequel to [83] and [85]. Establishes many-server heavy-traffic fluid limits for the steady-state distribution and the fitted birth and death rates in periodic $Mt/GI/infty$ models.

2016

Amini, Pedarsani, Skabardonis, and Varaiya [7]: Queue-length estimation using real-time traffic data.

Antunes, Jacinto, Pacheco, and Wichelhaus [8]: Uses a probing strategy to estimate the time dependent traffic intensity in an $M_t/G_t/1$ queue, where the arrival rate and the general service-time distribution change from one time interval to another, and derive statistical properties of the proposed estimator.

Anusha, Sharma, Vanajakshi, Subramanian, and Rilett [9]: Develops a model-based scheme to estimate the number of vehicles in queue and the total delay.

Cruz, Quinino and Ho [72]: Uses a Bayesian technique, the sampling/importance resampling method to estimate the parameters of multi-server queueing systems in which inter-arrival and service times are exponentially distributed.

Ghorbani-Mandolakani and Salehi Rad [104]: Derives the ML and Bayes estimators of traffic intensity and asymptotic confidence intervals for mean system size of a two-phase tandem queueing model with a second optional service and random feedback and two heterogeneous servers.

Morozov, Nekrasova, Peshkova, and Rumyantsev[195]: Develops a novel approach to confidence estimation of the stationary measures in high performance multi-server queueing systems.

Quinino and Cruz [214]: Describes a Bayesian method for sample size determination for traffic intensity estimation.

Zammit1, Fabri1 and Scerri1 [269]: A self-estimation algorithm is presented to jointly estimate the states and model parameters.

Zhang, Xu, and Mi [272]: Considers the hypothesis tests of performance measures for an $M/E_k/1$ queueing system.

References

[1] V.M. Abramov. Statistical analysis of single-server loss queueing systems. *Methodology and Computing in Applied Probability*, pages 1–19.

[2] S. K. Acharya, S. V. Rodríguez-Sánchez, and C. E. Villarreal-Rodríguez. Maximum likelihood estimates in an M/M/c queue with heterogeneous servers. *International Journal of Mathematics in Operational Research*, 5(4):537–549, 2013.

[3] S.K. Acharya. On normal approximation for maximum likelihood estimation from single server queues. *Queueing Systems*, 31(3):207–216, 1999.

[4] D.J. Aigner. Parameter estimation from cross-sectional observations on an elementary queuing system. *Operations Research*, 22(2):422–428, 1974.

[5] S. Alouf, P. Nain, and D. Towsley. Inferring network characteristics via moment-based estimators. In *INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 2, IEEE: 1045–1054. 2001.

[6] P. Amani, M. Kihl, and A. Robertsson. Multi-step ahead response time prediction for single server queuing systems. In *Computers and Communications (ISCC), 2011 IEEE Symposium on*: 950–955. IEEE, 2011.

[7] Z. Amini, R. Pedarsani, A. Skabardonis, and P. Varaiya. Queue-length estimation using real-time traffic data. In *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, IEEE: 1476–1481, 2016.

[8] N. Antunes, G. Jacinto, A. Pacheco , and C. Wichelhaus. Estimation of the traffic intensity in a piecewise-stationary $M_t/G_t t/1$ queue with probing. *ACM SIGMETRICS Performance Evaluation Review*, 44(2):3–5, 2016.

[9] S.P. Anusha, A. Sharma, L. Vanajakshi, S.C. Subramanian, and L.R. Rilett. Model-based approach for queue and delay estimation at signalized intersections with erroneous automated data. *Journal of Transportation Engineering*, 142(5):04016013, 2016.

[10] C. Armero. Bayesian analysis of $M/M/1/\infty/FIFO$ queues. *Bayesian statistics*, 2:613–618, 1985.

[11] C. Armero. Bayesian inference in Markovian queues. *Queueing Systems*, 15(1):419–426, 1994.

[12] C. Armero and M.J. Bayarri. Bayesian prediction in M/M/1 queues. *Queueing Systems*, 15(1):401–417, 1994.

[13] C. Armero and M.J. Bayarri. Prior assessments for prediction in queues. *The Statistician*: 139–153, 1994.

[14] C. Armero and M.J. Bayarri. A bayesian analysis of a queueing system with unlimited service. *Journal of Statistical Planning and Inference*, 58(2):241–261, 1997.

[15] C. Armero and D. Conesa. Prediction in Markovian bulk arrival queues. *Queueing Systems*, 34(1):327–350, 2000.

[16] S. Asmussen. Phase-type representations in random walk and queueing problems. *The Annals of Probability*: 772–789, 1992.

[17] S. Asmussen and M. Bladt. *Renewal Theory and Queueing Algorithms for Matrix-exponential Distributions*. University of Aalborg, Institute for Electronic Systems, Department of Mathematics and Computer Science, 1992.

[18] M.C. Ausín, R.E. Lillo, and M.P. Wiper. Bayesian control of the number of servers in a GI/M/c queueing system. *Journal of Statistical Planning and Inference*, 137(10):3043–3057, 2007.

[19] M.C. Ausín, M.P. Wiper, and R.E. Lillo. Bayesian estimation for the M/G/1 queue using a phase-type approximation. *Journal of Statistical Planning and Inference*, 118(1-2):83–101, 2004.

[20] M.C. Ausín, M.P. Wiper, and R.E. Lillo. Bayesian prediction of the transient behaviour and busy period in short-and long-tailed GI/G/1 queueing systems. *Computational Statistics & Data Analysis*, 52(3):1615–1635, 2008.

[21] D. Azriel, P. D. Feigin, and A. Mandelbaum. Erlang s: A data-based model of servers in queueing networks. Technical report, Working paper, 2014.

[22] F. Baccelli, B. Kauffmann, and D. Veitch. Inverse problems in queueing theory and internet probing. *Queueing Systems*, 63(1):59–107, 2009.

[23] F. Baccelli, B. Kauffmann, and D. Veitch. Towards multihop available bandwidth estimation. *ACM SIGMETRICS Performance Evaluation Review*, 37(2):83–84, 2009.

[24] M. A. Bakholdina and A. M. Gortsev. Optimal estimation of the states of modulated semi-synchronous integrated flow of events in condition of its incomplete observability. *Applied Mathematical Sciences*, 9(29):1433–1451, 2015.

