ESTIMATING CHARACTERISTICS OF QUEUEING NETWORKS USING TRANSACTIONAL DATA

RESEARCH THESIS

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE IN STATISTICS

SERGEY ZELTYN

SUBMITTED TO THE SENATE OF THE TECHNION – ISRAEL INSTITUTE OF TECHNOLOGY

Tebeth, 5756 Haifa January, 1996

THE WORK DESCRIBED HEREIN WAS SUPERVISED BY
PROF. AVISHAI MANDELBAUM UNDER THE AUSPICES OF
THE FACULTY OF INDUSTRIAL ENGINEERING AND
MANAGEMENT

THE GENEROUS FINANCIAL HELP OF
THE TECHNION – ISRAEL INSTITUTE OF TECHNOLOGY AND
MIRIAM AND AARON GUTWIRTH
IS GRATEFULLY ACNOWLEDGED

I AM DEEPLY THANKFUL TO PROFESSOR AVISHAI
MANDELBAUM FOR ENERGY, INSPIRATION AND TIME
THAT HE DEVOTED TO THIS WORK,
FOR HIS INCESSANT SUPPORT AND ENCOURAGEMENT

I AM ENDLESSLY INDEBTED TO MY WIFE OLGA FOR HER
LOVE AND FAITH

Contents

Abstract.

We consider an open queueing network. External arrivals to the network are Poisson (possibly time-inhomogeneous), service times are general and switches of customers between stations are Markovian (governed by a matrix of routing probabilities). Customers' transitions between stations may be either immediate or of exponentially distributed durations. First Come First Served (FCFS) discipline is assumed at all stations.

Each customer is supplied with an Identification Number (ID) upon entering the network. This ID accompanies him until exit, while being registered at each service on his route.

Suppose that we are able to register also service starts and service terminations, in addition to the corresponding ID's. Such type of data will be referred to as *transactional data.* (In some special cases, we also assume that external arrival times are also recorded.) The objective is to estimate the evolution of the queue during some time interval; specifically, to calculate

$$\hat{Q}(t) = \mathrm{E}[Q(t)/\mathcal{F}], \quad t \in [T_1, T_2],$$

where $\mathcal{F}$ represents the available information from the transactional data and Q is the queue length.

Previous research on the subject was devoted exclusively to the single-station case. (ID's are insignificant then.) The main approaches were the method of order statistics, used in the pioneering paper of Larson [12], and the method of Hidden Markov Models, developed by Daley and Servi [4].

First we consider the problem of Busy-Period Interpolation. In other words, we assume that all service starts and terminations during a busy period at a specific station are registered, and that estimation is carried out at the end of the busy period.

In the simplest case of immediate transition times between the stations and known external arrivals, we can calculate the queue length exactly. In all other cases, the technique of Hidden Markov Models is used. The case of immediate transitions and unknown external arrivals is similar to the single-station case. If transition times are exponential, one needs a more refined definition of the Hidden Markov Model. It turns out that order relations between truncated exponential random variables are involved in calculations of the transition probabilities of the Markov model.

The second class of the estimation problems includes Real-Time Estimation and General Interpolation. Here we estimate the queue during a busy period, using available information, until estimation time t within the busy period. In Real-Time Estimation, one infers the queue at time t, possibly updating the estimate as t varies over the busy period. (Bertsimas and Servi [2] performed such an analysis for a single station.) In the

case of General Interpolation, we infer the whole evolution of the queue from the start of the busy period until t, dynamically for $t \geq 0$.

These problems are, in general, more sophisticated than Busy-Period Interpolation. The case of immediate transitions is solved using specific methods; Hidden Markov Models are applicable if transitions are exponential. However, the state space of the Markov chains "explodes" and even networks that are moderate in size are, in general, not amenable for computations.

Our theoretical results are applied to a data set originating in a bank operation. Transition times between stations may be assumed immediate and external arrivals are registered. Therefore, with the availability of ID's, one can actually calculate the exact queues. As an experiment, we assume that external arrivals are unknown and use our Hidden Markov Model algorithm. Satisfactory correspondence between estimates and reality prevails in most cases, despite some theoretical assumptions (FCFS, for example) that do not exactly apply at the bank. Future additional applications may include communication and transportation networks.

Chapter 1 includes an introduction to the subject. We start with an informal description of the problem in Section 1. Section 2 includes a literature survey, an introduction to the methods of order statistics and Hidden Markov Models and examples of queue inference for the single-station case. In Section 3 we give an exact problem formulation, specifically description of the queueing network and objectives. Subsection 3.2 is devoted to a discussion of the structure of observations and Subsection 3.4 includes a short summary of our results.

Chapter 2 contains theoretical results. The Busy-Period Interpolation problems are considered in Section 1. Subsection 1.5 provides auxiliary results concerning truncated exponential random variables. Further, in Section 2, we study alternative representations of the observed information. These results are used in Section 3, devoted to Real-Time Estimation.

Finally, part of our theory is applied in **Chapter 3** to real transactional data of a bank branch. We conclude with some possible directions for future research.

Abbreviations.

ATM	Automatic Teller Machine
FCFS	First Come First Served
ID	Identification Number
IR-time	Invisible-Routing time
HMM	Hidden Markov Model
QIE	Queueing Inference Engine
VR-time	Visible-Routing time

Chapter I

Introduction

1 Motivation

We start with an informal introduction to the subject.

Suppose that one would like to measure the queue of an Automatic Teller Machine (ATM,"Bankomat" or "Caspomat" in Hebrew). In principle, it is possible to produce direct and exact measurements: one can install a camera or hire someone to stand near the ATM and register customers. However, both methods are likely to be expensive.

A different approach to the problem could be to record customers' *transactions*, i.e. service starts and terminations. (These times coincide with insertion and ejection times of the customers' ATM cards). Such information can be recorded easily by the ATM. However, since transactions do not include arrival times of customers, we do not know the queue exactly, hence we must estimate it.

There are many queueing systems which are, in some way, similar to the ATM example. Specifically, they share the following two features:

- *invisible queues*, namely queues for which direct measurements are either impossible or too expensive;
- *transactional data* is relatively easy to access.

Larson [12] was the first to introduce and solve the problem of queue-length inference, using transactional data. His paper contains convincing examples, included in the following list:

- Telephone service centers have a limited number of channels. Transactions correspond to starts and terminations of telephone calls. If customers that encounter a busy signal (blocked) call back with probability p, then the customers that do call back constitute a *retrial queue*. Usually, information about the arrival times of blocked customers is not available within the service center, but rather, at the level

of the telephone-call carriers ("Bezek" in Israel). Hence, such information may be impossible to obtain, and in any case expensive to monitor.

- Cellular phone networks ("Pelephone" in Hebrew) is another example of retrial queues. Here it is impossible to record arrival times of customers that encounter a busy signal.

- Transportation queues at traffic intersections. Transactions may correspond to times when vehicles cross a measuring cable. Such information is much more amenable for mathematical analysis then information obtained by installed cameras.

- Face-to-face services such as banks, government offices, health clinics, and so on. It is typically easier here to measure transactions rather then arrival times of customers. (These are the type of applications that motivated the present work, and to which our theory is being applied in Chapter 3.)

Note that the first two applications deal with single-station queues. The last two can be considered also within a queueing network context. As a matter of fact, previous research covered the single-station case exclusively, and the contribution here is an extension to a network setting.

2 Single-station case

2.1 Survey of Literature

As already mentioned, existing research has been devoted to single-station problems. The subject is only several years old, hence all the existing papers, which we are aware of, will now be surveyed.

Larson Larson opened up the subject in his pioneering paper [12], where he introduced the terminology Queueing Inference Engine (QIE) to describe the algorithm that infers queue length from transactional data.

Larson considered a service station with s servers working in parallel, Poisson arrivals and general distribution of service times ($M/G/s$ queue). Service times are truly general, as they are required to be neither independent nor identically distributed. (However, an assumption of the independence between services and arrivals had to be imposed.)

Transaction times were supposed known and the queue length was estimated through its expectation, conditioned on the available information.

Larson's simple key observation was that busy periods (time periods when all servers are working) coincide with the time intervals, over which new services begin instanta-

neously after service termination times (in practice — almost immediately). Note that a queue can arise only during a busy period.

Larson considered, what we call later, the problem of Busy-Period Interpolation. Specifically, his objective was to estimate the evolution of the queue length over a specified busy period, given the data (transactions) accumulated over that period. The estimation is carried out at the end of the period.

A technique, based on the order statistics property of the Poisson process, was used to develop an algorithm for queue inference. (We shall survey this technique in detail in Subsection 2.2.) It is notable that the algorithm does not require knowledge of the arrival rate λ.

The number of calculations for the initial algorithm was $O(n^5)$, where n is the number of transactions during the busy period. In [13] the algorithm was modified slightly and its number of calculations improved to $O(n^3)$.

Hall [8] attempts to incorporate information, in addition to transactions. Specifically, one is informed each time that the queue size hits 0 or M (some buffer space limit). Hall also calculated additional statistics, such as the density function of the arrival time for the k-th customer, as well as waiting time in queue. Several heuristic algorithms, that require less computations (down to $O(n^2)$ or $O(n)$) were developed.

Jones and Larson [11] derive results concerning order statistics and apply them to compute conditional distributions of some queue characteristics, for example, maximum queue length.

Daley & Servi An alternative approach was developed by Daley and Servi. In [3] they used the technique of taboo probabilities for Markov chains to solve Larson's problem and other problems of this type. For example, algorithms for $E_k/G/1$ (interarrival times are independent with Erlang-k distribution) and $M/G/s/m$ (finite buffer) systems were analyzed. (In every case, except $M/G/s$, the arrival rate λ must be known.)

An approximate $O(n^2 \ln n)$ algorithm was also developed. The approximation is based on a truncation of the conditional queue-length distribution.

In [4] more general methods were introduced. Specifically, Daley and Servi found that many models can be accommodated within the general framework of Markov Chain Boundary Value problems, which can be solved using taboo probabilities. A single algorithm was developed and applied to the following examples (in addition to [3]).

- Reneging (a customer leaves if the wait is too long);
- Bernoulli feedback (any customer completing service feeds back into the queue with probability p);

- Balking (any arriving customer leaves with probability p whenever the queue size is at or above a threshold M).

Another interesting result of [4] concerns a maximum-likelihood sample-path. An algorithm for its derivation was introduced and an illuminating example was given (see our Subsection 2.4 for an elaboration).

Markov Chain Boundary Value problems belong to an extensive class of models which are called Hidden Markov Models (HMM). The problem addressed in these models is the inference of different characteristics of a Markov chain, given partial information about its state. A clean account on HMM methods, with applications to speech recognition problems, is Rabiner [14].

In [5] asymptotic conditional mean queue length was studied for busy periods with a large number of customers. The $M/D/1$ queue (deterministic service times) was considered as a start, and then the results were generalized to $M/GI/1$ (i.i.d. service times). Brownian excursions were used for the approximations.

Bertsimas & Servi Several important results were introduced in [2]. First, an algorithm for queue estimation in real-time (within a busy period) was derived. This type of problems is very important since "if there is a possibility of real-time control of the service time, knowledge that the queue length *was* excessively large, but that it is currently zero, is not of value"[2].

Other results included:

- An original version of the $M/G/s$ problem (using multidimensional integrals).
- A solution of the $M_t/G/s$ problem (after a simple time-change, the initial $M/G/s$ algorithm is valid; the arrival rate $\lambda(t)$ must be known);
- A solution for the $GI/G/s$ problem (formulas include multidimensional integrals and convolutions that are, in general, non-computable).

Applications An attempt to apply the QIE algorithm to real data, coming from telecommunications, was made in Gawlick [7]. He compared true queue lengths with their estimates. Satisfactory results were reported, in spite of the fact that arrivals were not exactly Poisson.

2.2 Order statistics

Larson [12] considered an isolated busy period at an $M/G/s$ service station.

The available data consists of realizations of registration times (service starts and terminations) through the whole given busy period. They are denoted by $t_0, t_1, \ldots, t_n, t_{n+1}$.

Here t_0 is the first service start, t_n is the last service start and t_{n+1} is the last service termination (no service starts at t_{n+1}).

Let $A_0 = t_0, A_1, \ldots, A_n$ stand for the unknown arrival times of the customers that started service at t_0, $t_1, \ldots, t_n$, respectively. The cumulative number of arrivals during $(t_0, t]$ will be denoted by $A(t)$, $t_0 \leq t \leq t_{n+1}$ $(A(t_0) = 0,\ A(t_n) = A(t_{n+1}) = n)$.

The problem is to estimate the queue length $Q(t)$ over the given busy period, using the available data. It is equivalent to the estimation of $A(t)$, the number of arrivals up to time t, $t \in [t_0, t_n]$.

Our observations provide us with the following information about external arrivals:

- Arrival times precede service starts: $A_i \leq t_i,\ 1 \leq i \leq n$.
- Busy period terminates at t_{n+1}: $A_{n+1} > t_{n+1}$.

Therefore, the information which is relevant for queue estimation is given by the event

$$E_r = \{A_0 = t_0, A_1 \leq t_1, A_2 \leq t_2, \ldots, A_n \leq t_n, A_{n+1} > t_{n+1}\}$$

(The subject of relevant information will be treated rigorously in Subsection 2.1 of Chapter 2.)

From the well-known property of the Poisson process, external arrivals $A_1, \ldots, A_n$ are distributed as the order statistics of a uniform distribution on $[t_0, t_{n+1}]$, conditioned on the event $\{A_1 \leq t_1, \ldots, A_n \leq t_n\}$.

Larson derives an algorithm for calculating

$$\hat{A}(t) = \mathrm{E}[A(t)/E_r].$$

Then the queue estimate is given by:

$$\hat{Q}(t) = \hat{A}(t) - D(t),$$

where $D(t)$ is the observed number of cumulative departures during $(t_0, t]$.

The estimate $\hat{Q}(t)$ does not depend on the arrival rate λ since the joint distribution of order statistics does not depend on λ. (See Subsection 2.1 of Chapter 2 for the strict proof.)

2.3 Hidden Markov Models (HMM)

An HMM applies to a discrete time Markov chain $Y = \{Y_i,\ 0 \leq i \leq n\}$, on a countable state space $\mathcal{Y}$. (This Markov chain may be time-inhomogeneous.) In our special case, two elements of information are available for Y:

- Boundary conditions $\{Y_0 = i_0\}$ and $\{Y_n = i_n\}$.

- Taboo conditions $Y_i \notin \mathcal{B}_i$, $1 \leq i \leq n-1$, where $\mathcal{B}_i \subseteq \mathcal{Y}$ are called *taboo sets*. The complementary sets $\mathcal{G}_i = \mathcal{Y} \setminus \mathcal{B}_i$ will be referred to as *admissible sets*.

The HMM problem, whose algorithmic solutions is presented in Daley and Servi [4], is to calculate the conditional probabilities

$$\mathrm{P}\{Y_r = l / Y_0 = i_0;\ Y_n = i_n;\ Y_i \notin \mathcal{B}_i,\ 1 \leq i \leq n-1\}, \quad r = 1, \ldots, n-1;\ l \in \mathcal{Y}, \quad (2.1)$$

in terms of the transition probabilities of Y.

Details of the algorithm will be presented later, when we consider a specific problem in Subsection 1.2 of Chapter 2. The number of operations turns out to be $O(n^3)$ if $\mathcal{B}_i$ are bounded and the jump magnitude of the Markov chain is bounded either above or below. The algorithm can be slightly modified to cover more complicated boundary conditions such as initial or final distributions.

Larson's problem can be easily represented as HMM. Consider again an isolated busy period with the transaction times $t_0, t_1, \ldots, t_n, t_{n+1}$. Then, the sequence $A(t_0), A(t_1), \ldots, A(t_n)$ constitutes a Markov chain (time-inhomogeneous, in general).

The boundary conditions for this chain are given by $A(t_0) = 0$ and $A(t_n) = n$.

The taboo conditions are

$$A(t_i) \notin \{0, 1, \ldots, i-1\}, \quad i = 1, \ldots, n-1.$$

(It is identical to $A_i \leq t_i$ from the previous subsection.)

Then the general algorithm which calculates probabilities from (2.1) can be applied to computation of the conditional distribution of $A(t_i)$, $1 \leq i \leq n-1$, given boundary and taboo conditions.

Remark. This account is written "in the spirit of" Daley and Servi [4]. They refer to HMM but, most of the time, use the term: "Markov Chain Boundary Value Problem" which concerns the special case of our HMM problem (when the taboo sets $\mathcal{B}_i$ are identical for all i).

2.4 Examples

We present here several graphs of queue estimates, for the single-station case. The figures were obtained using the program QIE2 (see Chapter 3 for a detailed account on applications), and will now be briefly described.

Consider first a busy period, consisting of 20 deterministic service times equal to 1. Figure 2.1 shows the conditional expected value of the queue, together with its maximum likelihood estimate, i.e. the queue path with the maximum conditional probability (Daley and Servi [4] explain how to derive it in the case of deterministic services). Note that the expected queue length increases linearly between transaction points and, naturally,

decreases by one when a service terminates. The most likely queue path oscillates between zero and one. This example suggests that conditional expectation is a better representative of the sample path distribution then the most likely queue path (in the same way as the expected value is a better representative of a random variable distribution then the mode).

Figure 2.2 is a smoothed version of Figure 2.1. Note that the conditional expectation of the queue is slightly asymmetric. The explanation is given by Figure 2.3 which compares the cumulative number of departures (known) to the cumulative number of arrivals (estimated) during our busy period. The graph for cumulative arrivals is concave. Larson [12] established concavity for an arbitrary structure of transactions. In other words, the conditional arrival rate decreases during a busy period: more customers tend to arrive at the beginning and less at the end. However, when the number of customers in a busy period is large and services are deterministic, the queue estimate is asymptotically symmetric (see Daley and Servi [5]) and may be approximated using Brownian excursions. (Such asymptotic estimates will possibly be used in our future research.)

In Figures 2.4 and 2.5 we consider another busy period with 20 transactions. The first ten service times are equal to 1.5 and the last ten to 0.5. Figures 2.6 and 2.7 illustrate the opposite situation when the short services are at the beginning of the busy period.

It is seen that the queue estimate strongly depends on the transactions: the maximum of the expected queue length is twice as high in Figure 2.4 than Figure 2.6, the shapes of graphs are also very different. Note that, in the first case, the conditional expected number of arrivals (Figure 2.5) is very close to a straight line (namely, the unconditional expected arrivals) and in Figure 2.7 it is very concave. The intuitive explanation is that, given the departures in Figure 2.5, a queue forms even if arrivals are homogeneous during the considered period. As for Figure 2.7, a queue forms only if arrivals are very time-inhomogeneous.