[25] X.J. Ban, P. Hao, and Z. Sun. Real time queue length estimation for signalized intersections using travel times from mobile sensors. *Transportation Research Part C: Emerging Technologies*, 19(6):1133–1156, 2011.

[26] I.V. Basawa and B.R. Bhat. Sequential inference for single server queues. *OXFORD STATISTICAL SCIENCE SERIES in book "Queueing and Related Models, Bhat, Basawa"*, pages 325–325, 1992.

[27] I.V. Basawa, U.N. Bhat, and R. Lund. Maximum likelihood estimation for single server queues from waiting time data. *Queueing Systems*, 24(1):155–167, 1996.

[28] I.V. Basawa, U.N. Bhat, and J. Zhou. Parameter estimation using partial information with applications to queueing and related models. *Statistics & Probability Letters*, 78(12):1375–1383, 2008.

[29] I.V. Basawa, R. Lund, and U.N. Bhat. Estimating function methods of inference for queueing parameters. *Lecture Notes-Monograph Series*: 269–284, 1997.

[30] I.V. Basawa and N.U. Prabhu. Estimation in single server queues. *Naval Research Logistics Quarterly*, 28(3):475–487, 1981.

[31] I.V. Basawa and N.U. Prabhu. Large sample inference from single server queues. *Queueing Systems*, 3(4):289–304, 1988.

[32] V.E. Benes. A sufficient set of statistics for a simple telephone exchange model. *Bell System Tech. J*, 36:939–964, 1957.

[33] D.J. Bertsimas and L.D. Servi. Deducing queueing from transactional data: the queue inference engine, revisited. *Operations Research*, 40: S217–S228, 1992.

[34] U.N. Bhat. *An introduction to queueing theory: modeling and analysis in applications*. Birkhäuser, 2015.

[35] U.N. Bhat. *An introduction to queueing theory: modeling and analysis in applications*. Birkhauser, 2008.

[36] U.N. Bhat, G.K. Miller, and S.S. Rao. Statistical analysis of queueing systems. *Chapter in: Frontiers in queueing*: 351–394, 1997.

[37] U.N. Bhat and S.S. Rao. Statistical analysis of queueing systems. *Queueing Systems*, 1(3):217–247, 1987.

[38] P. Billingsley. *Statistical inference for Markov processes*, volume 2. University of Chicago Press Chicago, 1961.

[39] P. Billingsley. Statistical methods in Markov chains. *The Annals of Mathematical Statistics*, pages 12–40, 1961.

[40] N.H. Bingham and S.M. Pitts. Non-parametric estimation for the M/G/∞ queue. *Annals of the Institute of Statistical Mathematics*, 51(1):71–97, 1999.

[41] N.H. Bingham and S.M. Pitts. Nonparametric inference from M/G/l busy periods. *Stochastic Models*, 15(2):247–272, 1999.

[42] M. Bladt and M. Sørensen. Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):395–410, 2005.

[43] A. Borovkov. *Stochastic processes in queueing theory*, volume 4. Springer Science & Business Media, 2012.

[44] D.R. Brillinger. Cross-spectral analysis of processes with stationary increments including the stationary G/G/∞ queue. *The Annals of Probability*, pages 815–827, 1974.

[45] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center. *Journal of the American Statistical Association*, 100(469):36–50, 2005.

[46] M. Brown. An M/G/∞ estimation problem. *The Annals of Mathematical Statistics*, 41(2):651–654, 1970.

[47] D. O. Cahoy, F. Polito, and V. Phoha. Transient behavior of fractional queues and related processes. *Methodology and Computing in Applied Probability*, 17(3):739–759, 2015.

[48] G. Casale, P. Cremonesi, and R. Turrin. Robust workload estimation in queueing network performance models. In *Parallel, Distributed and Network-Based Processing, 2008. PDP 2008. 16th Euromicro Conference on*, pages 183–187. IEEE, 2008.

[49] Giuliano Casale, Eddy Z Zhang, and Evgenia Smirni. *Interarrival times characterization and fitting for Markovian traffic analysis*. Citeseer, 2008.

[50] M.E. Castellanos, J. Morales, A.M. Mayoral, R. Fried, and C. Armero. *On Bayesian design in finite source queues*. Centro de Investigación Operativa, Universidad Miguel Hernández, 2006.

[51] K Chandrs and L.K. Jones. Transactional data inference for telecommunication models. In *presentation at First Annual Technical Conference on Telecommunications R & D in Massachusetts, University of Massachusetts, Lowell, Massachusetts*, 1994.

[52] H. Chen, J.M. Harrison, A. Mandelbaum, A. Van Ackere, and L.M. Wein. Empirical evaluation of a queueing network model for semiconductor wafer fabrication. *Operations Research*, pages 202–215, 1988.

[53] N. Chen, Y. Yuan, and S. Zhou. Performance analysis of queue length monitoring of M/G/1 systems. *Naval Research Logistics (NRL)*, 58(8):782–794, 2011.

[54] N. Chen and S. Zhou. Simulation-based estimation of cycle time using quantile regression. *IIE Transactions*, 43(3):176–191, 2010.

[55] N. Chen and S. Zhou. Cusum statistical monitoring of M/M/1 queues and extensions. *Technometrics*, 57(2):245–256, 2015.

[56] T.M. Chen, J. Walrand, and D.G. Messerschmitt. Parameter estimation for partially observed queues. *Communications, IEEE Transactions on*, 42(9):2730–2739, 1994.

[57] S.E. Chick. Subjective probability and bayesian methodology. *Chapter in: Handbooks in Operations Research and Management Science*, 13:225–257, 2006.

[58] A. Choudhury and A.C. Borthakur. Bayesian inference and prediction in the single server Markovian queue. *Metrika*, 67(3):371–383, 2008.

[59] J. Chow. On observable chaotic maps for queuing analysis. *Transportation Research Record: Journal of the Transportation Research Board*, (2390):138–147, 2013.

[60] Y.K. Chu and J.C. Ke. Confidence intervals of mean response time for an M/G/1 queueing system: Bootstrap simulation. *Applied Mathematics and Computation*, 180(1):255–263, 2006.