Figure 2.1
Expectation vs. ML Estimate

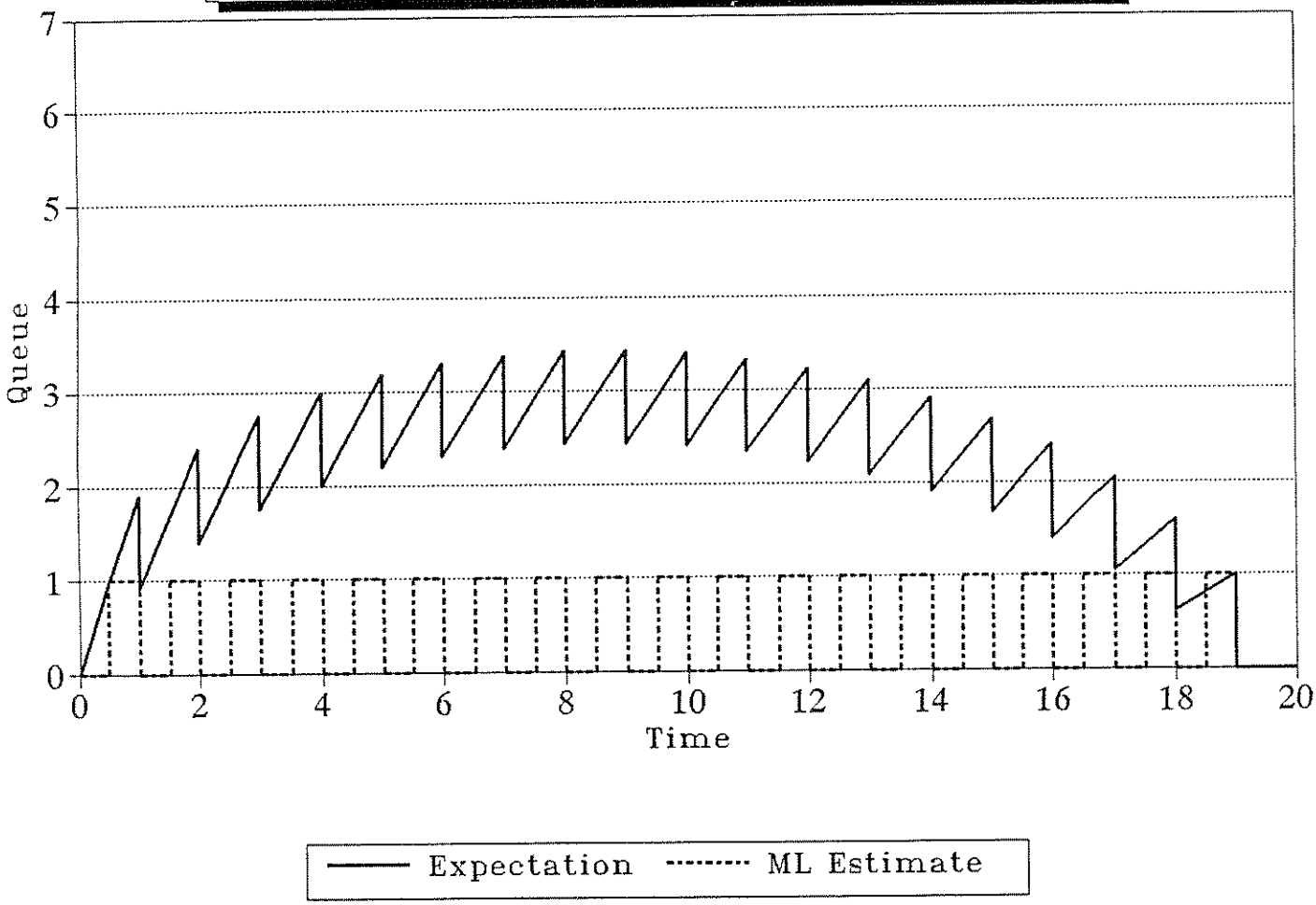


Figure 2.2
Expectation vs. ML Estimate: smoothed

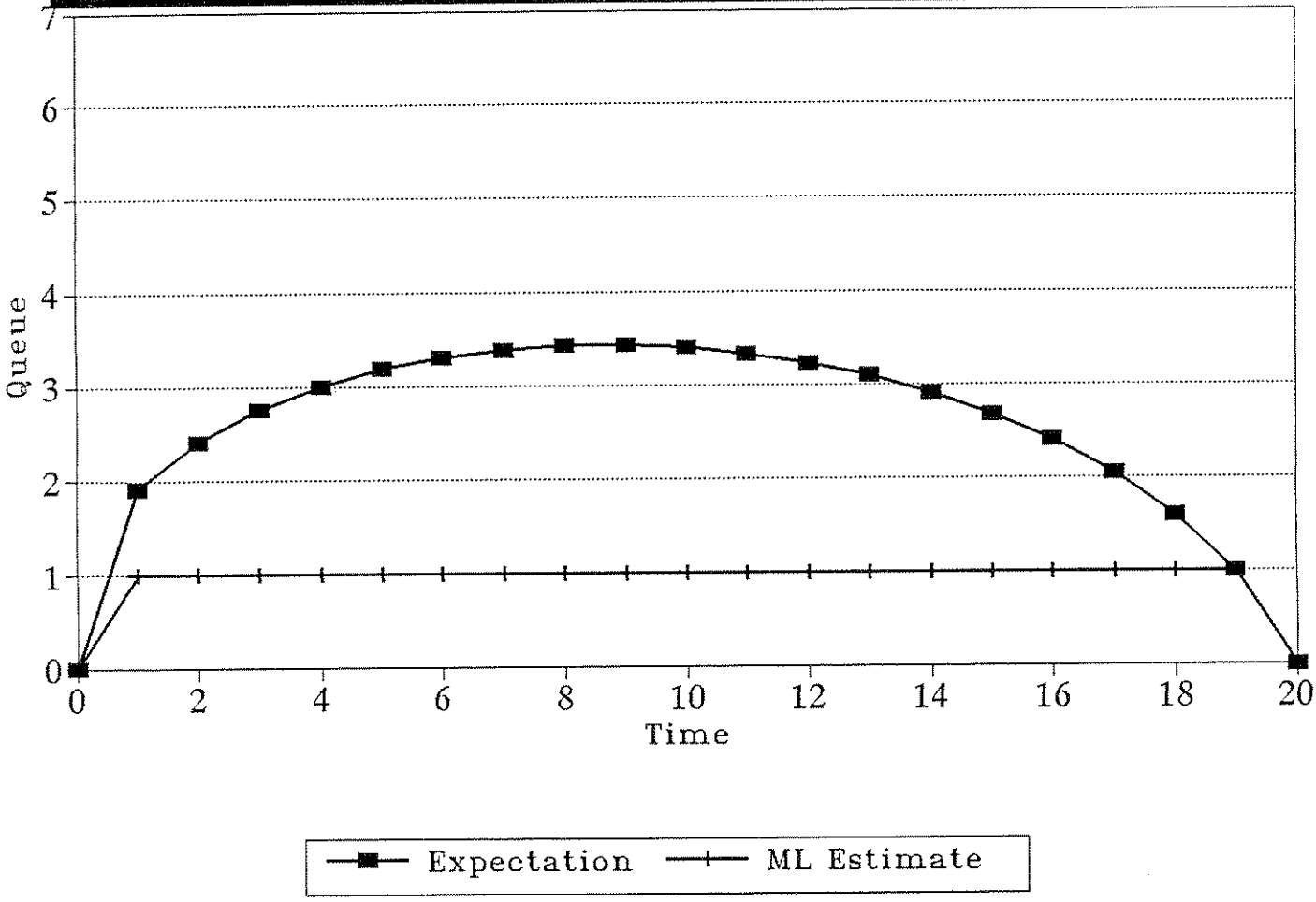


Figure 2.3
Arrivals vs. departures

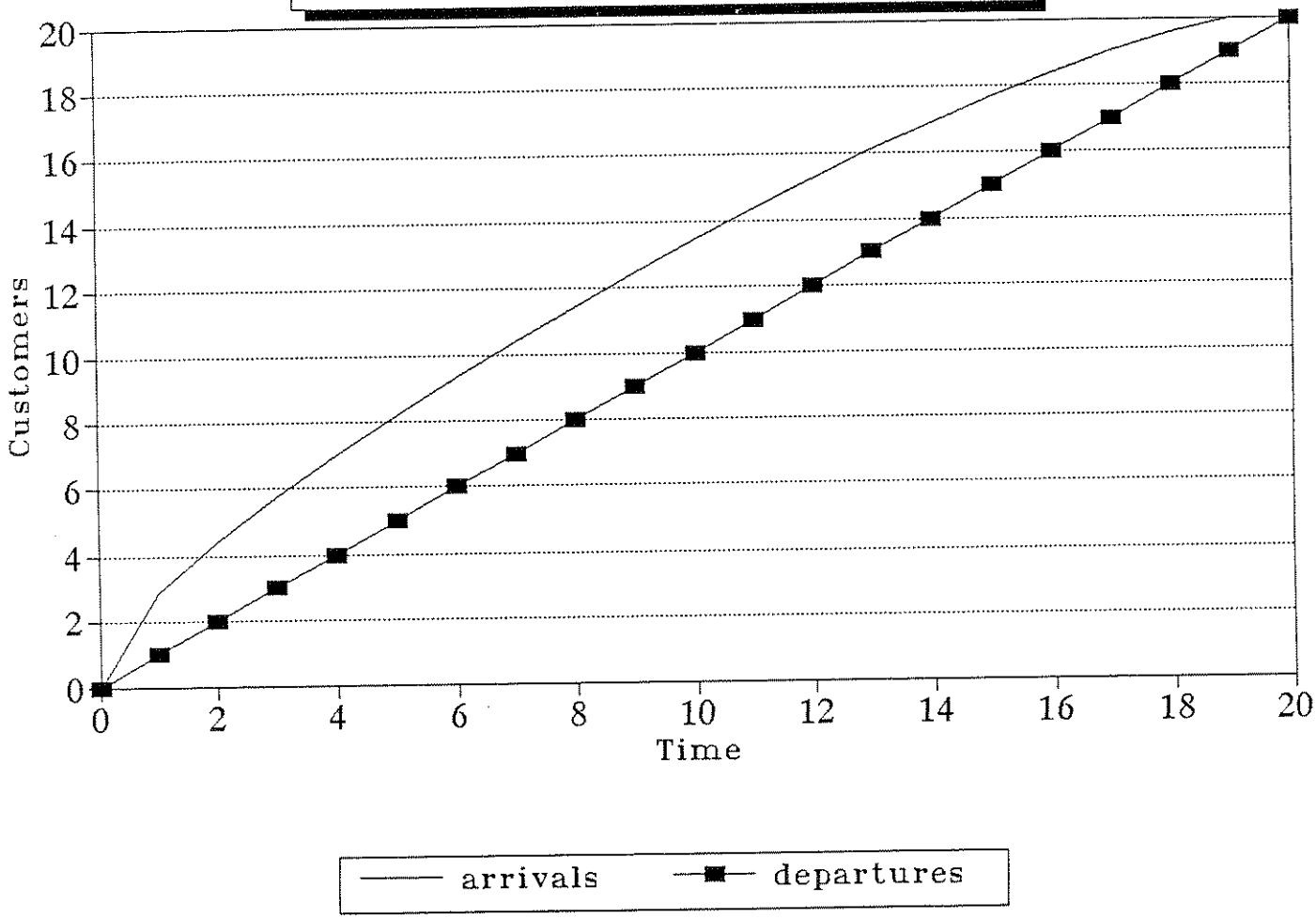


Figure 2.4
Long services at start

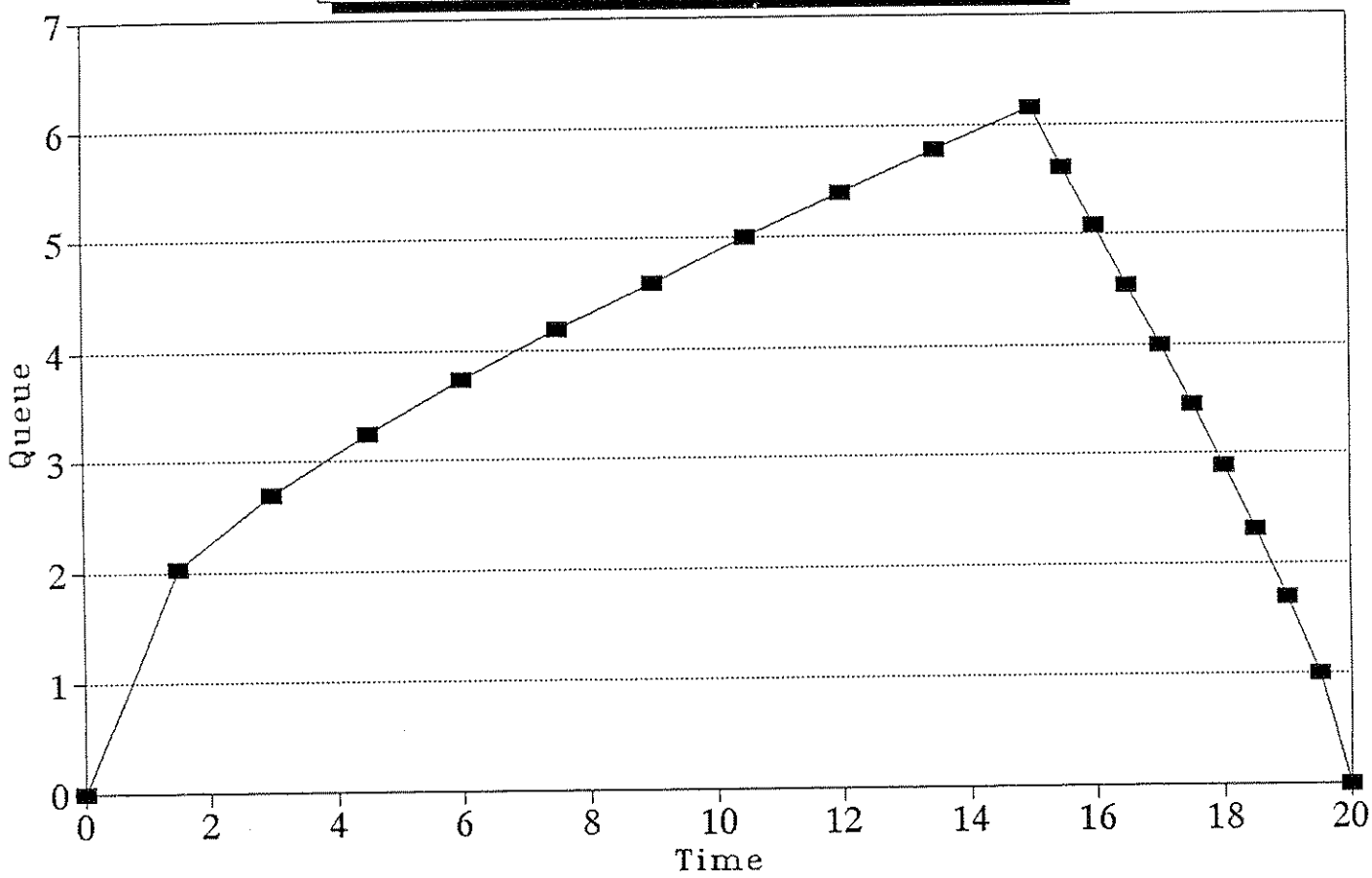


Figure 2.5
Long services at start: A(t) vs. D(t)

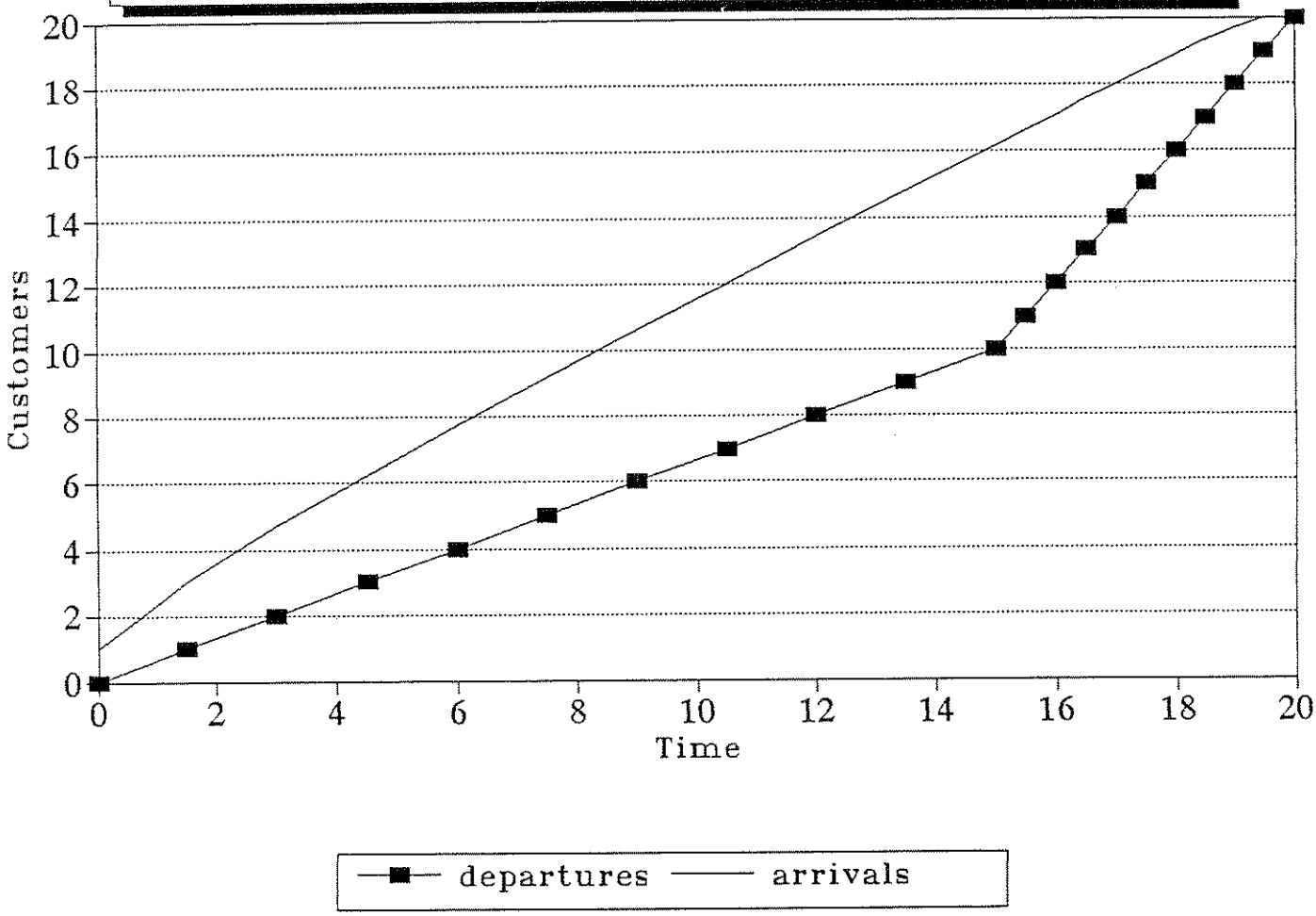


Figure 2.6
Long services at end

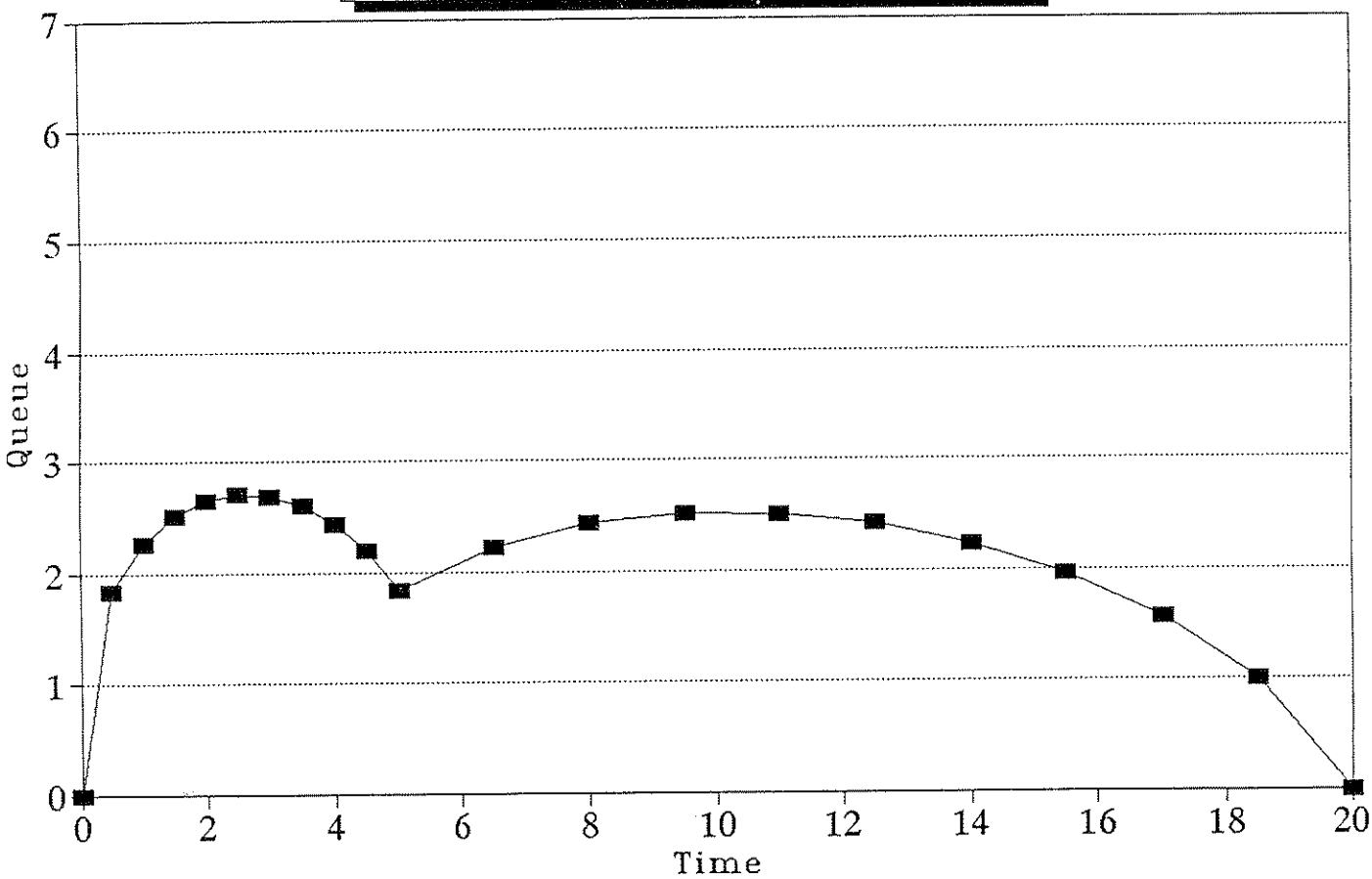
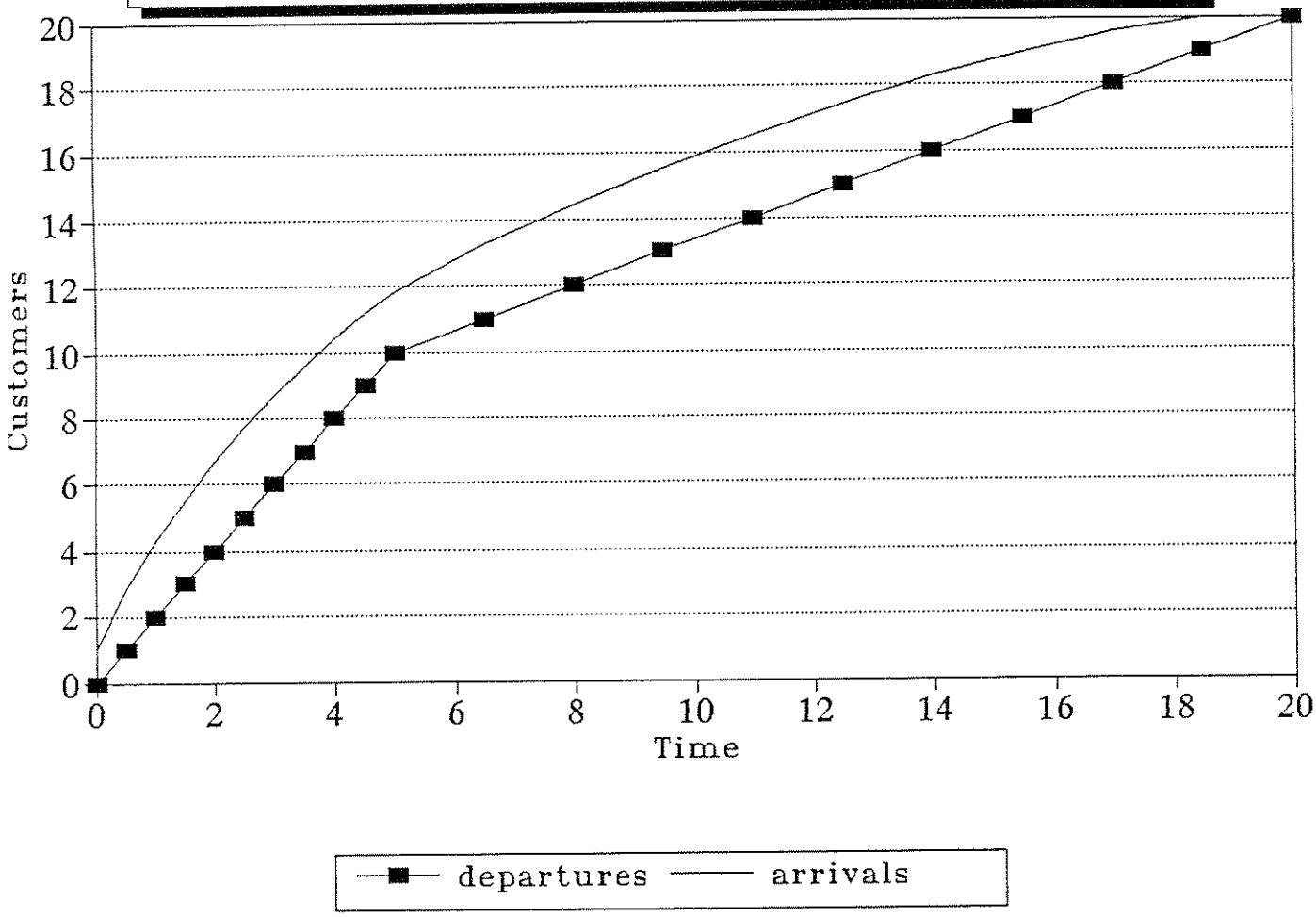


Figure 2.7
Long services at end: A(t) vs. D(t)



3 Formulation of Problems

3.1 Network Description

Consider an open queueing network with d stations, denoted as a set by $\mathcal{D} = \{1, 2, \ldots, d\}$. Prevailing assumptions about the network are as follows:

- External arrivals to the network constitute a Poisson process with a non-homogeneous arrival rate $\alpha = \{\alpha(t), t \geq 0\}$.
- Switches of customers between stations are governed by a d-dimensional routing matrix $P = [p_{jk},\ j, k \in \mathcal{D}]$.
- Service times $\xi_{ji},\ j \in \mathcal{D},\ 1 \leq i < \infty$ are completely general random variables (not necessarily independent or identically distributed).
- The stochastic components of the network, namely external arrivals, services and switches between stations, are independent of each other.
- Stations may be multi-server stations.
- First Come First Served (FCFS) queue discipline is assumed at all stations.
- Work-conserving principle prevails: a server cannot be idle if there is a customer waiting for service.
- An **Identification Number** (ID) is attached to each customer upon entering the network, and it accompanies him until exit.
- No simultaneous arrivals to stations take place.

In specific problems, some of these assumptions will be weakened or even omitted. The last technical assumption is used because, in the case of simultaneous arrivals, FCFS does not determine the service order.

It will become clear later that availability of ID's plays a key role in our approach and methods. However, knowledge of ID's considerably improves the queue estimates, as demonstrated in the applications of Chapter 3.

Two different models are treated:

- Transitions between stations are immediate.
- Transition times between stations are exponentially distributed with parameters $\eta_{jk},\ j, k \in \mathcal{D}$, independently of the other stochastic components of the network.

Note that the second case could be formalized, in terms of immediate transitions, by introducing on all routes, fictitious stations with an infinite number of servers and exponential service times.

We now elaborate on several features of the network. Most often we assume that the network has a single entrance, which is conveniently referred to as "station 0". (The extended set of stations $\{0, 1, \ldots, d\}$ is denoted by $\mathcal{D}_0$). Thus, at time t, new customers arrive to station 0 at a rate $\alpha(t)$, after which they switch to other stations according to the routing probabilities p_{0j}, $j \in \mathcal{D}$ (and transition rates η_{0j}, if transition times are exponential). Sometimes, however, the stations are assumed to have their own entrances and independent external arrival rates $\alpha_j(t)$. For the case of immediate transitions, the relation between the two models is clearly $\alpha_j(t) = p_{0j}\alpha(t)$.

It may be reasonable to assume sometimes that a measuring device is installed at the entrance of the network (see applications in Chapter 3, for example). Therefore, two cases are considered.

- External arrivals are registered.
- External arrivals are not registered.

As mentioned above, external arrivals constitute a non-homogeneous Poisson process. However it is worthwhile to consider separately external arrivals with a constant rate, mostly because in several special cases this rate can be assumed unknown. Solutions for the two cases are usually similar, so we choose the one more convenient or representative and then either extend or concretize briefly the solution for the second.

Non-external arrivals that come from other stations of the network will be called internal arrivals. Sometimes we shall also use the terminology "internal (external) customers" for customers that correspond to internal (external) arrivals.

3.2 Structure of Observations

The present subsection can be skipped without loss of continuity. Nevertheless, it may be helpful for general understanding of the problem and also for computer implementations (specifically, for the problems of data organization).

Suppose that the network is observed over the time interval $[0, t]$. We register service starts, service terminations and ID's of customers associated with these transactions. Sometimes external arrivals are also registered, and such cases will be treated separately.

While the structure of the observations is rather simple, we introduce four useful different points of view on the data. Only the third approach is used directly in our research, but they are all helpful for a better understanding of the network operation.

The first approach uses classical counting processes as in queueing theory. (All our random processes are assumed right-continuous with left limits.)

Counting. Let

$$A = (A_0,\ A_1, \ldots, A_d); \quad A_j = \{A_j(s),\ s \geq 0\}, \quad j \in \mathcal{D},$$

denote the arrival process of our network: $A_j(s)$ is the number of arrivals to station j, both external and internal, up to time s. For convenience, we denote $A_s = \{A_j(s), j \in \mathcal{D}\}$.

In the same way, let $V = (V_1, \ldots, V_d)$ and and $D = (D_1, \ldots, D_d)$ denote the service start and the departure processes respectively. Finally, let C_{jk} denote the identification number of the k-th customer that started service at the j-th station, and put

$$C = \{C_j(s),\ s \geq 0,\ j \in \mathcal{D}\},$$

where

$$C_j(s) = (C_{j1}, C_{j2}, \ldots, C_{jV_j(s)}).$$

If external arrivals are registered, then

$$C_0(s) = (C_{01}, C_{02}, \ldots, C_{0A_0(s)}).$$

The *available information* at time t is given by $\mathcal{F}_t = \sigma(V_s,\ D_s,\ 0 \leq s \leq t;\ C_t)$ or $\sigma(A_0(s),\ V_s,\ D_s,\ 0 \leq s \leq t;\ C_t)$. (To avoid obvious repetitions, we assume for the following representations that external arrivals are registered.)

Note that if, in addition, we observe arrival processes to stations, then it is possible to determine the states of all stations at all times in $[0, t]$. For example, we could calculate queue length (knowledge of A and D is sufficient for this), waiting times, et.c. Thus, our problem reduces to that of inferring arrivals to the stations.

Now suppose that our observations of the network are saved in a database. The data in such a hypothetical database can be organized by at least three methods: according to registrations of network transactions, according to busy periods and according to routes (or traces) of customers. The following three approaches to data organization are in line with these methods.

Registrations. Suppose that a computer registers observable events during the time interval $[0, t]$. The available information is then likely to be organized as follows: $\mathcal{F}_t = \sigma(Z_t)$, where

$$Z_t = (Z_0(t), Z_1(t), \ldots, Z_d(t)); \quad Z_j(t) = \{(t_{ji}, I^D_{ji}, K_{ji}),\ t_{ji} \leq t,\ i \geq 1\}, \quad j \in \mathcal{D}$$

and each triplet $(t_{ji}, I^D_{ji}, K_{ji})$ corresponds to a single registration. Here

t_{ji} — registration times, where registrations are external arrivals, service starts or service terminations.

I^D_{ji} — identification numbers,

K_{ji} — registration codes (0 stands for external arrivals, 1 for service starts and 2 for service terminations).

Busy periods. A busy period at a station is defined as a maximal time interval, over which all servers at that station are busy. We noted that a busy period can be also defined as a period of immediate start of new services after service terminations. In the case of a single-server station, the observed information can be represented as:

$$\mathcal{F}_t = \sigma(A_0(s),\ 0 \le s \le t;\ B_j(t),\ j \in \mathcal{D}),$$

where

$$B_j(t) = \{(t^j_{11},\ t^j_{12}, \ldots, t^j_{1n_1}), (t^j_{21}, \ldots, t^j_{2n_2}), \ldots, (t^j_{k1}, \ldots, t^j_{kn_k}),\ t^j_{mi} \le t\}.$$

The times in the brackets are registrations of different busy periods. Times within a busy period are marked by two ID's (the first ID belongs to a customer that terminated service and the second one is associated with a new start) and times that start or end a busy period and times of external arrivals are marked by a single ID. Note, that the last busy period at a station could be incomplete.

In the case of a single-server station, every registration belongs to some busy period. Multi-server station case is different since at the same time there may be busy and idle servers. Notation for this case would be more complicated.

The approach to data organization via Busy Periods is basic, helping us to develop various representations of the observed information in Sections 2 and 3 of Chapter 2.

Traces of customers. Suppose that the basic data unit in our database corresponds to a route, or *trace*, of a customer. Specifically

$$\mathcal{F}_t = \sigma(I^D_1(t), \ldots, I^D_{l_t}(t)),$$

where l_t denotes the number of customers that have been registered during $[0, t]$ and

$$I^D_k(t) = \{a_{k0},\ (n_{k1}, b_{k1}, e_{k1}), \ldots, (n_{km_k}, b_{km_k}, e_{km_k}),\ b_{ki} \le t,\ e_{ki} \le t,\ i \ge 1\}.$$

In the last expression, a_{k0} denotes the external arrival time of the k-th customer, n_{ki} are the stations on his route (chronologically ordered from entrance till departure), b_{ki} and e_{ki} are times of service starts and terminations.

3.3 Objectives

Introduce the right-continuous queue length process

$$Q = \{Q_s,\ s \ge 0\}; \quad Q_s = (Q_1(s), \ldots, Q_d(s)).$$

(We define a queue as a set of customers waiting for service.) Our objective is to estimate the unobserved queue process Q, given the available information $\mathcal{F}_t$, $t \ge 0$. Three different estimation problems are posed:

Busy-Period Interpolation Let $[T_1, T_2]$ denote a busy period at some station j. Based on the available information $\mathcal{F}_t$, $t \geq T_2$, estimate the sample path $Q_j(s)$, $s \in [T_1, T_2]$.

Real-Time Estimation Estimate in real-time $Q(t)$, based on the available information $\mathcal{F}_t$.

General Interpolation Estimate the sample path $Q(s)$, $0 \leq s \leq t$, based on the available information $\mathcal{F}_t$.

To summarize, we deal with three cases, distinguished by estimation objective, transition times (immediate or exponential) and external arrivals (known or unknown): all in all $3 \times 2 \times 2 = 12$ possibilities. Real-time and general interpolation will be treated together.

3.4 Summary of Results

Table 1. Summary of Results.

	Immediate transitions		Exponential transitions	
	External arrivals			
	known	unknown	known	unknown
Busy Periods Interpolation	Queue length can be computed exactly.	Internal arrivals known. Queue of external arrivals estimated using HMM technique.	HMM technique. Properties of truncated exponentials.	HMM; truncated exponentials; properties of $M/G/\infty$ queue.
Assumptions	None.	FCFS; Poisson external arrivals.	FCFS; transition rates η_{ij}.	FCFS; Poisson external arrivals; η_{ij}.
Real-time estimation. General interpolation	Simple combinatorics.	Real-time algorithm of Bertsimas, Servi. Matrix-type algorithm.	HMM with cumbersome state space; truncated exponentials.	HMM; truncated exponentials; Outline of approximate solution.
Assumptions	FCFS; routing probabilities p_{ij}.	FCFS; p_{ij}; external arrival rates λ_j.	FCFS; p_{ij}; transition rates η_{ij}.	FCFS; p_{ij}; η_{ij}; λ_j.

Remarks.

- The complexity of the problems and the number of assumptions increase rapidly when one moves from the upper left corner of the table to the lower right one.
- The two cases of exponential transitions and real-time estimation involve cumbersome computations. For the last case we provide only an outline of the solution.
- Homogeneous Poisson arrivals were assumed throughout the table. If arrivals are non-homogeneous Poisson then the arrival rate $\lambda(t)$ must be known for all real-time cells of the table. From a practical point of view, this is reasonable: estimating $\lambda(t)$, say a periodic function, requires data over many periods; such estimation should be carried out prior to queue-inference.

Chapter II

Theoretical Results

1 Busy-Period Interpolation

Busy-Period Interpolation problems cover a specific busy period, which terminated before our estimation time t. Here we focus on such a busy period, hence notation can and will be simplified (throughout this section only) by omitting the indices of the busy period and its associated station. Registrations during the busy period will be denoted by

$$t_0,\ t_1, \ldots, t_n, t_{n+1},$$

where
t_0 is the start of the busy period,
$t_1, \ldots, t_n$ are service starts, which are also service terminations and
t_{n+1} is the end of the busy period.
Without loss of generality, we assume that estimation time is $t = t_{n+1}$.

Note that in the multi-server case, a customer that starts service at t_i need not leave at t_{i+1}.

Let $A_0 = t_0, A_1, \ldots, A_n$ stand for the arrival times of the customers that started service at $t_0,\ t_1, \ldots, t_n$, respectively. The cumulative number of arrivals during $(t_0, t]$ will be denoted by $A(t)$, $t_0 \leq t \leq t_n$ $(A(t_0) = 0,\ A(t_n) = A(t_{n+1}) = n)$.

The following question is fundamental: which part of the available information $\mathcal{F}_t$ is relevant for our estimation of the queue length?

This problem will be treated rigorously only in Section 2 onwards. As for the current busy period interpolation problem, treated in Section 1, the structure of the relevant information is simple, hence only intuition-based statements will be presented. This simplicity is due to the availability of ID's of all the customers that were served during the busy period under consideration. The problem, then, reduces to making inferences about arrival times of a completely specified customer set. The knowledge of this set makes it also possible to omit any Markovian assumptions on customers routing.

1.1 Immediate Transitions, Known External Arrivals.

This problem is deterministic in the sense that one can calculate the queues exactly. To clarify this point, represent the queue process on the j-th node as

$$Q_j(t) = Q_j^{\alpha}(t) + Q_j^{\beta}(t),$$

where $Q_j^{\alpha}(t)$ is the part of the queue due to external customers (customers that arrived from station 0) and $Q_j^{\beta}(t)$ is due to internal ones.

The quantity $Q_j^{\alpha}(t)$ is known exactly, according to problem definition. Consider the second term. If a customer arrived from some other station, then his arrival time to the j-th station coincides with the termination time of his last service. But we know the time of the last termination because all the network is observed and all the ID's are registered. Information about arrivals to the station provides us with the queue size. Note that Poisson arrivals are not necessary here. This is also the only case where the FCFS queue discipline plays no role.

1.2 Immediate Transitions, Unknown External Arrivals.

Model Definition and Relevant Information We distinguish again between external and internal customers. The arrivals of internal ones are known, following the analysis in 1.1. We need, therefore, to estimate only external arrivals. Suppose that the external arrivals constitute a homogeneous Poisson process with a known arrival rate λ. (At the end of this subsection, we modify our results to accommodate the non-homogeneous case.) By the memoryless property of the exponential distribution, we can assume that this process starts at t_0.

Suppose that t_i is a service start of some internal customer; then the last service termination of this customer, which is known, will be denoted by s_i. The relevant information can now be represented as the intersection of the following four events:

E_1. $A_0 = t_0$ (the first arrival coincides with the start of our busy period);

E_2. $A_i = s_i$, if A_i is an internal arrival time;

E_3. $A_i \leq t_i$, if A_i is an external arrival time (a customer must arrive before his service starts);

E_4. If $A_k < A_l$ are any adjacent internal arrival times, then

$$s_k = A_k < A_{k+1} < A_{k+2} < \ldots < A_{l-1} < A_l = s_l.$$

If the first or last arrival of the busy period is not internal then

$$t_0 < A_1 < A_2 < \ldots < A_{l-1} < A_l = s_l,$$

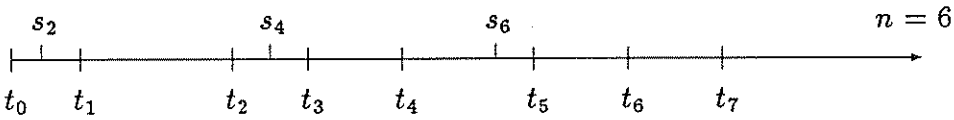
in case A_l is the first internal arrival and

$$s_k = A_k < A_{k+1} < A_{k+2} < \ldots < A_n \leq t_n,$$

in case A_k is the last internal arrival. (Event E_4 follows from event E_2 and FCFS, which implies $t_{i+1} > t_i \Rightarrow A_{i+1} > A_i$.)

Formulation as HMM Consider two adjacent internal arrival times $A_k = s_k$ and $A_l = s_l$. The interval $[0, t_n]$ is divided into intervals of the type $[s_k, s_l]$ and then cumulative arrivals $A(t)$ and queue length are estimated separately for every such interval (see Figure 1.1).

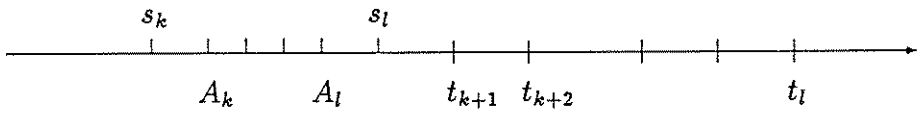
Figure 1.1. Immediate Transitions. Unknown External Arrivals. Example of Busy Period.



Customers 2, 4 and 6 are internal. Queue estimation is carried out separately on $[t_0, s_2)$, $[s_2, s_4)$, $[s_4, s_6)$ and $[s_6, t_6)$.

We consider two cases, according to the location of s_l relative to t_{k+1}.
First case: $t_{k+1} > s_l$ (see Figure 1.2).

Figure 1.2. Immediate Transitions. Unknown External Arrivals. Case 1.



Then event E_3 from the relevant-information list follows from event E_4, for $i \in \{k+1, \ldots, l-1\}$, and it can be omitted. From event E_4 and the order-statistics property of the Poisson process, we conclude that $A_{k+1}, \ldots, A_{l-1}$ are distributed as uniform order statistics on $[s_k, s_l]$. This means that for $t \in [s_k, s_l)$, $0 \leq i \leq l-k-1$,

$$\mathrm{P}\{A(t) = k + i/E_r\} = \binom{i}{l-k-1} \frac{(t-s_k)^i (s_l - t)^{l-i-1}}{(s_l - s_k)^{l-k-1}}, \tag{1.1}$$

$$\hat{A}(t) = \mathrm{E}[A(t)/E_r] = k + (l-k-1)\frac{t-s_k}{s_l - s_k}, \tag{1.2}$$

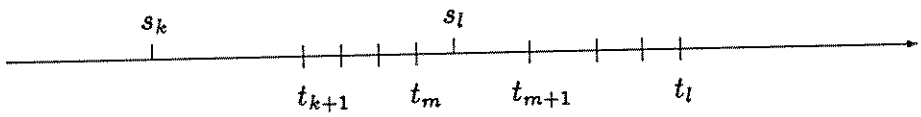
$$\mathrm{Var}[A(t)/E_r] = (l-k-1)\frac{(t-s_k)(s_l-t)}{(s_l-s_k)^2}, \tag{1.3}$$

where E_r is the relevant information. As for the queue length, if $t \in [t_i, t_{i+1})$, $0 \leq i \leq n-1$, then

$$\hat{Q}(t) = \hat{A}(t) - i \tag{1.4}$$

Second case: $t_{k+1} < \ldots < t_m < s_l < t_{m+1} < \ldots < t_l$ (see Figure 1.3).

Figure 1.3. Immediate Transitions. Unknown External Arrivals. Case 2.



Now $A_{k+1}, \ldots, A_{l-1}$ are distributed as uniform order-statistics on $[s_k, s_l]$, conditioned on $\{A_{k+1} \leq t_{k+1}, A_{k+2} \leq t_{k+2}, \ldots, A_m \leq t_m\}$. (We do not use explicitly this joint distribution of arrivals for queue calculation, though such an approach is feasible, as in Larson [12].) To formalize the problem in HMM terms, consider the non-homogeneous Markov chain

$$\{A(s_k), A(t_{k+1}), \ldots, A(t_m), A(s_l)\}. \tag{1.5}$$

The boundary conditions are $A(s_k) = k$ and $A(s_l) = l$.

The set of admissible states at t_i is given by

$$\mathcal{G}_i = \{i, i+1, \ldots, l-1\}, \quad k+1 \leq i \leq m,$$

since at least i arrivals must occur before the start of the i-th service.

The taboo conditions are given by $A(t_i) \in \mathcal{G}_i$.

Remark. Our Markov chain has been constructed in such a way that the events that constitute the boundary and taboo conditions are identical to the relevant information.

To simplify the presentation, we maintain index continuity by assigning

$$z_k = s_k,\ z_{k+1} = t_{k+1}, \ldots, z_m = t_m,\ z_{m+1} = s_l.$$

It is straightforward to calculate transition probabilities between admissible states of the Markov chain in (1.5):

$$\begin{aligned} \tilde{p}_i(u,v) &= \mathrm{P}\{A(z_i) = v / A(z_{i-1}) = u\} = \\ &= e^{-\lambda(z_i - z_{i-1})} \frac{(\lambda(z_i - z_{i-1}))^{v-u}}{(v-u)!}, \quad k+1 \leq i \leq m,\ i \leq u \leq v \leq l-1, \end{aligned} \tag{1.6}$$

$$\begin{aligned} \tilde{p}_{m+1}(u,v) &= \mathrm{P}\{A(z_{m+1}) = l / A(z_m) = u\} = \\ &= e^{-\lambda(z_l - z_m)} \frac{(\lambda(z_{m+1} - z_m))^{l-1-u}}{(l-1-u)!}, \quad m \leq u \leq l-1. \end{aligned} \tag{1.7}$$

All other transition probabilities between admissible states are equal to zero.