[61] Y.K. Chu and J.C. Ke. Interval estimation of mean response time for a G/M/1 queueing system: empirical laplace function approach. *Mathematical Methods in the Applied Sciences*, 30(6):707–715, 2007.

[62] Y.K. Chu and J.C. Ke. Mean response time for a G/G/1 queueing system: Simulated computation. *Applied Mathematics and Computation*, 186(1):772–779, 2007.

[63] Y.K. Chu and J.C. Ke. Analysis of intensity for a queueing system: bootstrapping computation. *International Journal of Services Operations and Informatics*, 4(3):195–211, 2009.

[64] A.B. Clarke. Maximum likelihood estimates in a simple queue. *The Annals of Mathematical Statistics*, 28(4):1036–1040, 1957.

[65] G. Comert and M. Cetin. Queue length estimation from probe vehicle location and the impacts of sample size. *European Journal of Operational Research*, 197(1):196–202, 2009.

[66] P.L. Conti. Large sample Bayesian analysis for Geo/G/1 discrete-time queueing models. *The Annals of Statistics*, 27(6):1785–1807, 1999.

[67] P.L. Conti. Nonparametric statistical analysis of discrete-time queues, with applications to ATM teletraffic data. *Stochastic Models*, 18(4):497–527, 2002.

[68] P.L. Conti. Bootstrap approximations for Bayesian analysis of Geo/G/1 discrete-time queueing models. *Journal of statistical planning and inference*, 120(1-2):65–84, 2004.

[69] P.L. Conti and L. De Giovanni. Queueing models and statistical analysis for ATM based networks. *Sankhyā: The Indian Journal of Statistics, Series B*: 50–75, 2002.

[70] D.R. Cox. The statistical analysis of congestion. *Journal of the Royal Statistical Society. Series A (General)*, 118(3):324–335, 1955.

[71] D.R. Cox. Some problems of statistical analysis connected with congestion. In *Proc. of the Symp. on Congestion Theory*, 1965.

[72] F.R.B Cruz, R.C. Quinino, and L.L. Ho. Bayesian estimation of traffic intensity based on queue length in a multi-server M/M/s queue. *Communications in Statistics-Simulation and Computation*, (just-accepted):00–00, 2016.

[73] D.J. Daley. Monte Carlo estimation of the mean queue size in a stationary GI/M/1 queue. *Operations Research*, pages 1002–1005, 1968.

[74] D.J. Daley. The serial correlation coefficients of waiting times in a stationary single server queue. *Journal of the Australian Mathematical Society*, 8(683-699):27, 1968.

[75] D.J. Daley and L.D. Servi. Exploiting Markov chains to infer queue length from transactional data. *Journal of Applied Probability*, 29(3):713–732, 1992.

[76] D.J. Daley and L.D. Servi. A two-point Markov chain boundary-value problem. *Advances in Applied Probability*, 25(3):607–630, 1993.

[77] D.J. Daley and L.D. Servi. Estimating waiting times from transactional data. *INFORMS Journal on Computing*, 9(2):224–229, 1997.

[78] D.J. Daley and L.D. Servi. Moment estimation of customer loss rates from transactional data. *Journal of Applied Mathematics and Stochastic Analysis*, 11(3):301–310, 1998.

[79] U. Dave and Y.K. Shah. Maximum likelihood estimates in a M/M/2 queue with heterogeneous servers. *Journal of the Operational Research Society*, 31(5):423–426, 1980.

[80] S. Dey. A note on bayesian estimation of the traffic intensity in M/M/1 queue and queue characteristics under quadratic loss function. *Data Science Journal*, 7(0):148–154, 2008.

[81] D.D. Dimitrijevic. Inferring most likely queue length from transactional data. *Operations Research Letters*, 19(4):191–199, 1996.

[82] T. V. Dinh, L. L.H. Andrew, and Y. Nazarathy. Architecture and robustness tradeoffs in speed-scaled queues with application to energy management. *International Journal of Systems Science*, 45(8):1728–1739, 2014.

[83] J. Dong and W. Whitt. Stochastic grey-box modeling of queueing systems: fitting birth-and-death processes to data. *Queueing Systems*, 79(3-4):391–426, 2015.

[84] J. Dong and W. Whitt. Using a birth-and-death process to estimate the steady-state distribution of a periodic queue. *Naval Research Logistics (NRL)*, 62(8):664–685, 2015.

[85] A. Doucet, L. Montesano, and A. Jasra. Optimal filtering for partially observed point processes using trans-dimensional sequential Monte Carlo. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, IEEE: V–V, 2006.

[86] N.G. Duffield. A large deviation analysis of errors in measurement based admission control to buffered and bufferless resources. *Queueing Systems*, 34(1):131–168, 2000.

[87] N. G. Duffield, J. T. Lewis, N. O'Connell, R. Russell, and F. Toomey. Entropy of ATM traffic streams: a tool for estimating QoS parameters. *Selected Areas in Communications, IEEE Journal on*, 13(6):981–990, 1995.

[88] K.R. Duffy and S.P. Meyn. Most likely paths to error when estimating the mean of a reflected random walk. *Performance Evaluation*, 67(12):1290–1303, 2010.

[89] K.R. Duffy and S.P. Meyn. Estimating Loynes exponent. *Queueing Systems*, 68(3-4):285–293, 2011.

[90] K.R. Duffy and S.P. Meyn. Large deviation asymptotics for busy periods. 2012.

[91] D.B. Edelman and D.E. McKellar. Comments on maximum likelihood estimates in a M/M/2 queue with heterogeneous servers. *Journal of the Operational Research Society*, 35(2):149–150, 1984.

[92] D. Efrosinin, A. Winkler, and P. Martin. Confidence intervals for performance measures of M/M/1 queue with constant retrial policy. *Asia-Pacific Journal of Operational Research*, 32(06):1550046, 2015.

[93] W. Eschenbach. Statistical inference for queueing models. *Series Statistics*, 15(3):451–462, 1984.

[94] I. N. Fabris-Rotelli, C. Kraamwinkel, et al. An investigation and historical overview of the G/M and M/G queueing processes. In *South African Statistical Journal Proceedings: Proceedings of the 54th Annual Conference of the South African Statistical Association: Congress 1*, Sabinet Online: 18–25, 2012.

[95] P. Fearnhead. Filtering recursions for calculating likelihoods for queues based on inter-departure time data. *Statistics and Computing*, 14(3):261–266, 2004.