Further introduce the events

$$B^{r_1,r_2} = \{A(t_i) \in \mathcal{G}_i,\ r_1 \leq i \leq r_2\}, \quad k+1 \leq r_1 < r_2 \leq m.$$

The taboo probability matrices are defined as $W^{r_1,r_2} = \{w_{u,v}^{r_1,r_2}\}$, and

$$w_{u,v}^{r_1,r_2} = \mathrm{P}\{A(z_{r_2}) = v; B^{r_1+1,r_2-1}/A(z_{r_1}) = u\}, \quad k \le r_1 \le r_2 \le m+1,$$

where u and v must be admissible states.

The algorithm for the calculation of the queue distribution at registration times is as follows:

Algorithm 1.1 (Follows Daley and Servi [4].)

1°. Using (1.6) and (1.7), define transition matrices between admissible states:

$$\tilde{P}_i = \{\tilde{p}_i(u,v),\ u \in \mathcal{G}_{i-1},\ v \in \mathcal{G}_i\}, \quad k+1 \le i \le m+1.$$

The dimensions of these matrices are $1 \times (l-k-1)$ for $i = k+1$, $(l-i+1) \times (l-i)$ for $k+2 \le i \le m$, and $(l-m) \times 1$ for $i = m+1$.

2°. Calculation of the taboo probabilities matrices:

$$\begin{aligned} W^{k,k+1} &= \tilde{P}_{k+1}, \\ W^{k,i} &= W^{k,i-1}\tilde{P}_i,\ k+2 \le i \le m+1, \\ W^{m,m+1} &= \tilde{P}_{m+1}, \\ W^{i,m+1} &= \tilde{P}_{i+1}W^{i+1,m+1},\ k+1 \le i \le m-1. \end{aligned}$$

3°. Calculation of the distribution for cumulative number of arrivals at registration times:

$$\begin{aligned} \mathrm{P}\{A(t_i) = u/E_r\} &= \mathrm{P}\{A(t_i) = u/A(s_k) = k; A(s_l) = l; A(t_r) \in \mathcal{G}_r,\ k+1 \le r \le m-1\} \\ &= \frac{w_{k,u}^{k,i} w_{u,l}^{i,m+1}}{W^{k,i}W^{i,m+1}}. \end{aligned}$$

(The denominator of the last expression is the scalar product of two vectors.)

4°. Calculation of the queue distribution at the registration times, using

$$\hat{Q}(t_i) = \hat{A}(t_i) - i, \quad k+1 \le i \le m.$$

Discussion and Extensions

- Suppose that t is not a registration point and one needs to estimate the queue at time t. If only conditional expectation of the queue is of interest, then one can use its linearity between the registration points (see [4]). Otherwise, t can be added to the domain of the Markov chain (1.5). If $t \in (t_i, t_{i+1})$, then the set of the admissible states for t is $\{i, i+1, \ldots, m\}$. The transition probabilities for $A(t_i) \to A(t)$ and $A(t) \to A(t_{i+1})$ are derived using the Poisson distribution.

- Given the conditional distribution of queue-length, it is easy to calculate conditional expectation, variance and other distributional characteristics.

- As in the single-station case, the queue-length distribution does not depend on λ. Indeed, this follows from formulae (1.1)-(1.4) for the first case. For the second case, the distribution of external arrivals is that of uniform order statistics, conditioned on an event whose probability does not depend on λ.

  The independence on λ is useful for debugging of any software implementation of the algorithm: simply substitute two distinct values of λ and check that probabilities do not change.

- Suppose that external arrivals to the station constitute a non-homogeneous Poisson process with a known arrival rate $\lambda(t)$. To accommodate such a generalization, let $\Lambda(t) = \int_0^t \lambda(s)ds$ stand for the cumulative arrival rate. The transition probabilities from (1.6) are now:

$$\tilde{p}_i(u,v) = e^{-(\Lambda(z_i)-\Lambda(z_{i-1}))}\frac{(\Lambda(z_i)-\Lambda(z_{i-1}))^{v-u}}{(v-u)!}, \quad k+1 \le i \le m,\ i \le u \le v \le l-1. \tag{1.8}$$

  Equation (1.7) is modified similarly.

- The jump magnitude of our hidden Markov chain is bounded, hence the number of calculations does not exceed $O(n^3)$, where n is the number of estimation points in an interval between two successive external arrivals. It is important, therefore, from a computational point of view, that our algorithm is executed separately on each interval between successive internal arrivals, since such intervals typically include less points than a whole busy period. Note also that for many queueing networks it is reasonable to expect that the number of internal arrivals in a busy period with n customers is of order n. For example, in the steady-state of a Jackson network, the long-run proportion of internal arrivals to station j is given by $(\lambda_j - \alpha_j)/\lambda_j$, where α_j is the external arrival rate and λ_j is the solution of flow conservation equations (see [16], for example).

  If every k-th arrival is internal, the number of calculations needed to estimate the queue at registration points will be reduced to $O(n)$. If internal arrivals are "approximately uniformly" dispersed between external ones, the same result can be expected. The knowledge of external arrival times enables, therefore, a decrease in the number of calculations even in comparison with the case of a single service station.

- Implementation of the algorithm for real data is presented in Chapter 3.

1.3 Exponential Transitions, Known External Arrivals

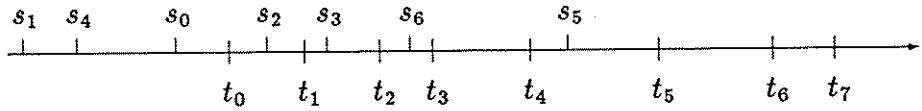
Model Definition and Relevant Information Suppose that the arrivals, service times and routing mechanism of our network are completely general. In the present model we eliminate the assumption of immediate transitions between stations. This assumption is indeed too restrictive for many applications, such as transportation models and human service systems. (It is appropriate, however, for many communication networks.)

Assume that the transition time of a customer, routed from station j to station k, is exponentially distributed with a known parameter η_{jk}, $j \in \mathcal{D}_0$, $k \in \mathcal{D}$. New customers are registered upon arrival to station 0 and then proceed to station j with probability p_{0j}, $j \in \mathcal{D}$. Finally, every transition time is assumed independent of all other stochastic components of the model. Note that our model allows overtaking between customers.

We now determine the part of the available information $\mathcal{F}_t$, which is relevant for queue estimation.

Consider a busy period at some station of the network, with registrations $t_0, t_1, \ldots, t_n, t_{n+1}$, as before. For the customer whose registration time is t_i, $0 \leq i \leq n$, let s_i denote the last registration time prior to t_i. These times will be referred to as *transition start times* (see Figure 1.4). They may be either service termination or arrival times to station 0. According to our model specification, $s_0, s_1, \ldots, s_n$ can be deduced from our data.

Figure 1.4. Exponential Transitions. Known External Arrivals. Example of Busy Period.



Enumerate by $1, \ldots, n$ the customers that started service at $t_1, \ldots, t_n$. Letting $A_1, \ldots, A_n$ denote their unknown arrival times as usual, we have

$$A_i = s_i + \tau_i, \quad 1 \leq i \leq n, \tag{1.9}$$

where $\tau_i \sim \exp(\gamma_i)$, γ_i are the corresponding known transition rates and τ_i are independent of each other.

Two important facts about arrival times are known:

E_1. Every customer arrived before his service start, namely

$$A_1 \leq t_1,\ A_2 \leq t_2, \ldots, A_n \leq t_n;$$

E_2. From the FCFS queueing discipline,

$$t_0 = A_0 < A_1 < A_2 < \ldots A_n \leq t_n.$$

The arrivals to the station are distributed according to (1.9), conditioned on the event $E_1 \bigcap E_2$.

Formulation as HMM Introduce a Markov chain

$$\tilde{A} = \{\tilde{A}(t_0), \ldots, \tilde{A}(t_n)\},$$

whose state at t_i is the set of customers that arrived during $(t_0, t_i]$. The set is ordered according to arrival times. In other words, the state space $\mathcal{A}$ is

$$\{\emptyset\} \bigcup \{m_1, \ldots, m_l\}, \quad m_j \in \{1, \ldots, n\}, \ m_i \neq m_j.$$

(The state space is more complicated than in the previous subsection since event E_2 must be accommodated in the taboo conditions.)

The boundary conditions for the busy period interpolation are:

$$\tilde{A}(t_0) = \{\emptyset\}; \ \tilde{A}(t_n) = \{1, \ldots, n\}.$$

The set of admissible states at t_i is given by

$$\mathcal{G}_i = \{ \{1, 2, \ldots, i\}, \{1, 2, \ldots, i, i+1\}, \{1, 2, \ldots, n\} \}.$$

All that is lacking for HMM are the transition probabilities between admissible states. We shall use capital letters as a convenient notation for these states. For example

$$I \stackrel{\text{def}}{=} \{1, 2, \ldots, i\}, \quad N \stackrel{\text{def}}{=} \{1, 2, \ldots, n\}, \quad K + L \stackrel{\text{def}}{=} \{1, 2, \ldots, k+l\}.$$

Then

$$\mathrm{P}\{\tilde{A}(t_{i+1}) = N / \tilde{A}(t_i) = N\} = 1, \quad 0 \leq i \leq n-1,$$

and for $0 \leq k \leq n-1, \ k+l \leq n$,

$$\begin{aligned} \mathrm{P}\{\tilde{A}(t_{i+1}) = K / \tilde{A}(t_i) = K\} &= \prod_{j=k+1}^{n} \exp\{-\gamma_j (t_{i+1} - \max(t_i, s_j))_+\}, \\ \mathrm{P}\{\tilde{A}(t_{i+1}) = K + L / \tilde{A}(t_i) = K\} &= 0, \quad \text{if } \exists s_m \in \{s_{k+1}, \ldots, s_{k+l}\} \ni \ s_m > t_{i+1}. \end{aligned}$$

Otherwise,

$$\begin{aligned} &\mathrm{P}\{\tilde{A}(t_{i+1}) = K + L / \tilde{A}(t_i) = K\} = \\ &= \mathrm{P}\{t_{i+1} < A_{k+l+1}, \ldots, t_{i+1} < A_n; t_i < A_{k+1} \leq t_{i+1}, \ldots, t_i < A_{k+l} \leq t_{i+1}; \\ &\quad A_{k+1} < A_{k+2} < \ldots < A_{k+l} / t_i < A_{i+1}, \ldots, t_i < A_n\} \\ &= \prod_{j=k+l+1}^{n} \exp\{-\gamma_j (t_{i+1} - \max(t_i, s_j))_+\} \prod_{j=k+1}^{k+l} (1 - \exp\{-\gamma_j (t_{i+1} - \max(t_i, s_j))\}) \ \times \\ &\times \ \mathrm{P}\{A_{k+1} < A_{k+2} < \ldots < A_{k+l} / t_i < A_{k+1} \leq t_{i+1}, \ldots, t_i < A_{k+l} \leq t_{i+1}\}. \end{aligned} \tag{1.10}$$

Consider the last term of (1.10). Given $\{t_i < A_j \leq t_{i+1}\}$, we have

$$A_j \sim \ \text{truncexp}(\gamma_j; \max(s_j, t_i), t_{i+1}).$$

A recursive algorithm for the calculation of the probabilities associated with this term is presented in Subsection 1.5. The key formulas are (1.14) and (1.15).

The following algorithm for calculating the queue distribution is very similar to Algorithm 1.1 in Subsection 1.2:

Algorithm 1.2

1°. Define transition probability matrices between the non-taboo states:

$$\tilde{P}_i = \{\tilde{p}_i(m,l),\ m \in \mathcal{G}_{i-1},\ l \in \mathcal{G}_i\},\ 1 \le i \le n.$$

The dimensions of these matrices are $1 \times n$ for $i = 1$ and $(n-i+2) \times (n-i+1)$ for $2 \le i \le n$.

2°. Calculation of the taboo probability matrices:

$$\begin{aligned} W^{0,1} &= \tilde{P}_1, \\ W^{0,i} &= W^{0,i-1}\tilde{P}_i,\ 2 \le i \le n, \\ W^{n-1,n} &= \tilde{P}_n, \\ W^{i,n} &= \tilde{P}_{i+1}W^{i+1,n},\ 0 \le i \le n-2. \end{aligned}$$

3°. Calculation of arrival number distribution at the registration times.

If $A(t)$ — cumulative number of arrivals to the station in $(t_0, t]$, then

$$\begin{aligned} &\mathrm{P}\{A(t_i) = k / A_1 \le t_1, \ldots, A_n \le t_n; t_0 < A_1 < A_2 < \ldots < A_n\} = \\ &= \mathrm{P}\{\tilde{A}(t_i) = K / \tilde{A}(t_i) \in \mathcal{G}_i,\ 0 \le i \le n\} = \\ &= \frac{w_{0,i}^{0,k} w_{k,n}^{i,n}}{W^{0,i}W^{i,n}} \end{aligned}$$

4°. Calculation of queue distribution at the registration times using $Q(t_i) = A(t_i) - i$.

Discussion

- Unlike the case in Subsection 1.2, it is no longer true that the expectation of the queue is linear between registration points.
- If queue distribution at some non-registration time t is of interest, then t must be added to the domain of the Markov chain. Specifically, if $t \in (t_i, t_{i+1})$, then the set of the admissible states for t is $\{I, \ldots, N\}$.

1.4 Exponential Transitions, Unknown External Arrivals

Model Definition and Relevant Information The treatment here resembles that of the previous model in Subsection 1.3. However, several points concerning external

arrivals must be modified. Suppose that external arrival times to the network constitute a Poisson process (possibly non-homogeneous) with a rate function $\alpha = \{\alpha(t),\ -\infty < t < \infty\}$. (Note that this subsection is the only one where we start the arrival process at the distant past. The reason is that in the case of exponential transitions arrivals to stations may depend on arrivals to the entrance of the network at the distance past. In all the other models, the start point of the arrival process is of no importance.)

New customers arrive to station 0 (there are no registrations of arrival times now) and then switch to other stations according to routing probabilities $p_{0j}, j \in \mathcal{D}$. Transition times between the stations are again exponentially distributed. Hence external arrivals to stations are departures from $M_t/M/\infty$ queues.

From the general theory of the $M_t/G/\infty$ queue (see [6]), the external input to any station j is Poisson with rate $\delta_j(t) = \mathrm{E}[\alpha_j(t-\tau_j)]$, where $\tau_j \sim \exp(\eta_{0j})$ is a transition time from station 0 to station j and $\alpha_j(t) = p_{0j}\alpha(t)$.

Two important special cases are:

1. Arrivals started at the distant past:

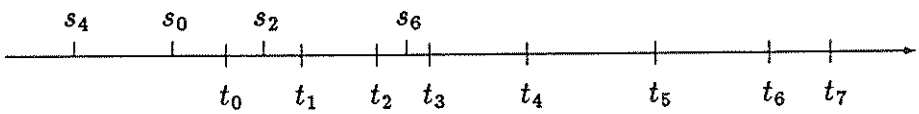
$$\alpha(t) = \lambda,\ t > -\infty \Rightarrow \delta_j(t) = \lambda_j \stackrel{\text{def}}{=} \lambda p_{0j},\ t > -\infty,\ j \in \mathcal{D}.$$

2. Arrivals start at time 0:

$$\alpha(t) = \lambda,\ t \geq 0 \Rightarrow \delta_j(t) = \lambda_j(1 - e^{-\eta_{0j}t}),\ t > 0, j \in \mathcal{D}.$$

As in all problems of interpolation, we consider a separate busy period at a specified station. Now, the transition start-times are known for internal customers only (see Figure 1.5). The representation of the relevant information is identical to that of the previous section.

Figure 1.5. Exponential Transitions. Unknown External Arrivals. Example of Busy Period.



Customers 0, 2, 4 and 6 are internal.

Customers 1, 3 and 5 are external, whose transition-start times are unknown.

Formulation as HMM We shall define the Hidden Markov Model and estimate the queue using the algorithm that was presented in Subsection 1.3. All definitions and calculations are preserved except calculations of transition probabilities between the admissible states. The Markov property of the process $\tilde{A} = \{\tilde{A}(t_0), \ldots, \tilde{A}(t_n)\}$ follows from the memoryless property of the exponential transition times and the independent increments of the Poisson process.

Define cumulative arrival rate $\Delta_j(t) = \int_0^t \delta_j(t)dt$ and suppose that our registrations $t_i \geq 0$ for all i. Let $\mathcal{I}(k,l)$ denote the subset of internal customers in the set $\{k, k+1, \ldots, l\}$ and let $\mathcal{E}(k,l)$ be the number of external customers in this set. Then for $0 \leq i \leq n-1$, $k < n$, $k+l \leq n$,

$$\begin{aligned}
\mathrm{P}\{\tilde{A}(t_{i+1}) = N / \tilde{A}(t_i) = N\} &= \exp\{-(\Delta(t_{i+1}) - \Delta(t_i))\}. \\
\mathrm{P}\{\tilde{A}(t_{i+1}) = K / \tilde{A}(t_i) = K\} &= \exp\{-(\Delta(t_{i+1}) - \Delta(t_i))\} \quad \times \\
&\quad \times \prod_{j \in \mathcal{I}(k+1,n)} \exp\{-\gamma_j (t_{i+1} - \max(t_i, s_j))_+\} \\
\mathrm{P}\{\tilde{A}(t_{i+1}) = K + L / \tilde{A}(t_i) = K\} &= 0, \quad \text{if } \exists s_m \in \mathcal{I}(k+1, k+l) \ni s_m > t_{i+1}.
\end{aligned}$$

Otherwise,

$$\begin{aligned}
&\mathrm{P}\{\tilde{A}(t_{i+1}) = K + L / \tilde{A}(t_i) = K\} = \\
&= \prod_{j \in \mathcal{I}(k+l+1,n)} \exp\{-\gamma_j (t_{i+1} - \max(t_i, s_j))_+\} \quad \times \\
&\times \prod_{j \in \mathcal{I}(k+1,k+l)} (1 - \exp\{-\gamma_j (t_{i+1} - \max(t_i, s_j))\}) \quad \times \\
&\times \exp\{-(\Delta(t_{i+1}) - \Delta(t_i))\} \frac{(\Delta(t_{i+1}) - \Delta(t_i))^{\mathcal{E}(k+1,k+l)}}{\mathcal{E}(k+1, k+l)!} \quad \times \\
&\times \mathrm{P}\{A_{k+1} < A_{k+2} < \ldots < A_{k+l} \ / \ t_i < A_{k+1} \leq t_{i+1}, \ldots, t_i < A_{k+l} \leq t_{i+1}\}.
\end{aligned}$$

The last probability is equal to

$$\int_{t_i}^{t_{i+1}} f_1(y_1) \int_{y_1}^{t_{i+1}} f_2(y_2) \ldots \int_{y_{n-1}}^{t_{i+1}} f_n(y_n) dy_n \, dy_{n-1} \ldots dy_1, \tag{1.11}$$

where f_j are densities of A_j. If A_j is an internal arrival, then

$$A_j \sim \text{ truncexp } (\gamma_j; \max(t_i, s_j), t_{i+1})$$

(see Subsection 1.5). Now consider external arrivals. From Hall [9], the time of an external arrival, which took place in $[t_i, t_{i+1}]$, has density

$$f_j(t) = \frac{\delta(t)}{\Delta(t_{i+1}) - \Delta(t_i)}, \quad t \in [t_i, t_{i+1}]. \tag{1.12}$$

Therefore in (1.11) we deal with order statistics of random variable with such a density. Thus, for external f_j we can substitute into (1.11) the density function from (1.12) and then multiply the result by $\mathcal{E}(k+1, k+l)!$.

Discussion

- For the first special case of constant external arrival rate, in which arrivals start in the distant past, the distribution of external arrivals to a station (and, therefore, the queue distribution) is independent of λ. Indeed, external arrivals are order statistics of the uniform distribution on $[t_0, t_n]$, conditioned on the event:

$$\{A_1 \le t_1,\ A_2 \le t_2, \ldots, A_n \le t_n\} \bigcap \{t_0 < A_1 < A_2 < \ldots A_n\},$$

  whose probability is not a function of λ.

- Consider the second special case of the constant external arrival rate, in which arrivals start at 0. If we make the time change

$$t \to \int_0^t (1 - e^{-\eta_{0j} s})ds = t - 1 + \frac{e^{-\eta_{0j} t}}{t},$$

  the external arrivals will be homogeneous Poisson and will have conditional distribution of uniform order statistics in the new scale. Note that the time change is independent of λ, therefore, the queue estimate is again independent of the arrival rate.

- Unfortunately, we have not been able to derive recursive algorithm for the calculation of formula (1.11), even for the special cases. The integral, however, can always be computed approximately by numerical methods.

Example. We demonstrate our algorithm with a simple example. Consider a short busy period with $n = 2$, whose registrations are denoted by $t_0 = 0, t_1$ and t_2. Suppose that the first customer (which starts service at t_1) is internal, his transition start time $s_1 < 0$ and the transition rate is γ. The second customer is external. We assume that external arrivals to the station constitute a Poisson process with a rate λ (special case 1, arrivals to the entrance of the network start at the distant past).

Let us calculate

$$\mathrm{P}\{A(t_1) = 1 \ / \ A_1 \le t_1,\ A_2 \le t_2;\ t_0 < A_1 < A_2\}$$

using the algorithm in Subsection 1.3.

The first step of the algorithm provides us with the transition probabilities:

$$\begin{aligned}
\tilde{P}_1(0,1) &= (1 - e^{-\gamma t_1})e^{-\lambda t_1} \\
\tilde{P}_1(0,2) &= (1 - e^{-\gamma t_1})\lambda t_1 e^{-\lambda t_1} \int_0^t \frac{\gamma e^{-\gamma s}}{1 - e^{-\gamma t_1}} \frac{t_1 - s}{t_1} ds = \\
&= (1 - e^{-\gamma t_1})\lambda t_1 e^{-\lambda t_1} (1 - \frac{1}{\gamma t_1} + \frac{e^{-\gamma t_1}}{1 - e^{-\gamma t_1}}) \\
\tilde{P}_2(1,2) &= \lambda(t_2 - t_1)e^{-\lambda(t_2 - t_1)} \\
\tilde{P}_2(2,2) &= e^{-\lambda(t_2 - t_1)}
\end{aligned}$$

The second step gives $W^{0,1} = \tilde{P}_1$ and $W^{1,2} = \tilde{P}_2$. Finally (step 3),

$$\begin{aligned} \mathrm{P}\{A(t_1) = 1 \,/\, A_1 \le t_1,\ A_2 \le t_2;\ t_0 < A_1 < A_2\} &= \frac{\tilde{P}_1(0,1)\tilde{P}_2(1,2)}{\tilde{P}_1(0,1)\tilde{P}_2(1,2) + \tilde{P}_1(0,2)\tilde{P}_2(2,2)} \\ &= \frac{t_2 - t_1}{t_2 - t_1(\frac{1}{\gamma t_1} - \frac{e^{-\gamma t_1}}{1-e^{-\gamma t_1}})}. \end{aligned} \tag{1.13}$$

Note, that

$$\mathrm{P}\{A(t_1) = 1 \,/\, t_0 \le A_1 \le t_1,\ t_0 \le A_2 \le t_2\} = \frac{t_2 - t_1}{t_2}.$$

It is straightforward to check that this expression is always larger then the probability in (1.13) (the condition $A_1 < A_2$ changes the distribution of A_2).

1.5 Truncated exponential random variables.

Notation. We define a truncated exponential random variable with rate λ and domain $[a, b]$ as a random variable ζ with density:

$$f(y) = \frac{\lambda e^{-\lambda y}}{e^{-\lambda a} - e^{-\lambda b}}.$$

It is denoted by $\zeta \sim \mathrm{truncexp}(\lambda; a, b)$ and its interpretation is that of an exponential random variable $\zeta_0 \sim \exp\lambda$, first shifted by a and then conditioned to have values in $[a, b]$.

Several facts about truncated exponential random variables are needed for the calculation of transition probabilities of the Hidden Markov Models, which are used in the case of exponential transition times between the stations.

Specifically, suppose that $\zeta_1, \zeta_2, \ldots, \zeta_n$ are independent truncated exponential random variables. We would like to calculate the probability $\mathrm{P}\{\zeta_1 < \zeta_2 < \ldots < \zeta_n\}$.

Case 1.

Let $\zeta_i \sim \mathrm{truncexp}(\lambda_i; 0, t)$, $1 \le i \le n$. This is the easiest case, considered mainly for a "warm-up".

Denote $P_\Lambda = \mathrm{P}\{\zeta_1 < \zeta_2 < \ldots < \zeta_n\}$, where $\Lambda = (\lambda_1, \ldots, \lambda_n)$. Then

$$\begin{aligned} P_\Lambda &= \int_0^t f_1(y_1) \int_{y_1}^t f_2(y_2) \ldots \int_{y_{n-1}}^t f_n(y_n) dy_n\, dy_{n-1} \ldots dy_1 = \\ &= \frac{1}{\prod_{i=1}^n (1 - e^{-\lambda_i t})} \int_0^t \lambda_1 e^{-\lambda_1 y_1} \ldots \int_{y_{n-2}}^t \lambda_{n-1} e^{-\lambda_{n-1} y_{n-1}} (e^{-\lambda_n y_{n-1}} - e^{-\lambda_n t}) dy_{n-1} \ldots dy_1 = \\ &= \frac{1}{1 - e^{-\lambda_n t}} \left[\frac{1}{\prod_{i=1}^{n-1} (1 - e^{-\lambda_i t})} \int_0^t \lambda_1 e^{-\lambda_1 y_1} \ldots \int_{y_{n-2}}^t \lambda_{n-1} e^{-(\lambda_{n-1} + \lambda_n) y_{n-1}} dy_{n-1} \ldots dy_1 - \right. \\ &- \left. e^{-\lambda_n t} \int_0^t \lambda_1 e^{-\lambda_1 y_1} \ldots \int_{y_{n-2}}^t \lambda_{n-1} e^{-\lambda_{n-1} y_{n-1}} dy_{n-1} \ldots dy_1 \right] = \\ &= \frac{1}{1 - e^{-\lambda_n t}} \left[\frac{1 - e^{-(\lambda_{n-1} + \lambda_n) t}}{1 - e^{-\lambda_{n-1} t}} \frac{\lambda_{n-1}}{\lambda_{n-1} + \lambda_n} P_{\Lambda_2} - e^{-\lambda_n t} P_{\Lambda_1} \right], \end{aligned}$$

where $\Lambda_1, \Lambda_2 \in R^{n-1}$ and

$$\Lambda_1 = (\lambda_1, \ldots, \lambda_{n-2}, \lambda_{n-1}),$$

$$\Lambda_2 = (\lambda_1, \ldots, \lambda_{n-2}, \lambda_{n-1} + \lambda_n).$$

So we have derived a recursive formula for the computation of P_Λ.

Case 2.

Let $\zeta_i \sim \text{truncexp}(\lambda_i; x_i, t)$, $1 \le i \le n$ and $x_1 \le x_2 \le \ldots \le x_n < t$.

Denote $P_\Lambda^X(t) = \Pr(\zeta_1 < \zeta_2 < \ldots < \zeta_n)$, where $\Lambda = (\lambda_1, \ldots, \lambda_n)$ and $X = (x_1, \ldots, x_n)$. Then

$$\begin{aligned} P_\Lambda^X(t) &= \int_{x_1}^t f_1(y_1) \int_{\max(y_1, x_2)}^t f_2(y_2) \ldots \int_{\max(y_{n-1}, x_n)}^t f_n(y_n) dy_n \, dy_{n-1} \ldots dy_1 = \\ &= \frac{1}{\prod_{i=1}^n (e^{-\lambda_i x_i} - e^{-\lambda_i t})} \int_{x_1}^t \lambda_1 e^{-\lambda_1 y_1} \ldots \int_{\max(y_{n-2}, x_{n-1})}^t \lambda_{n-1} e^{-\lambda_{n-1} y_{n-1}} \times \\ &\times [e^{-\lambda_n \max(y_{n-1}, x_n)} - e^{-\lambda_n t}] dy_{n-1} \ldots dy_1 = \\ &= S_1 - S_2. \end{aligned}$$

Here S_1 and S_2 correspond each to a term in the square brackets. Thus,

$$S_2 = \frac{e^{-\lambda_n t}}{e^{-\lambda_n x_n} - e^{-\lambda_n t}} P_{\Lambda_1}^{X_1}(t),$$

$$\vec{\Lambda_1} = (\lambda_1, \ldots, \lambda_{n-1}),$$

$$\vec{X_1} = (x_1, \ldots, x_{n-1}).$$

Some more work is required with

$$\begin{aligned} S_1 &= \frac{1}{\prod_{i=1}^{n-1} (e^{-\lambda_i x_i} - e^{-\lambda_i t})} \int_{x_1}^t \lambda_1 e^{-\lambda_1 y_1} \ldots \\ &\quad \int_{\max(y_{n-2}, x_{n-1})}^t \lambda_{n-1} e^{-\lambda_{n-1} y_{n-1} - \lambda_n \max(y_{n-1}, x_n)} dy_{n-1} \ldots dy_1 = \\ &= \frac{1}{\prod_{i=1}^{n-1} (e^{-\lambda_i x_i} - e^{-\lambda_i t})} \int_{x_1}^t \lambda_1 e^{-\lambda_1 y_1} \ldots \\ &\quad (\int_{\max(y_{n-2}, x_{n-1})}^{\max(y_{n-2}, x_n)} \ldots dy_{n-1} + \int_{\max(y_{n-2}, x_n)}^t \ldots dy_{n-1}) dy_{n-2} \ldots dy_1 = \\ &= S_3 + S_4 \end{aligned}$$

First we deal with S_4. Observe that $y_{n-1} \ge x_n$ on the domain of the integral.

$$\begin{aligned} S_4 &= \frac{1}{\prod_{i=1}^{n-1} (e^{-\lambda_i x_i} - e^{-\lambda_i t})} \int_{x_1}^t \lambda_1 e^{-\lambda_1 y_1} \ldots \int_{\max(y_{n-2}, x_n)}^t \lambda_{n-1} e^{-(\lambda_{n-1} + \lambda_n) y_{n-1}} dy_{n-1} \ldots dy_1 = \\ &= \frac{\lambda_{n-1}}{\lambda_{n-1} + \lambda_n} \frac{e^{-(\lambda_{n-1} + \lambda_n) x_n} - e^{-(\lambda_{n-1} + \lambda_n) t}}{e^{-\lambda_{n-1} x_{n-1}} - e^{-\lambda_{n-1} t}} P_{\Lambda_2}^{X_2}(t), \end{aligned}$$

where

$$\vec{\Lambda_2} = (\lambda_1, \ldots, \lambda_{n-1} + \lambda_n),$$
$$\vec{X_2} = (x_1, \ldots, x_{n-2}, x_n).$$

Before out last calculation we note that $\int_{\max(y_{n-2},x_{n-1})}^{\max(y_{n-2},x_n)}$ is non-zero only if $x_n > y_{n-2}$. The domain of the multi-dimensional integral is contained in the set $y_1 \le y_2 \le \ldots \le y_{n-1}$, therefore in our special case it is a subset of $y_i \le x_n$, $1 \le i \le n-1$. Back to S_3

$$\begin{aligned} S_3 &= \frac{1}{\prod_{i=1}^{n-1}(e^{-\lambda_i x_i} - e^{-\lambda_i t})} \int_{x_1}^{t} \lambda_1 e^{-\lambda_1 y_1} \int_{\max(y_1,x_2)}^{t} \lambda_2 e^{-\lambda_2 y_2} \ldots \\ &\quad \int_{\max(y_{n-2},x_{n-1})}^{x_n} \lambda_{n-1} e^{-\lambda_{n-1} y_{n-1} - \lambda_n x_n} dy_{n-1} \ldots dy_1 = \\ &= \frac{1}{\prod_{i=1}^{n-1}(e^{-\lambda_i x_i} - e^{-\lambda_i t})} \int_{x_1}^{x_n} \lambda_1 e^{-\lambda_1 y_1} \int_{\max(y_1,x_2)}^{x_n} \lambda_2 e^{-\lambda_2 y_2} \ldots \\ &\quad \int_{\max(y_{n-2},x_{n-1})}^{x_n} \lambda_{n-1} e^{-\lambda_{n-1} y_{n-1} - \lambda_n x_n} dy_{n-1} \ldots dy_1 = \\ &= e^{-\lambda_n x_n} \frac{\prod_{i=1}^{n-1}(e^{-\lambda_i x_i} - e^{-\lambda_i x_n})}{\prod_{i=1}^{n-1}(e^{-\lambda_i x_i} - e^{-\lambda_i t})} P_{\Lambda_1}^{X_1}(x_n) \end{aligned}$$

Now we summarize Case 2.

Proposition 1.1

$$\begin{aligned} P_{\Lambda}^{X}(t) &= \frac{1}{e^{-\lambda_n x_n} - e^{-\lambda_n t}} \left[e^{-\lambda_n x_n} \frac{\prod_{i=1}^{n-1}(e^{-\lambda_i x_i} - e^{-\lambda_i x_n})}{\prod_{i=1}^{n-1}(e^{-\lambda_i x_i} - e^{-\lambda_i t})} P_{\Lambda_1}^{X_1}(x_n) + \right. \\ &+ \left. \frac{\lambda_{n-1}}{\lambda_{n-1} + \lambda_n} \frac{e^{-(\lambda_{n-1}+\lambda_n)x_n} - e^{-(\lambda_{n-1}+\lambda_n)t}}{e^{-\lambda_{n-1} x_{n-1}} - e^{-\lambda_{n-1} t}} P_{\Lambda_2}^{X_2}(t) - e^{-\lambda_n t} P_{\Lambda_1}^{X_1}(t) \right], \quad (1.14) \end{aligned}$$

where

$$\begin{aligned} \Lambda_1 &= (\lambda_1, \ldots, \lambda_{n-1}), \\ X_1 &= (x_1, \ldots, x_{n-1}), \\ \Lambda_2 &= (\lambda_1, \ldots, \lambda_{n-1} + \lambda_n), \\ X_2 &= (x_1, \ldots, x_{n-2}, x_n). \end{aligned}$$

Case 3.

Now consider the final problem, directly connected with the calculations of the transitional probabilities in the Hidden Markov Models.

Let $\zeta_i \sim \text{truncexp}(\lambda_i; x_i, t)$, $1 \le i \le n$, where $0 \le x_i \le t$. Let

$$r_i = \max_{j \le i} x_j \quad \text{(see Figure 1.6) and} \quad R = (r_1, r_2, \ldots, r_n).$$

Figure 1.6. Illustration to Proposition 1.2.

Denote $P_\Lambda^X(t) = \mathrm{P}\{\zeta_1 < \zeta_2 < \ldots < \zeta_n\}$, where $\vec{\Lambda} = (\lambda_1, \ldots, \lambda_n)$ and $\vec{X} = (x_1, \ldots, x_n)$.

Proposition 1.2

$$P_\Lambda^X(t) = P_\Lambda^R(t) \prod_{i=1}^{n} \frac{e^{-\lambda_i r_i} - e^{-\lambda_i t}}{e^{-\lambda_i x_i} - e^{-\lambda_i t}} \tag{1.15}$$

Proof.

Introduce the event $A = \{\zeta_i \geq r_i,\ 1 \leq i \leq n\}$.

Then

$$\begin{aligned} P_\Lambda^X(t) &= \mathrm{P}\{\zeta_1 < \zeta_2 < \ldots < \zeta_n\} = \mathrm{P}\{\zeta_1 < \zeta_2 < \ldots < \zeta_n; A\} = \\ &= \mathrm{P}\{\zeta_1 < \zeta_2 < \ldots < \zeta_n / A\}\mathrm{P}\{A\} \end{aligned}$$

. Conditioning on A, $\zeta_1, \ldots, \zeta_n$ are truncated exponentials on $[r_i, t]$. Hence the first probability of the last formula is equal to $P_\Lambda^R(t)$, $\mathrm{P}\{A\}$ is easily computed and we get (1.15).

Remark. Consider Figure 1.6 again. In order to calculate the required probability, we first transform our x_i's to r_i's to get them in an ascending order and then use the algorithm of Case 2.

Number of operations.

Throughout the recursive calculations of Case 1, we need to compute P_Λ, where

$$\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_l, \lambda_{l+1} + \ldots + \lambda_k), \quad 0 < l < k \leq n.$$

There exists

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$$

different P_Λ of this type.

In Case 2 we compute $P_\Lambda^X(\tilde{t})$, where

$$\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_l, \lambda_{l+1} + \ldots + \lambda_k), \quad X = (x_1, x_2, \ldots, x_l, x_k), \quad \tilde{t} = x_m \text{ or } \tilde{t} = t, \ 0 < l < k < m \leq n.$$

Now

$$\sum_{k=1}^{n} \frac{k(k+1)}{2} = \frac{n(n+1)(2n+1)}{3}$$

different $P_\Lambda^X(\tilde{t})$ should be calculated.

Case 3 involves the same n-order of operations as Case 2.

2 Representations of an Observed Event.

As already mentioned in Section 1, our results and algorithms for the Interpolation case were not treated as carefully as they could have been. For example, we have worked with $t_0, t_1, \ldots, t_n$, which are fixed realizations of random registration times, ignoring the distinction between random variables and their realizations. Also, information that is relevant for queue inference was identified intuitively.

We have used, in essence, the terminology and approaches of the previous papers by Larson and Daley & Servi on single-station models [12], [3], [4]. It seemed inappropriate to complicate considerably our models and methods in order to derive rigorous proofs for facts that had been intuitively clear. However, the cases of Real-Time Estimation and General Interpolation are more difficult, in particular the structure of the relevant information is less intuitive, so formal methods will now be developed for these problems.

Representations of the observed event play an important role in our approach. They serve as a mathematical framework within which Interpolation and Real-Time problems are conveniently formulated and solved.

Three types of representations are introduced. The first, via *observations*, represents the real data that is actually observed. The third, via *stochastic components* (namely external arrivals, service times, switches between stations and transition times,) is helpful for our analysis: conditioning on the observed event, represented in this form, identifies the part of the event that is relevant for queue estimation. The second representation, via *arrivals* to stations helps to ascertain the equivalence between the two former representations.

Our approach is illustrated on the simplest example: we reconsider the basic assumptions of Larson [12] and rederive his representation of the observed event. We begin with service starts and terminations (observations) and conclude with external arrival times (stochastic components). Due to the model simplicity, the intermediate representation via arrivals is merely implicit.

2.1 M/G/1. Rederiving Larson's QIE.

Larson's model consists of a single-server station with Poisson arrivals, general service times and FCFS discipline.

Stochastic components.

- $A_0, A_1, \ldots, A_k, A_{k+1}, \ldots$ are the Poisson arrival times.
- $\xi_0, \xi_1, \ldots, \xi_k, \xi_{k+1}, \ldots$ are the corresponding service times.

Larson assumed that the service times are *completely general*: they may be dependent with an arbitrary general joint distribution. The following assumption must, however, be added: *service times are independent of arrival times.*

As before, service registrations of a specific busy period are denoted by $t_0 = A_0, t_1, \ldots, t_n, t_{n+1}$ and we need to estimate the queue length $Q(t)$ (or the cumulative number of arrivals $A(t)$) over the given busy period $[t_0, t_{n+1}]$, using the available data.

Let the random variable N stand for the length (number of service starts) of our busy period. The random variables T_i, $0 \leq i \leq N-1$, denote the start time of the i-th service and T_N denotes the last service termination of the busy period (the end of the busy period).

Let the event E stand for the information available at the end of the busy period. We introduce two representations of this event.

I. Representation of E via Observations.

$$E_o = \{N = n+1;\ T_0 = t_0, T_1 = t_1, \ldots, T_n = t_n, T_{n+1} = t_{n+1}\}.$$

II. Representation of E via Stochastic Components.

$$E_s = \{A_0 = t_0, A_1 \leq t_1, \ldots, A_n \leq t_n, A_{n+1} > t_{n+1};$$
$$\xi_0 = t_1 - t_0, \xi_1 = t_2 - t_1, \ldots, \xi_n = t_{n+1} - t_n\}.$$

Equivalence of representations I and II: $\mathbf{E_0 = E_s}$.

$E_o \Rightarrow E_s$. The event $\{A_0 = t_0\}$ is identical to $\{T_0 = t_0\}$. The event $\{A_i \leq t_i\}$, $1 \leq i \leq n$, prevails since, otherwise, $T_i = t_i$ could not be true (the i-th service could not have started since the corresponding arrival did not take place yet). The event $\{N = n+1\}$ implies $\{A_{n+1} > t_{n+1}\}$. Finally, $\{\xi_i = t_{i+1} - t_i\}$ follows from $T_{i+1} = t_{i+1}$ and $T_i = t_i$.

$E_s \Rightarrow E_o$. The event $\{T_i = t_i\}$, $1 \leq i \leq n$, follows from the events $\{\xi_j = t_{j+1} - t_j,\ 0 \leq j \leq i-1\}$, $\{A_j \leq t_j,\ 1 \leq j \leq i\}$ and $\{A_0 = t_0\}$. Specifically, the first event implies that potential service-start times are $t_0, t_1, \ldots, t_i$ (provided server starts services immediately following service terminations). The second event implies that arrival times precede potential service starts, therefore, actual service-start times coincide with potential service-start times. The event $\{T_{n+1} = t_{n+1}\}$ follows from $\{\xi_n = t_{n+1} - t_n\}$. Finally, $\{A_{n+1} > t_{n+1}\}$ gives $\{N = n+1\}$.

Now, if we condition cumulative number of arrivals on the available information:

$$\begin{aligned}
\mathrm{P}\{A(t) = k/E_o\} &= \mathrm{P}\{A(t) = k/E_s\} \\
&= \mathrm{P}\{A_k \leq t, A_{k+1} > t/A_0 = t_0, A_1 \leq t_1, \ldots, A_n \leq t_n, A_{n+1} > t_{n+1}; \\
&\quad \xi_0 = t_1 - t_0, \xi_1 = t_2 - t_1, \ldots, \xi_n = t_{n+1} - t_n\} \qquad (2.16) \\
&= \mathrm{P}\{A_k \leq t, A_{k+1} > t/A_0 = t_0, A_1 \leq t_1, \ldots, A_n \leq t_n, A_{n+1} > t_{n+1}\} \qquad (2.17) \\
&= \mathrm{P}\{A_k \leq t, A_{k+1} > t/A_0 = t_0, A_1 \leq t_1, \ldots, A_n \leq t_n, A(t_{n+1}) = n\} \qquad (2.18)
\end{aligned}$$

Formula (2.17) follows from the independence between arrivals and services. It is natural to refer to the event

$$E_r = \{A_0 = t_0, A_1 \leq t_1, A_2 \leq t_2, \ldots, A_n \leq t_n, A_{n+1} > t_{n+1}\}$$

as the *relevant information*, relevant in the sense that other components of the observed event (service times, in our case) are not necessary for queue estimation.

The condition in (2.18) implies that n Poisson arrivals took place in $(0, t_{n+1}]$. Using the well-known order-statistics property of the Poisson process, we see that the joint distribution of $A_1, A_2, \ldots, A_n$, given the available information, is that of order statistics from the uniform distribution on $[0, t_{n+1}]$, conditioned on the event

$$\{A_1 \leq t_1, A_2 \leq t_2, \ldots, A_n \leq t_n\}.$$

The expression in (2.18) is equal to:

$$\frac{\mathrm{P}\{A_k \leq t, A_{k+1} > t, A_1 \leq t_1, \ldots, A_n \leq t_n / A_0 = t_0, A(t_{n+1}) = n\}}{\mathrm{P}\{A_1 \leq t_1, \ldots, A_n \leq t_n / A_0 = t_0, A(t_{n+1}) = n\}} \tag{2.19}$$

Our estimate do not depend on λ since the numerator and the denominator of (2.19) do not depend on λ from the order-statistics property.