[96] K.W. Fendick and W. Whitt. Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proceedings of the IEEE*, 77(1):171–194, 1989.

[97] H. Feng, P. Dube, and L. Zhang. On estimation problems for the G/G/∞ queue. *ACM SIGMETRICS Performance Evaluation Review*, 39(3):40–42, 2011.

[98] J.C. Frey and E.H. Kaplan. Queue inference from periodic reporting data. *Operations Research Letters*, 38(5):420–426, 2010.

[99] M. C. Fu and J.-Q. Hu. *Conditional Monte Carlo: Gradient estimation and optimization applications*, volume 392. Springer Science & Business Media, 2012.

[100] A. Ganesh, P. Green, N. O'Connell, and S. Pitts. Bayesian network management. *Queueing Systems*, 28(1):267–282, 1998.

[101] N. Gans, N. Liu, A. Mandelbaum, H. Shen, H. Ye, et al. Service times in call centers: Agent heterogeneity and learning with some operational consequences. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, Institute of Mathematical Statistics: 99–123, 2010.

[102] D.P. Gaver and P.A. Jacobs. On inference concerning time-dependent queue performance: the M/G/1 example. *Queueing Systems*, 6(1):261–275, 1990.

[103] R. Gawlick. Estimating disperse network queues: The queue inference engine. *ACM SIGCOMM Computer Communication Review*, 20(5):111–118, 1990.

[104] M. Ghorbani-Mandolakani and M. R. Salehi Rad. ML and Bayes estimation in a two-phase tandem queue with a second optional service and random feedback. *Communications in Statistics-Theory and Methods*, 45(9):2576–2591, 2016.

[105] P.W. Glynn, B. Melamed, and W. Whitt. Estimating customer and time averages. *Operations research*: 400–408, 1993.

[106] P.W. Glynn and M. Torres. Parametric estimation of tail probabilities for the single-server queue. *Chapter in: Frontiers in queueing*: 449–462, 1997.

[107] P.W. Glynn and W. Whitt. Indirect estimation via $L = \lambda W$. *Operations Research*: 82–103, 1989.

[108] P.W. Glynn and A.J. Zeevi. Estimating tail probabilities in queues via extremal statistics. analysis of communication networks: Call centres, traffic and performance, 135–158. *Fields Inst. Commun*, 28.

[109] A Goldenshluger. Nonparametric estimation of service time distribution in the M/G/∞ queue and related estimation problems. *arXiv preprint arXiv:1508.00076*, 2015.

[110] A. Gorst-Rasmussen and M.B. Hansen. Asymptotic inference for waiting times and patiences in queues with abandonment. *Communications in Statistics–Simulation and Computation*, 38(2):318–334, 2009.

[111] T.L. Goyal and C.M. Harris. Maximum-likelihood estimates for queues with state-dependent service. *Sankhyā: The Indian Journal of Statistics, Series A*, 34(1):65–80, 1972.

[112] WK Grassmann. Technical note – the optimal estimation of the expected number in a M/D/∞ queueing system. *Operations Research*, 29(6):1208–1211, 1981.

[113] I. Greenberg. *Parameter estimation in a simple queue*. PhD thesis, Doctoral Thesis at New York University, 1964.

[114] I. Greenberg. The behavior of a simple queue at various times and epochs. *SIAM Review*, 9(2):234–248, 1967.

[115] D. Gross and C. Harris. Fundamentals of queueing theory. 1998.

[116] R. Grübel and H. Wegener. Matchmaking and testing for exponentiality in the M/G/∞ queue. *Journal of Applied Probability*, 48(1):131–144, 2011.

[117] I. Gurvich, J. Huang, and A. Mandelbaum. Excursion-based universal approximations for the Erlang-A queue in steady-state. *Mathematics of Operations Research*, 39(2):325–373, 2013.

[118] S. Halfin. Linear estimators for a class of stationary queueing processes. *Operations Research*, 30(3):515–529, 1982.

[119] P. Hall and J. Park. Nonparametric inference about service time distribution from indirect measurements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4):861–875, 2004.

[120] S.A. Hall and R.C. Larson. Using partial queue-length information to improve the queue inference engine's performance. 1991.

[121] M.B. Hansen and S.M. Pitts. Nonparametric inference from the M/G/1 workload. *Bernoulli*, 12(4):737–759, 2006.

[122] S.L. Hantler and Z. Rosberg. Optimal estimation for an M/M/c queue with time varying parameters. *Stochastic Models*, 5(2):295–313, 1989.

[123] K. Harishchandra and S.S. Rao. A note on statistical inference about the traffic intensity parameter in $M/E_k/1$ queue. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 144–148, 1988.

[124] C.M. Harris. Some new results in the statistical analysis of queues. Technical report, DTIC Document, 1973.

[125] D. He, R. Li, Q. Huang, and P. Lei. Maximum entropy principle based estimation of performance distribution in queueing theory. 2014.

[126] S. Heckmueller and B.E. Wolfinger. Reconstructing arrival processes to discrete queueing systems by inverse load transformation. *Simulation*, 2010.

[127] S. Heckmuller and B.E. Wolfinger. Reconstructing arrival processes to G/D/1 queueing systems and tandem networks. In *Performance Evaluation of Computer & Telecommunication Systems, 2009. SPECTS 2009. International Symposium on*, volume 41, pages 361–368. IEEE, 2009.

[128] X. Hei, B. Bensaou, and D.H.K. Tsang. A light-weight available bandwidth inference methodology in a queueing analysis approach. In *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, volume 1: 120–124. IEEE, 2005.

[129] X. Hei, B. Bensaou, and D.H.K. Tsang. Model-based end-to-end available bandwidth inference using queueing analysis. *Computer Networks*, 50(12):1916–1937, 2006.

[130] O. Hernàndez-Lerma and S.I. Marcus. Adaptive control of service in queueing systems. *Systems & Control Letters*, 3(5):283–289, 1983.

[131] O. Hernàndez-Lerma and S.I. Marcus. Optimal adaptive control of priority assignment in queueing systems. *Systems & Control Letters*, 4(2):65–72, 1984.

[132] C.C. Heyde. Some results on inference for stationary processes and queueing systems. *OXFORD STATISTICAL SCIENCE SERIES in book "Queueing and Related Models, Bhat, Basawa"*: 337–337, 1992.