Remark. In this work we sometimes condition on events whose probabilities can be equal to zero (see formulae (2.16) and (2.19)). For example, the denominator of (2.19) can be represented as

$$\lim_{h \to 0} \mathrm{P}\{A_1 \leq t_1, \ldots, A_n \leq t_n / A_0 \in [t_0 - h, t_0], A(t_{n+1}) = n\}.$$

In general, if we condition on event of the type $\{X_1 = x_1, \ldots, X_n = x_n\}$ and assume that the random vector $(X_1, \ldots, X_n)$ has a positive probability in the neighborhood of $(x_1, \ldots, x_n)$, a similar definition can be applied.

2.2 Representations via Registration Times.

In Section 3.2 of Chapter 1 several different approaches to the observed information up to time t were introduced (all were based on the real data). Here we shall elaborate on one of the approaches, that which is based on busy periods. It is technically simpler to work with the observed event up to time t, which is denoted by E_t, instead of the observed information $\mathcal{F}_t$. (E_t is an element of $\mathcal{F}_t$.)

For simplicity of notation (which is somewhat cumbersome anyway), we shall assume everywhere throughout Sections 2 and 3 of Chapter 2 that only a single server is present at every station. However, all our results remain valid for the multi-server case.

Recall that we always register start and termination times of services with ID's of customers and sometimes also arrival times and ID's of external customers.

At first, a notation for registration times will be introduced. Let N_k^j denote the number of customers that were served during the k-th busy period at station j (without counting the customer that starts the busy period, since customers at the k-th busy period are numbered by $0, 1, 2, \ldots, N_k^j$). Then T_{ki}^j, $0 \leq i \leq N_k^j + 1$, is the time of the i-th registration

during busy period number k at station j. In particular, T^j_{k0} and T^j_{k,N^j_k+1} stand for the service start of customer 0 and for the service termination of customer N^j_k respectively.

Suppose that we observe the network during the time interval $[0,t]$. Let $B_j(t)$, $j \in \mathcal{D}$, denote the number of busy periods, which started at station j up to time t. Let $\mathcal{D}_b(t)$ stand for the *set of busy stations* at time t (i.e. stations whose servers are busy). For $j \in \mathcal{D}_b(t)$ introduce $M_j(t)$ to be the number of customers that started service before t during the current busy period. (Here, again, we do not count customer 0, which started this busy period.)

Finally, introduce a notation for external arrivals: let $N^0(t)$ denotes the number of external arrivals up to t and T^0_i denotes the time of the i-th external arrival to the network.

We now develop a theoretical framework for describing the information embodied in the ID's of customers. Using ID's, it is possible to obtain the trace of a customer (the sequence of stations visited up to time t) and the times of registrations up to t. To this end, introduce two routing operators. The *Routing operator* $\mathcal{R}$ locates the next service start of a customer, the *Backward Routing operator* $\mathcal{R}^{-1}$ locates his previous service termination.

Definition of the Routing Operator $\mathcal{R}$. The domain of $\mathcal{R}$ consists of triplets of natural numbers (n,m,l) and pairs of the type $(0,l)$. The value of the routing operator is given by

$$\mathcal{R}(n,m,l) = (j,k,i) \quad (\text{ or } \mathcal{R}(0,l) = (j,k,i)\,) \tag{2.20}$$

if the customer, which terminated service at time T^n_{ml} at station n (or entered the network at T^0_l), starts his next service at time T^j_{ki} at station j. If that customer leaves the network, let

$$\mathcal{R}(n,m,l) = 0 \quad (\text{or } \mathcal{R}(0,l) = 0\,).$$

We introduce also a special notation for the station to where the customer switches: if formula (2.20) prevails,

$$R^n_{ml} = j \;\; (\text{ or } R^0_l = j\,).$$

Definition of Backward Routing operator $\mathcal{R}^{-1}$ The domain of $\mathcal{R}^{-1}$ is formed by triplets of natural numbers. It takes the value

$$\mathcal{R}^{-1}(j,k,i) = (n,m,l)$$

if the customer, which starts service at time T^j_{ki}, terminated his previous service at T^n_{ml}. If the service that starts at time T^j_{ki} was the first one for that customer let

$$\mathcal{R}^{-1}(j,k,i) = (0,l),$$

where l is the index of the customer's registration at station 0.

In addition, auxiliary operators for the previous station and the previous service termination time are introduced:

$$\mathcal{R}_S^{-1}(j,k,i) = n, \qquad \mathcal{R}_T^{-1}(j,k,i) = T_{ml}^n.$$

Representation 1 of E_t via Observations.

Known External Arrivals.

Roughly speaking, the information embodied in ID's is identical to the information about values of the Backward Routing operator. Indeed, an ID of a customer enables the derivation of all his previous registrations. But it is possible to get that same information recursively from the Backward Routing operator.

The observed event E_t can be represented as the intersection of the eight events 1.1-1.8 below. Here and later we adopt a uniform style of event representations: small letters will be used for realizations of random variables; a verbal description of an event is followed by formulas; the main statements are printed in bold.

1.1 The number of busy periods at every station is known.
$\mathbf{B_j(t) = b_j(t)}, \quad j \in \mathcal{D}.$

1.2 The set of busy stations is known.
$\mathcal{D}_\mathbf{b}(\mathbf{t}) = \mathbf{d_b(t)}.$

1.3 The number of registrations is known for all complete busy periods.
$\mathbf{N_k^j = n_k^j}, \quad j \in \mathcal{D},$
$1 \le k \le b_j(t)$ for $j \notin d_b(t)$,
$1 \le k \le b_j(t) - 1$ for $j \in d_b(t)$.

1.4 The number of registrations until t for incomplete busy periods is known.
$\mathbf{M_j(t) = m_j(t)}, \quad j \in d_b(t).$

1.5 The registrations are known.
$\mathbf{T_{ki}^j = t_{ki}^j}, \quad j \in \mathcal{D},\ 1 \le k \le b_j(t),$
$0 \le i \le m_j(t)$ for $j \in d_b(t),\ k = b_j(t),$
$0 \le i \le n_k^j + 1,$ otherwise.

1.6 Complete information about the backward routing.
$\mathcal{R}^{-1}(\mathbf{j,k,i}) = (\mathbf{n,m,l})$ **or** $\mathcal{R}^{-1}(\mathbf{j,k,i}) = (\mathbf{0,l}),$ for $(j,k,i) \ni t_{ki}^j \le t.$

1.7 The number of external arrivals is known.
$\mathbf{N^0(t) = n^0(t)}.$

1.8 The registrations at station 0 are known.
$\mathbf{T_i^0 = t_i^0},\ 1 \le i \le n^0(t).$

Representation 1' of $\mathbf{E_t}$ via Observations.
Unknown External Arrivals.

Here the information about the backward routing can be incomplete. If the ID of a registered customer has never appeared earlier, we know that he switched to the current station from station 0 but no information is available about his arrival time to station 0.

Events 1.1'-1.5' of Representation 1' coincide with their counterparts 1.1-1.5 from the previous case. The last two events from Representation 1 are omitted and event 1.6 is changed in the following way.

1.6' Partial information about the backward routing.
$\mathcal{R}^{-1}(\mathbf{j},\mathbf{k},\mathbf{i}) = (\mathbf{n},\mathbf{m},\mathbf{l})$ **or** $\mathcal{R}_{\mathbf{S}}^{-1}(\mathbf{j},\mathbf{k},\mathbf{i}) = \mathbf{0}$, for $(j,k,i) \ni t_{ki}^{j} \le t$.
($\mathcal{R}_S^{-1}(j,k,i) = 0$ when the ID of a customer was never registered earlier.)

2.3 Representations via Arrival Times.

Recall that t_{ki}^{j} denotes the realization of the i-th registration time from the k-th busy period at station j. Let ξ_{ki}^{j} denote the duration of service which started at t_{ki}^{j}. Let also A_{ki}^{j} stand for the arrival time to station j of the customer, which started service at t_{ki}^{j}. Finally, $\bar{A}_{ml}^{n}$ stands for the arrival time to his next station of the customer that terminated service at t_{ml}^{n}. A connection between the two last definitions is given by:

$$\mathcal{R}(n,m,l) = (j,k,i) \;\Rightarrow\; \bar{A}_{ml}^{n} = A_{ki}^{j}.$$

Consider a service termination time t_{ml}^{n}. It will be called a Visible-Routing service termination time (or VR-time in short) if

$$\exists\,(j,k,i) \ni t_{ki}^{j} \le t,\ \mathcal{R}(n,m,l) = (j,k,i).$$

In words, this customer registered at least once during the time interval $(t_{ml}^{n}, t]$. Otherwise, time t_{ml}^{n} will be called Invisible-Routing (IR-time). A time of service termination is IR-time if either the corresponding customer leaves the network or if his next service beginning takes place after t. We use the notation $(n,m,l) \in V_R(t)$ and $(n,m,l) \in I_R(t)$ for VR and IR-times respectively. The corresponding customers are called **visible customers** and **invisible customers.**

The definition of VR and IR-times can be easily extended to the realizations of times of external arrivals t_i^0.

In order to derive more convenient representation of the observed event we need to "decompose" the Routing Operator $\mathcal{R}(n,m,l) = (j,k,i)$. The knowledge of this operator for VR-times is equivalent to the knowledge of the following events (provided that all registrations $t_{ki}^{j} \le t$ are also known).

- **The switches of the visible customers are known.** (See event 2.6 of the following representation.)
- **Truncation rule.** The visible customers arrive to stations before their services start. (See event 2.4.)
- **FCFS discipline.** The visible customers arrive in the order at which they are served. The invisible customers do not arrive to stations before the visible ones. (See events 2.5 and 2.7.)

The exact formulations of these events are given in the following representation.

Representation 2 of $\mathbf{E_t}$ via Arrival Times. Known External Arrivals.

Here we identify conditions that must be imposed on service times and arrivals (internal and external) in order to get the observed event from Representation 1.

2.1 Services that terminated up to t are known.
$\xi^{\mathbf{j}}_{\mathbf{ki}} = \mathbf{t}^{\mathbf{j}}_{\mathbf{k,i+1}} - \mathbf{t}^{\mathbf{j}}_{\mathbf{ki}}$, $j \in \mathcal{D}$, $1 \le k \le b_j(t)$,
$0 \le i \le m_j(t) - 1$, for $j \in d_b(t)$, $k = b_j(t)$,
$0 \le i \le n^j_k$, otherwise.

2.2 Last service at a busy station have not terminated until t.
For $j \in d_b(t)$ $\xi^{\mathbf{j}}_{\mathbf{b_j(t),m_j(t)}} > \mathbf{t} - \mathbf{t}^{\mathbf{j}}_{\mathbf{b_j(t),m_j(t)}}$.

2.3 The start of a busy period coincides with the first arrival.
For $(n,m,l) \in V_R(t) \ni \mathcal{R}(n,m,l) = (j,k,0)$
$\bar{\mathbf{A}}^{\mathbf{n}}_{\mathbf{ml}} = \mathbf{t}^{\mathbf{j}}_{\mathbf{k0}}$, $j \in \mathcal{D}$, $1 \le k \le b_j(t)$.

2.4 Truncation rule. Arrivals take place before corresponding service starts.
For $(n,m,l) \in V_R(t) \ni \mathcal{R}(n,m,l) = (j,k,i)$, $i \ge 1$
$\bar{\mathbf{A}}^{\mathbf{n}}_{\mathbf{ml}} \le \mathbf{t}^{\mathbf{j}}_{\mathbf{ki}}$, $j \in \mathcal{D}$, $1 \le k \le b_j(t)$, where
$1 \le i \le m_j(t)$, for $j \in d_b(t)$, $k = b_j(t)$,
$1 \le i \le n^j_k$, otherwise.

2.5 FCFS discipline.
For $(n_i,m_i,l_i) \in V_R(t) \ni \mathcal{R}(n_i,m_i,l_i) = (j,k,i)$,
$\mathbf{t}^{\mathbf{j}}_{\mathbf{k0}} \le \bar{\mathbf{A}}^{\mathbf{n_1}}_{\mathbf{m_1,l_1}} \le \bar{\mathbf{A}}^{\mathbf{n_2}}_{\mathbf{m_2,l_2}} \le \ldots \le \bar{\mathbf{A}}^{\mathbf{n_u}}_{\mathbf{m_u,l_u}}$, $j \in \mathcal{D}$, $1 \le k \le b_j(t)$, $1 \le i \le u$, where
$u = m_j(t)$, for $j \in d_b(t)$, $k = b_j(t)$,
$u = n^j_k$, otherwise.

2.6 Complete information about the switches of the visible customers.
For $(n,m,l) \in V_R(t)$
$\mathbf{R}^{\mathbf{n}}_{\mathbf{ml}} = \mathbf{j}$.

2.7 Partial information about the invisible routing.

For $(n,m,l) \in I_R(t)$:

$\mathbf{if\, j} \in \mathbf{d_b(t)}, \quad \mathbf{R^n_{ml} = j} \Rightarrow \mathbf{\bar{A}^n_{ml} \geq A^j_{b_j(t),m_j(t)}},$

$\mathbf{if\, j} \notin \mathbf{d_b(t)}, \quad \mathbf{R^n_{ml} = j} \Rightarrow \mathbf{\bar{A}^n_{ml} > t}.$ (Event $A \Rightarrow B = A^c \bigcup B.$)

2.8 The number of external arrivals is known.

$\mathbf{N^0(t) = n^0(t)}.$

2.9 The times of external arrivals are known.

$\mathbf{A^0_i = t^0_i}, \quad 1 \leq i \leq n^0(t).$

Remark. Events 2.3-2.6 must include also external arrivals $(0,l) \in V_R(t)$. They are treated exactly the same as $(n,m,l) \in V_R(t)$. Event 2.7 is true also for $(0,l) \in I_R(t)$.

Representation 2' of $\mathbf{E_t}$ via Arrival Times. Unknown External Arrivals.

Events 2.1', 2.2', 2.6' and 2.7' coincide with their counterparts 2.1, 2.2, 2.6 and 2.7 from Representation 2. We supplement events 2.3-2.5 with the statements of the type:

2.3' For $(j,k,0) \ni \mathcal{R}_S^{-1}(j,k,0) = 0,$
$\exists\, l \ni \mathcal{R}(0,l) = (j,k,0), \bar{A}^0_l = t^j_{k0}.$

The last two events of Representation 2 are omitted.

Equivalence of Representations 1 and 2.

First we show that the event that constitutes Representation 1 implies Representation 2.

Event 2.1 follows from the definition of T^j_{ki} (times of service starts and service terminations). Events 1.2, 1.4, 1.5 determine event 2.2 (the last service on a busy station have not terminated before t).

Then note that complete information about the values of the Backward Routing operator (event 1.6) is identical to complete information about the values of the Routing operator for the VR-times. Events 2.3-2.5 are implied by the knowledge of $\mathcal{R}(n,m,l)$ for $(n,m,l) \in V_R(t)$ and by the apparent facts written at the headings of these events. Event 2.6 follows from 1.6.

The first statement of event 2.7 means that if an invisible customer switched to a busy station, then his arrival time took place later then the arrival time of the last serviced customer. (This fact follows from the FCFS queue discipline.) The second statement means that if he switched to an idle station, then his arrival there took place after t. (This is the second part of the event).

Finally, events 1.7 and 1.8 coincide with events 2.8 and 2.9.

Now, for the converse, suppose all events from Representation 2 are taking place. Then arrival times and service times uniquely define service starts and terminations from Representation 1. Specifically, 2.1-2.4 and 2.6 allow service starts at t_{ki}^{j}, 2.7 guarantees that no arrivals, which can prevent terminations of busy periods at the proper times, take place during $[0, t]$ and 2.5 implies 1.6 (the proper customers start services at the proper times).

The equivalence of Representations 1' and 2' can be derived similarly.

3 Real-Time Estimation and General Interpolation.

3.1 Stochastic Components of the Network.

The stochastic components of our network were introduced in Subsection 1.1 of Chapter 1. Recall that there are four components:

- external Poisson arrivals to the network;
- service times;
- switches of customers between the stations;
- transition times between the stations (in the case of exponential transitions).

We now briefly summarize the components that are known exactly and those which are necessary to infer, for various special cases. We actually cover all problems: those that were already analyzed (Busy-Period Interpolation) and the new ones that will be analyzed in this section. Note that, in all cases, we have complete information about the service times up to the estimation time t.

Unknown stochastic components.

	Immediate transitions		Exponential transitions	
	External arrivals			
	known	unknown	known	unknown
Busy-Period Interpolation.	All stochastic components are known.	External arrivals.	Transition times.	External arrivals; transition times.
Real-time Estimation. General Interpolation.	Switches.	Switches; external arrivals.	Switches; transition times.	Switches; external arrivals; transition times.

We see that the second row differs from the first one only in the knowledge of customers' switches between the stations.

In Busy-Period Interpolation, queue estimation was carried out separately for every station; customers served over a busy period were all known, and the objective was to infer, if necessary, their arrival times (internal or external). The critical parameter, from a computational point of view, was the length of the busy period, while the number of stations in the network had minor effect.

In contrast, real-time problems require also inference about the set of customers that stand in queue at a specific station. Indeed, the current location of a specific customer, having been served, is often unknown. A basic concept here turns out to be the set of *invisible customers*, namely those customers that possibly wait in some queue at the time of estimation. The "double inference" nature of real-time problems, i.e. the need to infer both the set of customers standing in the queue, and their arrival times increases considerably the computational complexity.

3.2 General Interpolation. The Single-Station Case.

Bertsimas and Servi [2] carried out the analysis of the Real-Time problem in the single-station case. However, the General Interpolation problem (queue path reconstruction during a busy period that has not terminated yet) was not considered in the papers on the subject. The results of the current subsection will be applied in Subsection 3.5 in the network setting.

Consider the following problem formulation for a single station. Assume that the busy period under consideration begins at time 0. The times $t_0 = 0, t_1, \ldots, t_n, \ldots$ denote the realizations of service starts. As before, $A_0 = 0, A_1, \ldots, A_n$ stand for arrivals to the station and $A(t)$ is the cumulative number of arrivals in $(0, t]$. Arrivals to the station constitute a homogeneous Poisson process with rate λ. (The case of the non-homogeneous Poisson input is similar: perform the time change $t \to \Lambda(t) = \int_0^t \lambda(s)ds$, then, apply the results for the homogeneous case with $\lambda = 1$ and, finally, perform the reverse time change.) Introduce the event

$$B_n = \{A(t_1) \geq 1, A(t_2) \geq 2, \ldots, A(t_n) \geq n\} = \{A_1 \leq t_1, A_2 \leq t_2, \ldots, A_n \leq t_n\}.$$

The meaning of event B_n is that the considered busy period has not terminated up to t_n.

Our objective is the recursive calculation of $\mathrm{E}\,[A(t)/B_{n+1}]$, for a fixed $t > 0$. Specifically, we update our estimate in view of a new service completion.

This problem was studied previously in Bertsimas and Servi [2]. They, however, assumed that $t = t_n$ (real-time estimation) or $t > t_n$ (extrapolation), while we present an algorithm for general t and clarify the difference between extrapolation and interpolation $(t \leq t_n)$.

Define the events

$$C_n = \{A(t_1) \geq 1, A(t_2) \geq 2, \ldots, A(t_{n-1}) \geq n-1, A(t_n) = n\}$$

and

$$D_n = C_n \bigcap \{A(t_{n+1}) = n\}.$$

Algorithm 3.1. General Interpolation of a single-station.

— Main formula.

$$\mathrm{E}\,[A(t)/B_{n+1}] = \frac{\mathrm{E}\,[A(t)/B_n]\mathrm{P}\{B_n\} - \mathrm{E}\,[A(t)/D_n]\mathrm{P}\{C_n\}e^{-\lambda(t_{n+1}-t_n)}}{\mathrm{P}\{B_n\} - \mathrm{P}\{C_n\}e^{-\lambda(t_{n+1}-t_n)}}. \tag{3.1}$$

$$\mathrm{E}\,[A(t)/B_0] = \lambda t.$$

— Recursive calculation of $\mathrm{P}\{B_{n+1}\}$.

$$\mathrm{P}\{B_{n+1}\} = \mathrm{P}\{B_n\} - \mathrm{P}\{C_n\}e^{-\lambda(t_{n+1}-t_n)}. \tag{3.2}$$

$$\mathrm{P}\{B_0\} = 0.$$

— Recursive calculation of $\mathrm{P}\{C_{n+1}\}$.

$$\mathrm{P}\{C_{n+1}\} = e^{-\lambda t_{n+1}}\lambda^{n+1} g(t_1, t_2, \ldots, t_n, t_{n+1}), \tag{3.3}$$

where

$$g(t_1, t_2, \ldots, t_n, t_{n+1}) \stackrel{\mathrm{def}}{=} \int_{x_1=0}^{t_1}\int_{x_2=x_1}^{t_2}\cdots\int_{x_n=x_{n-1}}^{t_n}\int_{x_{n+1}=x_n}^{t_{n+1}} dx_{n+1}dx_n\ldots dx_2dx_1 \tag{3.4}$$

and

$$g(t_1, t_2, \ldots, t_n, t_{n+1}) = \sum_{k=0}^{n}(-1)^{n-k}\frac{t_{k+1}^{n+1-k}}{(n+1-k)!}g(t_1, t_2, \ldots, t_k). \tag{3.5}$$

$$\mathrm{P}\{C_0\} = g_0 = 1.$$

— Calculation of $\mathrm{E}\,[A(t)/D_n]$. If $t \geq t_n$ (extrapolation)

$$\mathrm{E}\,[A(t)/D_n] = n + \lambda(t - t_{n+1})_+. \tag{3.6}$$

If $t < t_n$ (interpolation), specifically if $t_i \leq t < t_{i+1}$, $i \leq n-1$:

$$\mathrm{E}\,[A(t)/D_n] = \frac{1}{\mathrm{P}\{C_n\}}e^{-\lambda t_n}\lambda^n[\sum_{k=i}^{n}\frac{t^k}{(k-1)!}\,g(t_{k+1}-t, t_{k+2}-t, \ldots, t_n-t) -$$

$$- \sum_{k=i}^{n} k\,g(t_{k+1}-t, t_{k+2}-t, \ldots, t_n-t)]\sum_{m=0}^{i-1}\frac{(t-t_{m+1})^{k-m}}{(k-m)!}g(t_1, \ldots, t_m). \tag{3.7}$$

Verification of Algorithm 3.1.

Formula (3.2) follows from the definitions of B_n, C_n and D_n, equalities

$$B_{n+1} = B_n \setminus \{C_n \bigcap \{A(t_{n+1}) - A(t_n) = 0\}\} = B_n \setminus D_n,$$

and independence between C_n and $A(t_{n+1}) - A(t_n)$.

Formula (3.3) prevails since

$$\begin{aligned}
\mathrm{P}\{C_{n+1}\} &= \mathrm{P}\{A_1 \le t_1, \ldots, A_n \le t_n, A_{n+1} \le t_{n+1}, A(t_{n+1}) = n+1\} \\
&= \mathrm{P}\{A_1 \le t_1, \ldots, A_n \le t_n, A_{n+1} \le t_{n+1} / A(t_{n+1}) = n+1\} e^{-\lambda t_{n+1}} \frac{(\lambda t_{n+1})^{n+1})}{(n+1)!} \\
&\qquad \text{(conditional joint density of uniform order statistics is } \frac{(n+1)!}{t_{n+1}^{n+1}}) \\
&= e^{-\lambda t_{n+1}} \lambda^{n+1} g(t_1, t_2, \ldots, t_n, t_{n+1}).
\end{aligned}$$

Formula (3.5) is almost identical to formula 42 (page 225) of Bertsimas and Servi [2]. We present a simple proof, by computing multidimensional integrals according to definition (3.4):

$$\begin{aligned}
&2.\ g(t_1, t_2, \ldots, t_n, t_{n+1}) \\
&= \int_{x_1=0}^{t_1} \int_{x_2=x_1}^{t_2} \cdots \int_{x_n=x_{n-1}}^{t_n} (t_{n+1} - x_n) dx_n \ldots dx_1 \\
&= t_{n+1} g(t_1, \ldots, t_n) - \int_{x_1=0}^{t_1} \int_{x_2=x_1}^{t_2} \cdots \int_{x_{n-1}=x_{n-2}}^{t_{n-1}} \frac{t_n^2 - x_{n-1}^2}{2} dx_{n-1} \ldots dx_1 \\
&= t_{n+1} g(t_1, \ldots, t_n) - \frac{t_n^2}{2} g(t_1, \ldots, t_{n-1}) + \int_{x_1=0}^{t_1} \cdots \int_{x_{n-2}=x_{n-3}}^{t_{n-2}} \frac{t_{n-1}^3 - x_{n-2}^3}{3!} dx_{n-2} \ldots dx_1 \\
&= \ldots = \sum_{k=0}^{n} (-1)^{n-k} \frac{t_{k+1}^{n+1-k}}{(n+1-k)!} g(t_1, t_2, \ldots, t_k).
\end{aligned}$$

Remark. We showed that $\mathrm{P}(B_{n+1})$ and $\mathrm{P}(C_{n+1})$ can be computed recursively using $O(n)$ operations for every recalculation of each event.

Formula (3.1) follows from:

$$\begin{aligned}
\mathrm{E}\,[A(t)/B_{n+1}] &= \mathrm{E}\,[A(t) \,/\, \{B_n \setminus D_n\}] = \\
&= \frac{\mathrm{E}\,[A(t)/B_n]\mathrm{P}\{B_n\} - \mathrm{E}\,[A(t)/D_n]\mathrm{P}\{D_n\}}{\mathrm{P}\{B_n\} - \mathrm{P}\{D_n\}} = \\
&= \frac{\mathrm{E}\,[A(t)/B_n]\mathrm{P}\{B_n\} - \mathrm{E}\,[A(t)/D_n]\mathrm{P}\{C_n\} e^{-\lambda(t_{n+1} - t_n)}}{\mathrm{P}\{B_n\} - \mathrm{P}\{C_n\} e^{-\lambda(t_{n+1} - t_n)}}.
\end{aligned}$$

Formula (3.6) is true since for $t \ge t_n$:

$$\begin{aligned}
\mathrm{E}\,[A(t)/D_n] &= \mathrm{E}\,[A(t_n)/D_n] + \mathrm{E}\,[(A(t) - A(t_n))/D_n] = \\
&= n + \lambda(t - t_{n+1})_+.
\end{aligned}$$

We still need to prove formula (3.7). For $t_i \le t < t_{i+1} \le t_n$,

$$\mathrm{E}\,[A(t)/D_n] = \mathrm{E}\,[A(t)/C_n].$$

Then

$$\begin{aligned}
\mathrm{E}\,[A(t)/C_n] &= \sum_{k=i}^{n} k\mathrm{P}\{A(t)=k/C_n\} = \\
&= \frac{1}{\mathrm{P}\{C_n\}} \sum_{k=i}^{n} k\mathrm{P}\{A(t)=k\}\mathrm{P}\{C_n/A(t)=k\} = \\
&= \frac{1}{\mathrm{P}\{C_n\}} \sum_{k=i}^{n} e^{-\lambda t}\frac{(\lambda t)^k}{(k-1)!}\mathrm{P}\{B_i/A(t)=k\}\mathrm{P}\{C_n \setminus B_k \,/\, A(t)=k\}. \qquad (3.8)
\end{aligned}$$

Formula (3.8) follows from conditional independence of B_i and $C_n \setminus B_i$ given $A(t)$ and also from

$$\mathrm{P}\{C_n \setminus B_i \,/\, A(t)=k\} = \mathrm{P}\{C_n \setminus B_k \,/\, A(t)=k\}.$$

Now we calculate the two probabilities from (3.8) separately.

$$\begin{aligned}
\mathrm{P}\{B_i/A(t)=k\} &= \mathrm{P}\{A_1 \le t_1, \ldots, A_i \le t_i, \ldots, A_{i+1} \le t, A_k \le t\} \\
&= \frac{k!}{t^k}\int_{x_1=0}^{t_1}\int_{x_2=x_1}^{t_2}\cdots\int_{x_i=x_{i-1}}^{t_i}\int_{x_{i+1}=x_i}^{t}\cdots\int_{x_k=x_{k-1}}^{t} dx_k\ldots dx_1 \qquad (3.9) \\
&= \frac{k!}{t^k}\int_{x_1=0}^{t_1}\int_{x_2=x_1}^{t_2}\cdots\int_{x_i=x_{i-1}}^{t_i}\frac{(t-x_i)^{k-i}}{(k-i)!}dx_i\ldots dx_1 \qquad (3.10) \\
&= \frac{k!}{t^k}\Big(\frac{t^k}{k!} - \sum_{m=0}^{i-1}\frac{(t-t_{m+1})^{k-m}}{(k-m)!}g(t_1,\ldots,t_m)\Big) \qquad (3.11) \\
&= 1 - \frac{k!}{t^k}\sum_{m=0}^{i-1}\frac{(t-t_{m+1})^{k-m}}{(k-m)!}g(t_1,\ldots,t_m).
\end{aligned}$$

Here (3.9) follows from the order statistics property, Lemma 1 of Bertsimas and Servi [2] implies (3.10), and (3.11) is derived by integration of (3.10).

As for the second probability from (3.8),

$$\mathrm{P}\{C_n \setminus B_k \,/\, A(t)=k\} = e^{-\lambda(t_n-t)}\lambda^{n-k}g(t_{k+1}-t, t_{k+2}-t, \ldots, t_n-t), \qquad (3.12)$$

and $g(t_{k+1}-t, t_{k+2}-t, \ldots, t_n-t)$ can be computed recursively using (3.5).

Formulas (3.8), (3.11) and (3.12) imply (3.7).

Remark. Note that extrapolation requires $O(n)$ computations for each updating of the estimate. However, if the interpolation problem is considered, the number increases to $O(n^2)$. (The calculation of $\mathrm{E}\,[A(t)/D_n]$ is critical.)

3.3 Immediate Transitions, Known External Arrivals.

Our first goal is to introduce the third representation of the observed event E_t, namely via stochastic components of the network. Then $Q_j(t)$, $j \in \mathcal{D}$, will be estimated at a given $t \geq 0$ and the general interpolation problem of $Q_j(s)$ estimation ($s < t$) will be solved. Finally, a dynamical problem of $Q_j(t)$ reestimation in real-time will be considered: we take some $h > t$ and recalculate $\mathrm{E}[Q_j(h)/E_h]$.

Notation and Definitions. Recall the stochastic components of the network: A_i^0 stand for the external arrivals, ξ_{ki}^j denote the service durations and $R_{ml}^n \in \mathcal{D}_0$ (or $R_l^0 \in \mathcal{D}_0$) stand for the switches of customers.

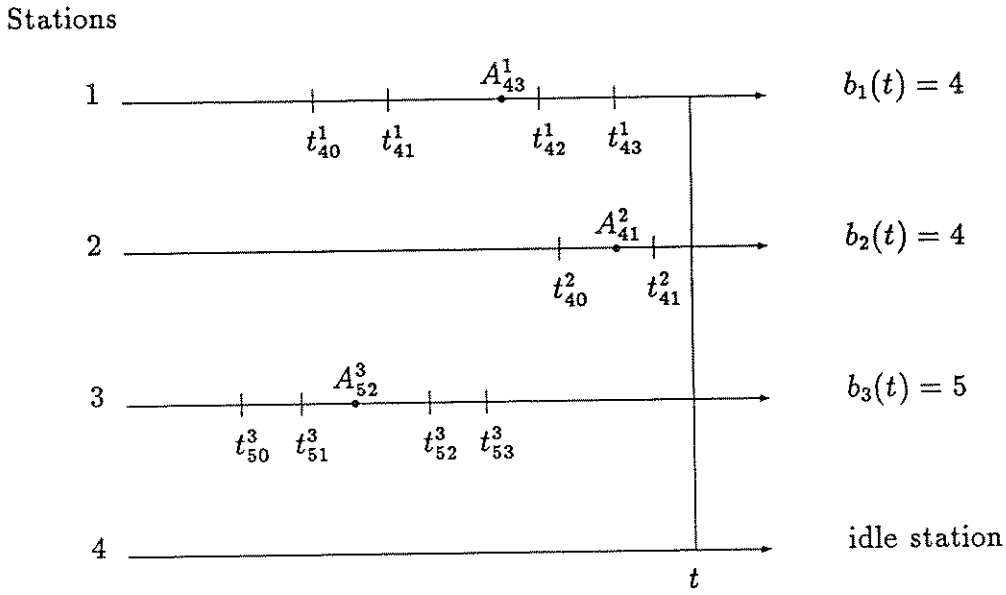
Then introduce a d-dimensional stochastic process, $C = \{C(t), t \geq 0\}$, where the j-th coordinate of $C(t)$, $C_j(t)$ denotes the arrival time to station j, of the customer who is served there at time t. If station j is idle, put $C_j(t) = t$. Using the notation of Section 2, we can write $C_j(t) = A^j_{b_j(t),m_j(t)}$ for $j \in d_b(t)$. Note that $C_j(t)$ is determined by the observed event E_t. Indeed, information about the history of this customer provides us with his last registration time at the predecessor station to j, which is precisely $C_j(t)$ because transitions in the network are immediate.

The customers that terminated service at IR-times (see section 2 for definition) are called **invisible customers**. For every IR-time t_{ml}^n we introduce the **available stations process**

$$\mathcal{M}_t(t_{ml}^n) = \{\{j : C_j(t) < t_{ml}^n\} \bigcup \{0\}\}.$$

The following figure illustrates the definition.

Figure 3.1. Illustration to Definitions of $C_j(t)$ and $\mathcal{M}_t(t^n_{ml})$



Registrations of the last incomplete busy periods at several stations are given. Here $C_1(t) = A^1_{43}$, $C_2(t) = A^2_{41}$, $C_3(t) = A^3_{52}$, $C_4(t) = t$ and, for example, $\mathcal{M}_t(t^2_{41}) = \{0, 1, 2, 3\}$, $\mathcal{M}_t(t^3_{52}) = \{0, 3\}$, $\mathcal{M}_t(t^3_{51}) = \{0\}$.

If a customer stands in some queue at t, then his last registration has been either arrival to the network or service termination (it could not have been service start). If this registration occurred before $C_j(t)$, he can not queue at station j according to the FCFS queue discipline and the assumption of immediate transitions. This argument provides us with two important statements.

- Only invisible customers occupy the queues of the network.
- Invisible customer that terminated service at t^n_{ml} must reside at one of the stations of $\mathcal{M}_t(t^n_{ml})$.

Remark 3.1. If the IR-time $t^n_{ml} < \min_{j \in \mathcal{D}} C_j(t)$ we know definitely that $\mathcal{M}(t^n_{ml}) = \{0\}$, therefore, the corresponding invisible customer have left the network.

Representation 3 of $\mathbf{E_t}$ via Stochastic Components.
Immediate Transitions, Known External Arrivals.

The parenthesized letters that end an index of an event correspond to the stochastic components that are involved in its description. Specifically:

A — external arrivals;

S — service times;

R — routing between stations;

T — transition times (in the case of exponential transition times).

3.1(S) Services that terminated up to t are known.
$\xi^{\mathbf{j}}_{\mathbf{ki}} = \mathbf{t}^{\mathbf{j}}_{\mathbf{k,i+1}} - \mathbf{t}^{\mathbf{j}}_{\mathbf{ki}}$, $j \in \mathcal{D}$, $1 \le k \le b_j(t)$,
$0 \le i \le m_j(t) - 1$, for $j \in d_b(t)$, $k = b_j(t)$,
$0 \le i \le n^j_k$, otherwise.

3.2(S) Last service at a busy station have not terminated until t.
For $j \in d_b(t)$ $\xi^{\mathbf{j}}_{\mathbf{b_j(t),m_j(t)}} > \mathbf{t} - \mathbf{t}^{\mathbf{j}}_{\mathbf{b_j(t),m_j(t)}}$.

3.3(R) Complete information about the switches of the visible customers.
For $(n,m,l) \in V_R(t)$ and $(0,l) \in V_R(t)$
$\mathbf{R}^{\mathbf{n}}_{\mathbf{ml}} = \mathbf{j}$ $(\mathbf{R}^{\mathbf{0}}_{\mathbf{l}} = \mathbf{j})$.

3.4(R) Partial information about the invisible routing.
For $(n,m,l) \in V_R(t)$ and $(0,l) \in I_R(t)$
$\mathbf{R}^{\mathbf{n}}_{\mathbf{ml}} \in \mathcal{M}_{\mathbf{t}}(\mathbf{t}^{\mathbf{n}}_{\mathbf{ml}})$ $(\mathbf{R}^{\mathbf{0}}_{\mathbf{l}} \in \mathcal{M}_{\mathbf{t}}(\mathbf{t}^{\mathbf{0}}_{\mathbf{l}}))$.

3.5(A) The number of external arrivals is known.
$\mathbf{N^0(t)} = \mathbf{n^0(t)}$.

3.6(A) The times of external arrivals are known.
$\mathbf{A}^{\mathbf{0}}_{\mathbf{i}} = \mathbf{t}^{\mathbf{0}}_{\mathbf{i}}$, $1 \le i \le n^0(t)$.

Equivalence of Representations 2 and 3. In the beginning, note that events 3.1, 3.2, 3.3, 3.5 and 3.6 coincide with events 2.1, 2.2, 2.6, 2.8 and 2.9 respectively. The assumption of immediate transitions between the stations can be expressed as

$$R(n,m,l) = (j,k,i) \Rightarrow A^j_{ki} = \bar{A}^n_{ml} = t^n_{ml}. \tag{3.13}$$

In order to prove the equivalence of events 2.7 and 3.4, recall the first part of 2.7:

$$\text{if } j \in d_b(t), \quad R^n_{ml} = j \Rightarrow \bar{A}^n_{ml} \ge A^j_{b_j(t),n_j(t)}.$$

Using $C_j(t) = A^j_{b_j(t),m_j(t)}$, $\bar{A}^n_{ml} = t^n_{ml}$ and definition of $\mathcal{M}_t$ we get the equivalence. The second part of 2.7 is irrelevant since in the case of immediate transitions an invisible customer cannot switch to an idle station.

Therefore, Representation 3 follows from 2.

As for the second direction, the arrivals $\bar{A}^n_{ml}$, where $t^n_{ml} \in V_R(t)$, are completely determined by the stochastic components from events 3.1-3.5. Hence events 2.3-2.5 will stand automatically if 3.1-3.5 stand.

Queue Estimation at Time t. We proceed to compute the conditional expectation and the variance of the queue at station j.

Proposition 3.2.

$$\mathrm{E}[Q_j(t)/E_t] = \sum_{t^n_{ml}\in I_R(t)} 1_{\{j\in\mathcal{M}_t(t^n_{ml})\}}\frac{p_{nj}}{\sum_{k\in\mathcal{M}_t(t^n_{ml})} p_{nk}}, \tag{3.14}$$

$$\mathrm{Var}[Q_j(t)/E_t] = \sum_{t^n_{ml}\in I_R(t)} 1_{\{j\in\mathcal{M}_t(t^n_{ml})\}}\frac{p_{nj}(1-p_{nj})}{(\sum_{k\in\mathcal{M}_t(t^n_{ml})} p_{nk})^2}. \tag{3.15}$$

Proof.

$$\mathrm{E}[Q_j(t)/E_t] = \mathrm{E}[\sum_{t^n_{ml}\in I_R(t)} 1_{\{R^n_{ml}=j\}}/E_t] = \tag{3.16}$$

$$= \sum_{t^n_{ml}\in I_R(t)} \mathrm{P}\{R^n_{ml}=j/E_t\} =$$

$$= \sum_{t^n_{ml}\in I_R(t)} \mathrm{P}\{R^n_{ml}=j/R^n_{ml}\in\mathcal{M}_t(t^n_{ml})\} = \tag{3.17}$$

$$= \sum_{t^n_{ml}\in I_R(t)} 1_{\{j\in\mathcal{M}(t^n_{ml})\}}\frac{p_{nj}}{\sum_{k\in\mathcal{M}(t^n_{ml})} p_{nk}} \tag{3.18}$$

Formula (3.16) is true since only invisible customers occupy the queues of the network. Routing of the invisible customers is involved in representation 3 only in event 3.4. This fact and independence of the stochastic components of the network imply (3.17). Finally, (3.18) follows from a straightforward calculation.

Formula (3.15) prevails since transitions between the stations are conditionally independent given E_t.

General Interpolation. The queue is estimated at station j at time s, before observation time t.

Proposition 3.3.
If $s \le C_j(t)$, $Q_j(s)$ is known exactly given E_t and

$$Q_j(s) = A_j(s) - A_j(C_j(s)). \tag{3.19}$$

If $s > C_j(t)$,

$$Q_j(s) = [A_j(C_j(t)) - A_j(C_j(s))] + [A_j(s) - A_j(C_j(t))].$$

The first term is known exactly given E_t and

$$\mathrm{E}[A_j(s) - A_j(C_j(t))/E_t] = \sum_{t^n_{ml}\in I_R(t)\ \mathrm{s.t.}\ t^n_{ml}<s} 1_{\{j\in\mathcal{M}_t(t^n_{ml})\}}\frac{p_{nj}}{\sum_{k\in\mathcal{M}_t(t^n_{ml})} p_{nk}}.$$

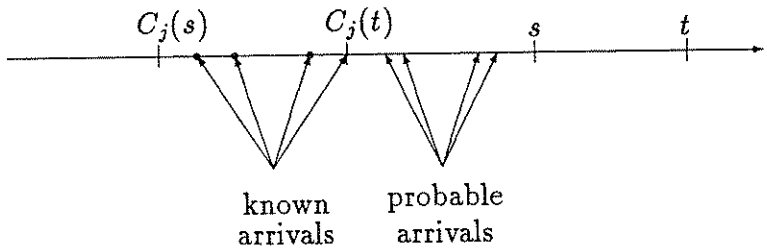
Proof.

Formula (3.19) prevails since customers that stand in queue at s arrived in $[C_j(s), s]$ due to FCFS discipline. All arrivals to station j which took place before $C_j(t)$ are known because the corresponding customers started service before t and we can trace their previous registrations using ID. Therefore, $A_j(C_j(t)) - A_j(C_j(s))$ is known exactly.

Arrivals that took place during $[C_j(t), s]$ (if such interval exists) are estimated using the same argument as for formula (3.14).

Figure 3.2 illustrates our approach.

Figure 3.2. Illustration to Proposition 3.3.



Queue Estimation in Real-Time. Dynamic Algorithm. Suppose that we observe the network in real-time and estimate the queue conditioning on the observed event (in the sense of conditional expected value). Then an algorithm for the queue reestimation at the successive estimation times is needed.

We assume that the last estimation of the queue was carried out at time t and the reestimation takes place at time $h > t$ when *no registrations occur between t and h* (but there can be registrations at h itself).

The following information must be saved in the memory of our computer for every estimation time t.

- The matrix of the routing probabilities P.
- The set of the busy nodes $d_b(t)$.
- Information concerning invisible customers.
  From Remark 3.1 only those invisible customers that were registered last time after $\min_{j \in \mathcal{D}} C_j(t)$ can possibly wait in some queue at time t. We use notation $\mathcal{L}_t$ for the set of such customers at time t. Then for each customer the following characteristics are stored.
  - $s_k(t)$, $k \in \mathcal{L}_t$, are times of the last registration of invisible customers.
  - $n_k(t)$, $k \in \mathcal{L}_t$, denote the last stations where invisible customers have been observed ($n_k(t) = 0$ is possible).

– $\mathcal{M}_k(t)$, $k \in \mathcal{L}_t$, are sets of available stations for each customer.