[133] Y.C. Ho and C.G. Cassandras. Perturbation analysis for control and optimization of queueing systems: an overview and the state of the art. *Chapter in: Frontiers in Queueing"(J. Dshalalow, Ed.), CRC Press*: 395–420, 1997.

[134] Y. Hu and Lee. Parameter estimation for a reflected fractional Brownian motion based on its local time. 2012.

[135] M.L. Huang and P. Brill. On estimation in M/G/c/c queues. *International Transactions in Operational Research*, 8(6):647–657, 2001.

[136] R. Ibrahim. *Real-Time Delay Prediction in Customer Service Systems*. PhD thesis, Doctoral Thesis at Columbia University, 2010.

[137] R. Ibrahim and W. Whitt. Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management*, 11(3):397–415, 2009.

[138] R. Ibrahim and W. Whitt. Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations Research-Baltimore*, 59(5):1106–1118, 2011.

[139] D.R. Insua, M. Wiper, and F. Ruggeri. Bayesian analysis of M/Er/1 and M/H$_k$/1 queues. *Queueing Systems*, 30(3):289–308, 1998.

[140] S. Jain. Comparison of confidence intervals of traffic intensity for M/E$_k$/1 queueing systems. *Statistical Papers*, 32(1):167–174, 1991.

[141] S. Jain. Relative efficiency of a parameter for a M/G/1 queueing system based on reduced and full likelihood functions. *Communications in Statistics-Simulation and Computation*, 21(2):597–606, 1992.

[142] S. Jain. Estimating changes in traffic intensity for M/M/1 queueing systems. *Microelectronics and Reliability*, 35(11):1395–1400, 1995.

[143] S. Jain. An autoregressive process and its application to queueing model. *Metron-International Journal of Statistics*, 58(1-2):131–138, 2000.

[144] S. Jain and T.S.S. Rao. Statistical inference for the GI/G/1 queue using diffusion approximation. *International Journal of Information and Management Sciences*, 11(2):1–12, 2000.

[145] S. Jain and J.G.C. Templeton. Statistical inference for G/M/1 queueing system. *Operations Research Letters*, 7(6):309–313, 1988.

[146] S. Jain and J.G.C. Templeton. Problem of statistical inference to control the traffic intensity. *Sequential Analysis*, 8(2):135–146, 1989.

[147] S. Jain and J.G.C. Templeton. Confidence interval for M/M/2 queue with heterogeneous servers. *Operations Research Letters*, 10(2):99–101, 1991.

[148] J. Jang and C.R. Liu. Waiting time estimation in a manufacturing system using the number of machine idle periods. *European Journal of Operational Research*, 78(3):426–440, 1994.

[149] J. Jang, J. Suh, and C.R. Liu. A new procedure to estimate waiting time in GI/G/2 system by server observation. *Computers & Operations Research*, 28(6):597–611, 2001.

[150] J.H. Jenkins. The relative efficiency of direct and maximum likelihood estimates of mean waiting time in the simple queue, M/M/1. *Journal of Applied Probability*: 396–403, 1972.

[151] L.K. Jones. Inferring balking behavior from transactional data. *Operations Research*: 778–784, 1999.

[152] L.K. Jones. Remarks on queue inference from departure data alone and the importance of the queue inference engine. *Operations Research Letters*, 2012.

[153] L.K. Jones and R.C. Larson. Efficient computation of probabilities of events described by order statistics and applications to queue inference. 1994.

[154] B. Kauffmann. Inverse problems on bandwidth sharing networks.

[155] B. Kauffmann. *Inverse problems in Networks*. PhD thesis, Doctoral Thesis at Université Pierre et Marie Curie-Paris VI, 2011.

[156] J.C. Ke and Y.K. Chu. Nonparametric and simulated analysis of intensity for a queueing system. *Applied Mathematics and Computation*, 183(2):1280–1291, 2006.

[157] J.C. Ke and Y.K. Chu. Comparison on five estimation approaches of intensity for a queueing system with short run. *Computational Statistics*, 24(4):567–582, 2009.

[158] J.C. Ke, M.Y. Ko, and K.C. Chiou. Analysis of 2^4 factorial design for a controllable M/G/1 system. *Mathematical and Computational Applications*, 13(3):165–174, 2008.

[159] J.C. Ke, M.Y. Ko, and S.H. Sheu. Estimation comparison on busy period for a controllable M/G/1 system with bicriterion policy. *Simulation Modelling Practice and Theory*, 16(6):645–655, 2008.

[160] N. Keiding. Maximum likelihood estimation in the birth-and-death process. *The Annals of Statistics*, 3(2):363–372, 1975.

[161] P.C. Kiessler and R. Lund. Technical note: Traffic intensity estimation. *Naval Research Logistics (NRL)*, 56(4):385–387, 2009.

[162] S. Kim and W. Whitt. Statistical analysis with Little's law. *preparation. Columbia University*, 2012.

[163] S. Kim and W. Whitt. Statistical analysis with Little's law, supplementary material: More on the call center data. *Supplementary Material to Preprint*, 2012.

[164] S.H. Kim and W. Whitt. Appendix to estimating waiting times with the time-varying Little's law. 2012.

[165] S.H. Kim and W. Whitt. Estimating waiting times with the time-varying Little's law. 2012.

[166] Y.B. Kim and J. Park. New approaches for inference of unobservable queues. In *Proceedings of the 40th Conference on Winter Simulation*: 2820–2825. Winter Simulation Conference, 2008.

[167] I. N Kovalenko. In recovering the characteristics of a system from observations of the outgoing flow (in Russian). *Dokl. Akad. Nauk SSSR*, 164(5):979–981, 1965.

[168] S. Kraft, S. Pacheco-Sanchez, G. Casale, and S. Dawson. Estimating service resource consumption from response time measurements. In *Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools*, page 48. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2009.

[169] A. Kumar. On the average idle time and average queue length estimates in an M/M/1 queue. *Operations Research Letters*, 12(3):153–157, 1992.

[170] R.C. Larson. The queue inference engine: deducing queue statistics from transactional data. *Management Science*: 586–601, 1990.

[171] R.C. Larson. The queue inference engine: addendum. *Management science*, 37(8): 1062, 1991.

[172] L. Li, X. Chen, Z. Li, and L. Zhang. Freeway travel-time estimation based on temporal–spatial queueing model. *Intelligent Transportation Systems, IEEE Transactions on*, 14(3):1536–1541, 2013.

[173] H.W. Lilliefors. Some confidence intervals for queues. *Operations Research*: 723–727, 1966.

[174] H. Liu, X. Wu, and P. Michalopoulos. Improving queue size estimation for Minnesota's stratified zone metering strategy. *Transportation Research Record: Journal of the Transportation Research Board*, 2015.

[175] H. X. Liu, X. Wu, W. Ma, and H. Hu. Real-time queue length estimation for congested signalized intersections. *Transportation research part C: emerging technologies*, 17(4):412–427, 2009.

[176] X. Liu, J. Heo, L. Sha, and X. Zhu. Adaptive control of multi-tiered web applications using queueing predictor. In *Network Operations and Management Symposium, 2006. NOMS 2006. 10th IEEE/IFIP*, pages 106–114. IEEE, 2006.

[177] Z. Liu, L. Wynter, C.H. Xia, and F. Zhang. Parameter inference of queueing models for it systems using end-to-end measurements. *Performance Evaluation*, 63(1):36–60, 2006.

[178] F. Machihara. Carried traffic estimate errors for delay systems. *Electronics and Communications in Japan (Part I: Communications)*, 67(12):49–58, 1984.

[179] A. Mandelbaum and S. Zeltyn. Estimating characteristics of queueing networks using transactional data. *Queueing Systems*, 29(1):75–127, 1998.

[180] M. Mandjes and R. van de Meent. Inferring traffic burstiness by sampling the buffer occupancy. *NETWORKING 2005. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems*: 233–240, 2005.

[181] M. Mandjes and R. van De Meent. Resource dimensioning through buffer sampling. *IEEE/ACM Transactions on Networking (TON)*, 17(5):1631–1644, 2009.

[182] M. Mandjes and P. Żuraniewski. A queueing-based approach to overload detection. *Network Control and Optimization*, pages 91–106, 2009.

[183] M. Mandjes and P. Zuraniewski. M/g/[infinity] transience, and its applications to overload detection. *Performance Evaluation*, 2011.

[184] D. Manjunath and M.L. Molle. Passive estimation algorithms for queueing delays in lans and other polling systems. In *INFOCOM'96. Fifteenth Annual Joint Conference of the IEEE Computer Societies. Networking the Next Generation. Proceedings IEEE*, volume 1: 240–247. IEEE, 1996.

[185] M. Manoharan and J.K. Jose. Markovian queueing system with random balking. *OPSEARCH*, 38(3):1–11, 2011.

[186] Y. Masuda. Exploiting partial information in queueing systems. *Operations Research*, 43(3):530–536, 1995.

[187] B.P.M. McCabe, G.M. Martin, and D. Harris. Efficient probabilistic forecasts for counts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2011.

[188] M.F. Mcgrath, D. Gross, and N.D. Singpurwalla. A subjective Bayesian approach to the theory of queues I–modeling. *Queueing Systems*, 1(4):317–333, 1987.

[189] M.F. McGrath and N.D. Singpurwalla. A subjective Bayesian approach to the theory of queues II–inference and information in M/M/1 queues. *Queueing Systems*, 1(4):335–353, 1987.

[190] R. McVinish and P.K. Pollett. Constructing estimating equations from queue length data. *preprint*.

[191] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[192] A. H. Mohajerzadeh, M. H. Yaghmaee, and A. Zahmatkesh. Efficient data collecting and target parameter estimation in wireless sensor networks. *Journal of Network and Computer Applications*, 57:142–155, 2015.

[193] A. Mohammadi and M.R. Salehi-Rad. Bayesian inference and prediction in an M/G/1 with optional second service. *Communications in Statistics-Simulation and Computation*, 41(3):419–435, 2012.

[194] J. Morales, M. Eugenia Castellanos, A.M. Mayoral, R. Fried, and C. Armero. Bayesian design in queues: An application to aeronautic maintenance. *Journal of statistical planning and inference*, 137(10):3058–3067, 2007.

[195] E. Morozov, R. Nekrasova, I. Peshkova, and A. Rumyantsev. A regeneration-based estimation of high performance multiserver systems. In *International Conference on Computer Networks*: 271–282. Springer, 2016.

[196] M.V. Muddapur. Bayesian estimates of parameters in some queueing models. *Annals of the Institute of Statistical Mathematics*, 24(1):327–331, 1972.

[197] C. Muthu and V.S. Sampathkumar. Estimation of parameters in a particular finite capacity priority queueing model. *Optimization*, 34(4):359–363, 1995.

[198] S.Y. Nam, S. Kim, and D.K. Sung. Estimation of available bandwidth for an M/G/1 queueing system. *Applied Mathematical Modelling*, 33(8):3299–3308, 2009.

[199] SR Neal and A. Kuczura. A theory of traffic measurement errors for loss systems with renewal input. *BSTJ*, 52:967–990, 1973.

[200] N. Nelgabats, Y. Nov, and G. Weiss. Sojourn Time Estimation in an M/G/∞ Queue with Partial Information. *preprint*.

[201] B. L. Nelson. *Stochastic modeling: analysis and simulation*. Courier Corporation, 2012.

[202] M. Neuts. Reflections on statistical methods or complex stochastic systems. *Modeling Uncertainty*: 751–760, 2005.

[203] A. Novak and R. Watson. Determining an adequate probe separation for estimating the arrival rate in an M/D/1 queue using single-packet probing. *Queueing Systems*, 61(4):255–272, 2009.

[204] A. Nozari and W. Whitt. Estimating average production intervals using inventory measurements: Little's law for partially observable processes. *Operations research*, 36(2):308–323, 1988.

[205] A.G. Pakes. The serial correlation coefficients of waiting times in the stationary GI/M/1 queue. *The Annals of Mathematical Statistics*, 42(5):1727–1734, 1971.

[206] J. Park. On the choice of an auxiliary function in the M/G/∞ estimation. *Computational Statistics & Data Analysis*, 51(12):5477–5482, 2007.

[207] J. Park, Y. B. Kim, and T. R. Willemain. Analysis of an unobservable queue using arrival and departure times. *Computers & Industrial Engineering*, 61(3):842–847, 2011.

[208] I.C. Paschalidis and S. Vassilaras. On the estimation of buffer overflow probabilities from measurements. *Information Theory, IEEE Transactions on*, 47(1):178–191, 2001.

[209] J. Pickands and R.A. Stine. Estimation for an M/G/∞ queue with incomplete information. *Biometrika*, 84(2):295–308, 1997.

[210] F. Pin, D. Veitch, and B. Kauffmann. Statistical estimation of delays in a multicast tree using accelerated em. *Queueing Systems*, 66(4):1–44, 2010.

[211] S.M. Pitts. Nonparametric estimation of the stationary waiting time distribution function for the GI/G/1 queue. *The Annals of Statistics*, pages 1428–1446, 1994.

[212] N. U. Prabhu. *Stochastic storage processes: queues, insurance risk, dams, and data communication*, volume 15. Springer Science & Business Media, 2012.

[213] J.E. Prieger. Estimation of a simple queuing system with units-in-service and complete data. *Working Papers*, 2005.

[214] R.C. Quinino and F.R.B. Cruz. Bayesian sample sizes in an M/M/1 queueing systems. *The International Journal of Advanced Manufacturing Technology*: 1–8, 2016.

[215] M.F. Ramalhoto. Some statistical problems in random translations of stochastic point processes. *Annals of Operations Research*, 8(1):229–242, 1987.

[216] P. Ramirez, R.E. Lillo, and M.P. Wiper. Bayesian analysis of a queueing system with a long-tailed arrival process. *Communications in Statistics–Simulation and Computation*, 37(4):697–712, 2008.

[217] P. Ramirez-Cobo, R.E. Lillo, S. Wilson, and M.P. Wiper. Bayesian inference for double Pareto lognormal queues. *The Annals of Applied Statistics*, 4(3):1533–1557, 2010.

[218] H. Ren and G. Wang. Bayes estimation of traffic intensity in M/M/1 queue under a precautionary loss function. *Procedia Engineering*, 29:3646–3650, 2012.

[219] H.P. Ren and J.P. Li. Bayes estimation of traffic intensity in M/M/1 queue under a new weighted square error loss function. *Advanced Materials Research*, 485:490–493, 2012.

[220] J.F. Reynolds. On estimating the parameters of a birth-death process. *Australian Journal of Statistics*, 15(1):35–43, 1973.

[221] A. Rodrigo. Estimators of the retrial rate in M/G/1 retrial queues. *ASIA PACIFIC JOURNAL OF OPERATIONAL RESEARCH*, 23(2):193, 2006.

[222] A. Rodrigo and M. Vazquez. Large sample inference in retrial queues. *Mathematical and Computer Modelling*, 30(3-4):197–206, 1999.

[223] J. Rodrigues and J.G. Leite. A note on bayesian analysis in M/M/1 queues derived from confidence intervals. *Statistics: A Journal of Theoretical and Applied Statistics*, 31(1):35–42, 1998.

[224] A.M. Ross and J.G. Shanthikumar. Estimating effective capacity in Erlang loss systems under competition. *Queueing Systems*, 49(1):23–47, 2005.

[225] J.V. Ross, T. Taimre, and P.K. Pollett. Estimation for queues from queue length data. *Queueing Systems*, 55(2):131–138, 2007.

[226] S.M. Ross. Identifiability in GI/G/k queues. *Journal of Applied Probability*, 7(3):776–780, 1970.

[227] G. Rubin and D.S. Robson. A single server queue with random arrivals and balking: confidence interval estimation. *Queueing Systems*, 7(3):283–306, 1990.

[228] L. Schruben and R. Kulkarni. Some consequences of estimating parameters for the M/M/1 queue. *Operations Research Letters*, 1(2):75–78, 1982.

[229] A. Senderovich, S. Leemans, S. Harel, A. Gal, A. Mandelbaum, and W. V.D. Aalst. Discovering queues from event logs with varying levels of information. *Lect Notes Bus Inf (forthcoming)*, 2015.

[230] A. Senderovich, M. Weidlich, A. Gal, and A. Mandelbaum. Queue mining for delay prediction in multi-class service processes. *Information Systems*, 2015.

[231] B. Sengupta. Markov processes whose steady state distribution is matrix-exponential with an application to the GI/PH/1 queue. *Advances in Applied Probability*, pages 159–180, 1989.

[232] V. Sharma. Estimating traffic intensities at different nodes in networks via a probing stream. In *Global Telecommunications Conference, 1999. GLOBECOM'99*, volume 1, pages 374–380. IEEE, 1999.

[233] V. Sharma and R. Mazumdar. Estimating traffic parameters in queueing systems with local information. *Performance Evaluation*, 32(3):217–230, 1998.

[234] N.D. Singpurwalla. Discussion of Thiruvaiyaru and Basawa's "Empirical Bayes estimation for queueing systems and networks". *Queueing Systems*, 11(3):203–206, 1992.

[235] W. L. Smith. On the distribution of queueing times. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 49, pages 449–461. Cambridge Univ Press, 1953.

[236] S.Y. Sohn. Empirical bayesian analysis for traffic intensity: M/M/1 queues with covariates. *Queueing Systems*, 22(3):383–401, 1996.

[237] S.Y. Sohn. Influence of a prior distribution on traffic intensity estimation with covariates. *Journal of Statistical Computation and Simulation*, 55(3):169–180, 1996.

[238] S.Y. Sohn. Robust design of server capability in M/M/1 queues with both partly random arrival and service rates. *Computers & Operations Research*, 29(5):433–440, 2002.

[239] M.E. Sousa-Vieira. Suitability of the M/G/∞ process for modeling scalable h. 264 video traffic. *Analytical and Stochastic Modeling Techniques and Applications*, pages 149–158, 2011.

[240] V Srinivas and B.K. Kale. Ml and umvu estimation in the M/D/1 queuing system. *Communications in Statistics-Theory and Methods*, (just-accepted), 2015.

[241] V. Srinivas, S.S. Rao, and B.K. Kale. Estimation of measures in M/M/1 queue. *Communications in Statistics-Theory and Methods*, 40(18):3327–3336, 2011.

[242] S. Subba Rao and K. Harishchandra. On a large sample test for the traffic intensity in GI/G/s queue. *Naval Research Logistics Quarterly*, 33(3):545–550, 1986.

[243] H. Y. Sutarto and E. Joelianto. Modeling, identification, estimation, and simulation of urban traffic flow in Jakarta and Bandung. *Journal of Mechatronics, Electrical Power, and Vehicular Technology*, 6(1):57–66, 2015.

[244] C. Sutton and M.I. Jordan. Probabilistic inference in queueing networks. In *Proceedings of the Third conference on Tackling computer systems problems with machine learning techniques*: 6–6. USENIX Association, 2008.

[245] C. Sutton and M.I. Jordan. Learning and inference in queueing networks. 2010.

[246] C. Sutton and M.I. Jordan. Bayesian inference for queueing networks and modeling of internet services. *The Annals of Applied Statistics*, 5(1):254–282, 2011.

[247] T.R. Thiagarajan and C.M. Harris. Statistical tests for exponential service from M/G/1 waiting-time data. *Naval Research Logistics Quarterly*, 26(3):511–520, 1979.

[248] D. Thiruvaiyaru and I.V. Basawa. Empirical Bayes estimation for queueing systems and networks. *Queueing Systems*, 11(3):179–202, 1992.

[249] D. Thiruvaiyaru, I.V. Basawa, and U.N. Bhat. Estimation for a class of simple queueing networks. *Queueing Systems*, 9(3):301–312, 1991.

[250] H. Toyoizumi. Sengupta's invariant relationship and its application to waiting time inference. *Journal of Applied Probability*, 34(3):795–799, 1997.

[251] S. Vorobeychikov. Cusum algorithms for parameter estimation in queueing systems with jump intensity of the arrival process. In *Information Technologies and Mathematical Modelling-Queueing Theory and Applications: 14th International Scientific Conference, ITMM 2015, named after AF Terpugov, Anzhero-Sudzhensk, Russia, November 18-22, 2015, Proceedings*, volume 564, page 275. Springer, 2015.

[252] J. Walrand. Filtering formulas and the ./M/1 queue in a quasi-reversible network. *Stochastics: An International Journal of Probability and Stochastic Processes*, 6(1):1–22, 1981.

[253] J. Walrand. *An introduction to queueing networks*, volume 21. Prentice Hall Englewood Cliffs, NJ, 1988.

[254] T.Y. Wang, J.C. Ke, K.H. Wang, and S.C. Ho. Maximum likelihood estimates and confidence intervals of an M/M/R queue with heterogeneous servers. *Mathematical Methods of Operations Research*, 63(2):371–384, 2006.

[255] W. Wang and G. Casale. Maximum likelihood estimation of closed queueing network demands from queue length data. *ACM SIGMETRICS Performance Evaluation Review*, 43(2):45–47, 2015.

[256] K.H. Wang, S.C. Chen and J.C. Ke. Maximum likelihood estimates and confidence intervals of an M/M/R/N queue with balking and heterogeneous servers. *RAIRO – Operations Research*, 38(3):227–241, 2004.

[257] W. Wang, J. F. Pérez, and G. Casale. Filling the gap: a tool to automate parameter estimation for software performance models. In *Proceedings of the 1st International Workshop on Quality-Aware DevOps*, pages 31–32. ACM, 2015.

[258] R. Warfield and G. Foers. Application of Bayesian teletraffic measurement to systems with queueing or repeated attempts. In *Proceedings of the Eleventh International Teletraffic Congress*, 1985.

[259] RE Warfield and GA Foers. Application of Bayesian methods to teletraffic measurement and dimensioning. *Australian Telecommunications Research*, 18:51–58, 1984.

[260] A. Weerasinghe and A. Mandelbaum. Abandonment versus blocking in many-server queues: asymptotic optimality in the QED regime. *Queueing Systems*, 75(2-4):279–337, 2013.

[261] W. Whitt. Fitting birth-and-death queueing models to data. *Statistics and Probability Letters*, 82:998–1004, 2012.

[262] W. Whitt. Many-server limits for periodic infinite-server queues. *Columbia University*, 2015.

[263] M.P. Wiper. Bayesian analysis of Er/M/1 and Er/M/c queues. *Journal of Statistical Planning and Inference*, 69(1):65–79, 1998.

[264] R.W. Wolff. Problems of statistical inference for birth and death queuing models. *Operations Research*: 343–357, 1965.

[265] C.M. Woodside, D.A. Stanford, and B. Pagurek. Optimal prediction of queue lengths and delays in GI/M/M multiserver queues. *Operations research*: 809–817, 1984.

[266] X. Xu, Q. Zhang, and X. Ding. Hypothesis testing and confidence regions for the mean sojourn time of an M/M/1 queueing system. *Communications in Statistics–Theory and Methods*, 40(1):28–39, 2010.

[267] G. B. Yom-Tov and A. Mandelbaum. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299, 2014.

[268] S.-Z. Yu. *Hidden Semi-Markov Models: Theory, Algorithms and Applications*. Morgan Kaufmann, 2015.

[269] L. C. Zammit, S. G. Fabri, and K. Scerri. Joint state and parameter estimation for a macro traffic junction model. In *Control and Automation (MED), 2016 24th Mediterranean Conference on*: 1152–1157. IEEE, 2016.

[270] L. Zhang, C.H. Xia, M.S. Squillante, and W.N. Mills III. Workload service requirements analysis: A queueing network optimization approach. In *Modeling, Analysis and Simulation of Computer and Telecommunications Systems, 2002. MASCOTS 2002. Proceedings. 10th IEEE International Symposium on*: 23–32. IEEE, 2002.

[271] Q. Zhang and X. Xu. Confidence intervals of performance measures for an m/g/1 queueing system. *Communications in Statistics–Simulation and Computation®*, 39(3):501–516, 2010.

[272] Q. Zhang, X. Xu, and S. Mi. A generalized p-value approach to inference on the performance measures of an $M/E_k/1$ queueing system. *Communications in Statistics-Theory and Methods*, 45(8):2256–2267, 2016.

[273] P. Zuraniewski, M. Mandjes, and M. Mellia. Empirical assessment of VoIP overload detection tests. In *Next Generation Internet (NGI), 2010 6th EURO-NF Conference on*: 1–8. IEEE, 2010.