- The matrix of the conditional routing probabilities $\tilde{P}_t = \tilde{p}_{kj}(t)$, $k \in \mathcal{L}_t$, $j \in \mathcal{D}$, where

$$\tilde{p}_{kj}(t) = 1_{\{j \in \mathcal{M}_k(t)\}} \frac{p_{n_k,j}}{\sum_{l \in \mathcal{M}_k(t)} p_{n_k,l}}. \tag{3.20}$$

Rewriting formula (3.14), we get the queue estimate:

$$\mathrm{E}[Q_j(t)/E_t] = \sum_{k=1}^{I_t} \tilde{p}_{kj}(t). \tag{3.21}$$

Now introduce dynamic reestimation algorithm. Several special cases of events, which could happen at h, are considered. The correctness of the algorithm can be verified from (3.20), (3.21) and the preceding definitions.

Algorithm 3.2

- *No registrations took place at h.* The queue estimate does not change. Network data, stored in memory, is also preserved. (Further if we do not mention some part of the data then it does not change when we move from t to h.)

- *Customer number k arrived to station 0 at h and immediately started service at station v.*

$$d_b(h) = d_b(t) \bigcup \{v\}.$$

- *Customer number k arrived to station 0 at h.*

$$\mathcal{L}_h = \mathcal{L}_t \bigcup \{k\},$$

$$s_k(h) = h, \quad N_k(h) = 0, \quad \mathcal{M}_k(h) = d_b(h) = d_b(t),$$

$$\tilde{p}_{kj}(t) = \frac{p_{0j}}{\sum_{v \in \mathcal{M}_k(h)} p_{0v}}, \quad j \in d_b(t),$$

$$\mathrm{E}[Q_j(h)/E_h] = \mathrm{E}[Q_j(t)/E_t] + \tilde{p}_{kj}, \quad j \in \mathcal{D}.$$

- *Customer number k terminated service at station u at time h and immediately started a new one at station v.*

$$d_b(h) = \{d_b(t) \setminus \{u\}\} \bigcup \{v\}.$$

For $l \in \mathcal{L}_h$ s.t. $u \in \mathcal{M}_l(t)$ let

$$\mathcal{M}_l(h) = \mathcal{M}_l(t) \setminus \{u\} \text{ and}$$

$$\tilde{p}_{lj}(h) = \tilde{p}_{lj}(t) \frac{\sum_{m \in \mathcal{M}_l(t)} p_{n_l,m}}{\sum_{m \in \mathcal{M}_l(h)} p_{n_l,m}}, \quad j \in \mathcal{M}_l(h).$$

$$\mathrm{E}[Q_j(h)/E_h] = \sum_{k=1}^{I_h} \tilde{p}_{kj}(h).$$

- *Customer number k terminated service at station u at time h.*

$$d_b(h) = d_b(t) \setminus \{u\},$$

$$\mathcal{L}_h = \mathcal{L}_t \bigcup \{k\},$$

$$s_k(h) = h, \quad N_k(h) = u, \quad \mathcal{M}_k(h) = d_b(h),$$

$$\tilde{p}_{kj}(t) = \frac{p_{0j}}{\sum_{v \in \mathcal{M}_k(h)} p_{0v}}, \quad j \in d_b(t)$$

For $l \in \mathcal{L}_h$ s.t. $u \in \mathcal{M}_l(t)$ let

$$\mathcal{M}_l(h) = \mathcal{M}_l(t) \setminus \{u\} \text{ and}$$

$$\tilde{p}_{lj}(h) = \tilde{p}_{lj}(t) \frac{\sum_{m \in \mathcal{M}_l(t)} p_{n_l,m}}{\sum_{m \in \mathcal{M}_l(h)} p_{n_l,m}}, \quad j \in \mathcal{M}_l(h).$$

$$\mathrm{E}[Q_j(h)/E_h] = \sum_{k=1}^{I_h} \tilde{p}_{kj}(h).$$

- *Customer number k terminated service at station u at time h and immediately started a new one at station v. At the same moment customer number g started service at station u.*

$$d_b(h) = d_b(t) \bigcup \{v\},$$

$$\mathcal{L}_h = \mathcal{L}_t \setminus \{g\},$$

For $l \in \mathcal{L}_h$ s.t. $u \in \mathcal{M}_l(t)$, $s_u(t) < s_g(t)$ let

$$\mathcal{M}_l(h) = \mathcal{M}_l(t) \setminus \{u\} \text{ and}$$

$$\tilde{p}_{lj}(h) = \tilde{p}_{lj}(t) \frac{\sum_{m \in \mathcal{M}_l(t)} p_{n_l,m}}{\sum_{m \in \mathcal{M}_l(h)} p_{n_l,m}}, \quad j \in \mathcal{M}_l(h).$$

$$\mathrm{E}[Q_j(h)/E_h] = \sum_{k=1}^{I_h} \tilde{p}_{kj}(h).$$

- *Customer number k terminated service at station u at time h. At the same moment customer number g started service at station u.*

$$\mathcal{L}_h = \{\mathcal{L}_t \setminus \{g\}\} \bigcup \{k\},$$

$$s_k(h) = h, \quad N_k(h) = u, \quad \mathcal{M}_k(h) = d_b(h) = d_b(t),$$

$$\tilde{p}_{kj}(t) = \frac{p_{0j}}{\sum_{v \in \mathcal{M}_k(h)} p_{0v}}, \quad j \in d_b(t)$$

For $l \in \mathcal{L}_h$ s.t. $u \in \mathcal{M}_l(t)$, $s_u(t) < s_g(t)$ let

$$\mathcal{M}_l(h) = \mathcal{M}_l(t) \setminus \{u\} \text{ and}$$

$$\tilde{p}_{lj}(h) = \tilde{p}_{lj}(t)\frac{\sum_{m\in\mathcal{M}_l(t)} p_{n_l,m}}{\sum_{m\in\mathcal{M}_l(h)} p_{n_l,m}}, \quad j \in \mathcal{M}_l(h).$$

$$\mathrm{E}[Q_j(h)/E_h] = \sum_{k=1}^{I_h} \tilde{p}_{kj}(h).$$

Remark. Note, that for the last four special cases considerable recalculations of our estimates are needed.

Multi-Server Case. All our results remain valid for the multi-server case. The only alteration concerns the definition of process C. In this case $C_j(t)$ denotes the arrival time to station j, of the customer who has started service last among all customers currently being served there.

3.4 Exponential Transitions, Known External Arrivals.

First, we introduce new stochastic components of the network. Set τ^n_{ml} and τ^0_l stand for the transition time of the customer, which terminated service at t^n_{ml} (t^0_l). If the customer switched to station j then $\tau^n_{ml} \sim \exp(\eta_{nj})$.

Arrivals to the stations are given by:

$$\begin{aligned} R(n,m,l) = (j,k,i) \quad &\Rightarrow \quad A^j_{ki} = t^n_{ml} + \tau^n_{ml} \\ (\text{ or } R(0,l) = (j,k,i) \quad &\Rightarrow \quad A^j_{ki} = t^0_l + \tau^0_l \ ) \end{aligned} \tag{3.22}$$

Representation 4 of $\mathbf{E_t}$ via Stochastic Components.
Exponential Transitions, Known External Arrivals.

4.1(S) Services that terminated up to t are known.
$\xi^{\mathbf{j}}_{\mathbf{ki}} = \mathbf{t^j_{k,i+1} - t^j_{ki}}$, $j \in \mathcal{D}$, $1 \le k \le b_j(t)$,
$0 \le i \le m_j(t) - 1$, for $j \in d_b(t)$, $k = b_j(t)$,
$0 \le i \le n^j_k$, otherwise.

4.2(S) Last service at a busy station have not terminated until t.
For $j \in d_b(t)$ $\xi^{\mathbf{j}}_{\mathbf{b_j(t),m_j(t)}} > \mathbf{t} - \mathbf{t^j_{b_j(t),m_j(t)}}$.

4.3(T) The start of a busy period coincides with the first arrival.
For $(n,m,l) \in V_R(t) \ni \mathcal{R}(n,m,l) = (j,k,0)$
$\mathbf{t^n_{ml} + \tau^n_{ml} = t^j_{k0}}$, $j \in \mathcal{D}$, $1 \le k \le b_j(t)$.

4.4a(T) Truncation rule for complete busy periods.
For $(n,m,l) \in V_R(t) \ni \mathcal{R}(n,m,l) = (j,k,i)$, $i \ge 1$
$\mathbf{t^n_{ml} + \tau^n_{ml} \le t^j_{ki}}$, $j \in \mathcal{D}$,

$1 \le k \le b_j(t), \quad$ for $j \notin d_b(t)$,
$1 \le k \le b_j(t) - 1, \quad$ for $j \in d_b(t)$,
$1 \le i \le n_k^j$.

4.4b(T) Truncation rule for incomplete busy periods.
For $(n,m,l) \in V_R(t) \ni \mathcal{R}(n,m,l) = (j, b_j(t), i)$, $i \ge 1$, $j \in d_b(t)$,
$\mathbf{t^n_{ml} + \tau^n_{ml} \le t^j_{b_j(t),i}}$.

4.5a(T) FCFS discipline for complete busy periods.
For $(n_i, m_i, l_i) \in V_R(t) \ni \mathcal{R}(n_i, m_i, l_i) = (j, k, i)$
$\mathbf{t^j_{k0} \le t^{n_1}_{m_1,l_1} + \tau^{n_1}_{m_1,l_1} \le t^{n_2}_{m_2,l_2} + \tau^{n_2}_{m_2,l_2} \le \dots \le t^{nu}_{mu,lu} + \tau^{nu}_{mu,lu}}$, $\quad j \in \mathcal{D}$,
$1 \le k \le b_j(t), \quad$ for $j \notin d_b(t)$,
$1 \le k \le b_j(t) - 1, \quad$ for $j \in d_b(t)$,
$1 \le i \le u = n_k^j$.

4.5b(T) FCFS discipline for incomplete busy periods.
For $(n_i, m_i, l_i) \in V_R(t) \ni \mathcal{R}(n_i, m_i, l_i) = (j, b_j(t), i)$
$\mathbf{t^j_{b_j(t),0} \le t^{n_1}_{m_1,l_1} + \tau^{n_1}_{m_1,l_1} \le t^{n_2}_{m_2,l_2} + \tau^{n_2}_{m_2,l_2} \le \dots \le t^{nu}_{mu,lu} + \tau^{nu}_{mu,lu}}$, $\quad j \in d_b(t)$,
$1 \le i \le u = m_j(t)$.

4.6(R) Complete information about the switches of the visible customers.
For $(n,m,l) \in V_R(t)$ and $(0,l) \in V_R(t)$
$\mathbf{R^n_{ml} = j}$ $(\mathbf{R^0_l = j})$.

4.7(RT) Partial information about the invisible routing.
For $(n,m,l) \in I_R(t)$:
$\mathbf{if\, j \notin d_b(t)}, \quad \mathbf{R^n_{ml} = j \Rightarrow t^n_{ml} + \tau^n_{ml} > t_{w_j}}$,
$\mathbf{if\, j \in d_b(t)}, \quad \mathbf{R^n_{ml} = j \Rightarrow t^n_{ml} + \tau^n_{ml} \ge t_{v_j,u_j} + \tau^{w_j}_{v_j,u_j}}$, where
$(w_j, v_j, u_j) = \mathcal{R}^{-1}(j, b_j(t), m_j(t))$.
(The similar relations are true for $(0,l) \in I_R(t)$.)

4.8(A) The number of external arrivals is known.
$\mathbf{N^0(t) = n^0(t)}$.

4.9(A) The times of external arrivals are known.
$\mathbf{A^0_i = t^0_i}$, $1 \le i \le n^0(t)$.

Equivalence of Representations 2 and 4. The equivalence immediately follows from formula (3.22). In Representation 4 $\bar{A}^n_{ml}$ from Representation 2 is replaced by $t^n_{ml} + \tau^n_{ml}$. In addition, each of the events 2.4 and 2.5 is divided into two separate events.

Queue estimation at time t. Relevant information. We introduce again the set of invisible customers. It includes the customers, which terminated service or arrived to the network at IR-times. As in Subsection 3.3, we state that only invisible customers occupy the queues of the network. Therefore,

$$\mathrm{E}[Q_j(t)/E_t] = \sum_{t^n_{ml}\in I_R(t)} \mathrm{P}\{R^n_{ml}=j/E_t\} = \tag{3.23}$$

$$= \sum_{t^n_{ml}\in I_R(t)} \mathrm{P}\{R^n_{ml}=j/E^7_t, E^{5b}_t, E^{4b}_t\} \tag{3.24}$$

$$= \sum_{t^n_{ml}\in I_R(t)} \mathrm{P}\{R^n_{ml}=j/E_r\}, \tag{3.25}$$

where events E^7_t, E^{5b}_t and E^{4b}_t correspond to events 4.7, 4.5b and 4.4b of Representation 4 and $E_r = E^{4b}_t \bigcap E^{5b}_t \bigcap E^{7b}_t$ is the relevant information. Is is straightforward to check that all other events include stochastic components, which are independent of $R^n_{ml} \ni t^n_{ml} \in I_R(t)$ and of the stochastic components, involved in events 4.7, 4.5b and 4.4b. Dependence of the Invisible Routing on event 4.7 is obvious. Dependence on events 4.5b and 4.4b appears through $\tau^{wj}_{v_j,u_j}$.

Mathematical Formulation of the Estimation Problem. In the beginning, we are concerned with notational arrangements. The time index t will be omitted whenever convenient. (Our algorithm is static, therefore it will make no harm.) Registrations of incomplete busy periods $t^j_{b_j(t),0},\ldots,t^j_{b_j(t),m_j(t)}$ are called $t^j_0,\ldots,t^j_{m_j}$. Let $s^j_0,\ldots,s^j_{m_j}$ stand for the transition start times of the corresponding customers. (They are provided by operator R^{-1}_T). The last transition times of these customers are called $\tau^j_0,\ldots,\tau^j_{m_j}$ respectively.

Suppose that K invisible customers exist at time t. Their last transition start times are denoted by $s^0_k,\ \ 1\le k\le K$ and $R_k,\ \ 1\le k\le K$ stands for the station where they switch. In the end, set $\tau^0_k,\ 1\le k\le K$ denote the transition times of the invisible customers.

Using the new notation the relevant events of Representation 4 are expressed as follows:

4.4b $s^j_i+\tau^j_i\le t^j_i,\quad j\in d_b,\ 1\le i\le m_j.$

4.5b $t^j_0\le s^j_1+\tau^j_1\le\ldots\le s^j_{m_j}+\tau^j_{m_j},\quad j\in d_b.$

4.7 if $j\in d_b,\ \ R_k=j\Rightarrow s^0_k+\tau^0_k\ge s^j_{m_j}+\tau^j_{m_j},$
if $j\notin d_b,\ \ R_k=j\Rightarrow s_k+\tau_k>t.$

We need to calculate

$$\mathrm{E}[Q_j(t)/E_r]=\sum_{k=1}^{K}\mathrm{P}\{R_k=j/E_r\}.$$

Formulation as Hidden Markov Chain Problem. We shall solve the estimation problem using the approach from section 1.3. However, the Markov chain in the Real-Time case will be considerably more complicated. (We cannot focus on a specific station, all the stations together are considered.)

Time domain of the Markov chain. Set $M = \max_{j \in d_b} m_j$ and $t_k(j) = t^j_{\min(k,m_j)}$, $0 \leq k \leq M$, $j \in d_b$. Then our Markov chain is defined on $\tilde{t}_0, \tilde{t}_1, \ldots, \tilde{t}_M, \tilde{t}_{M+1}$, where

$$\tilde{t}_0 = \{t^j_0,\ j \in d_b(t)\},\quad \tilde{t}_k = \{t_k(j),\ j \in d_b(t)\},\ 1 \leq k \leq M,\quad \tilde{t}_{M+1} = \{t, t, \ldots, t\}.$$

State space. The Markov chain can be represented as:

$$\tilde{A} = \{\tilde{A}(\tilde{t}_0), \tilde{A}(\tilde{t}_1), \ldots, \tilde{A}(\tilde{t}_M), \tilde{A}(\tilde{t}_{M+1})\},$$

where

$$\tilde{A}(\tilde{t}_k) = \{\tilde{A}_j(t_k(j)),\ j \in d_b(t)\},$$

and

$$\tilde{A}_j(\tilde{t}_k(j)) = \{\tilde{A}^V_j(t_k(j)) \bigcup \tilde{A}^I_j(t_k(j))\}.$$

In the last expression, $\tilde{A}^V_j(t_k(j))$ is the ordered set of visible customers, which arrived to station j until $t_k(j)$. (Customers are ordered according to their arrival times.) Finally, $\tilde{A}^I_j(t_k(j))$ is the unordered set of invisible customers, which arrived to station j until $t_k(j)$.

Boundary condition.

$$\tilde{A}(\tilde{t}_0) = \{\emptyset, \ldots, \emptyset\}.$$

Admissible states.

Let (j,i) denote an ID of a customer that started service at t^j_i. Then, if $t_k(j) = t^j_i$,

$$\tilde{A}^V_j(t_k(j)) \in \{\{(j,1),(j,2),\ldots,(j,i)\},\{(j,1),(j,2),\ldots,(j,i{+}1)\},\ldots,\{(j,1),(j,2),(j,m_j)\}\}$$

(customer (j,i) arrives before his service starts; FCFS principle prevails).

If $(j,m_j) \notin \tilde{A}^V_j(t_k(j))$, then $\tilde{A}^I_j(t_k(j)) = \emptyset$ (all visible customers must appear before any invisible one arrives).

Transition probabilities.

We must calculate

$$\mathrm{P}\{\tilde{A}(\tilde{t}_{k+1}) = C_2 / \tilde{A}(\tilde{t}_k) = C_1\}$$

for admissible states C_1 and C_2.

Define the j-th component of C_1 and C_2 as C^j_1 and C^j_2, $j \in d_b$. Then we can rewrite the last probability as

$$\mathrm{P}\{\tilde{A}_j(t_{k+1}(j)) = C^j_2,\ j \in d_b / \tilde{A}_j(t^j_k) = C^j_1,\ j \in d_b\}.$$

We can represent the considered event as the intersection of the following four events: $B_1 = \{$ Visible customers that do not belong to C^j_2 did not arrive during $[t_k(j), t_{k+1}(j)]$, $j \in$

d_b },
B_2 = { Invisible customers that belong to C_2^j and do not belong to C_1^j arrived during $[t_k(j), t_{k+1}(j)]$, $j \in d_b$. Their arrival times are ordered according to their service start times },
B_3 = { Invisible customers that do not belong to C_2^j did not arrive during $[t_k(j), t_{k+1}(j)]$, $j \in d_b$ },
B_4 = { Invisible customers that belong to C_2^j arrived during $[t_k(j), t_{k+1}(j)]$, $j \in d_b$. Their arrivals took place after arrivals of all visible customers.}.

Events that correspond to two sentences from event B_2 definition will be called B_2^1 (visible customers arrived) and B_2^2 (order relation). In the same way we define events B_4^1 and B_4^2.

Let M_1^j stand for a set of customers that appear in event B_1 in connection with station j. Sets M_2^j, M_3 and M_4^j correspond to events B_2, B_3 and B_4.

In the following reasoning we use the chain formula:

$$\begin{aligned} \mathrm{P}\{\tilde{A}(\tilde{t}_{k+1}) = C_2 / \tilde{A}(\tilde{t}_k) = C_1\} &= \mathrm{P}\{B_1 B_2 B_3 B_4 / \tilde{A}(\tilde{t}_k) = C_1\} = \\ &= \mathrm{P}\{B_1 / \tilde{A}(\tilde{t}_k) = C_1\} \mathrm{P}\{B_2 / B_1, \tilde{A}(\tilde{t}_k) = C_1\} \mathrm{P}\{B_3 / B_1, B_2, \tilde{A}(\tilde{t}_k) = C_1\} \times \\ &\times \mathrm{P}\{B_4 / B_1, B_2, B_3, \tilde{A}(\tilde{t}_k) = C_1\}. \end{aligned} \tag{3.26}$$

We shall consider in turn the four probabilities from formula (3.26).

$$\mathrm{P}\{B_1 / \tilde{A}(\tilde{t}_k) = C_1\} = \prod_{j \in d_b} \prod_{l \in M_1^j} \exp\{-\gamma_{lj}(t_{k+1}(j) - \max(t_k(j), s_l))_+\}. \tag{3.27}$$

Here γ_{lj} and s_l are transition rates and transition start-times of the corresponding customers from M_1^j. Specifically $\gamma_{lj} = \eta_{n_l j}$, where n_l is the transition-start station of the invisible customer l.

$$\mathrm{P}\{B_2 / B_1, \tilde{A}(\tilde{t}_k) = C_1\} = \mathrm{P}\{B_2 / \tilde{A}(\tilde{t}_k) = C_1\} = \tag{3.28}$$

$$\begin{aligned} &= \mathrm{P}\{B_2^1 / \tilde{A}(\tilde{t}_k) = C_1\} \times \mathrm{P}\{B_2^2 / B_2^1, \tilde{A}(\tilde{t}_k) = C_1\} = \\ &= \prod_{j \in d_b} \prod_{l \in M_2^j} (1 - \exp\{-\gamma_{lj}(t_{k+1}(j) - \max(t_k(j), s_l))_+\}) \times \\ &\times \prod_{j \in d_b} \mathrm{P}\{A_1^j < A_2^j < \ldots < A_{|M_2^j|}^j\}, \end{aligned} \tag{3.29}$$

where A_l^j from formula (3.29) are truncated exponential random variables (truncexp(γ_l; min($s_l, t_k(j)$), $t_{k+1}(j)$)) that correspond to customers from M_2^j and $|M_2^j|$ stands for a number of customers in M_2^j. Independence between stochastic components implies formula (3.28).

Now consider

$$\mathrm{P}\{B_3/B_1,B_2,\tilde{A}(\tilde{t}_k)=C_1\} = \mathrm{P}\{B_3/\tilde{A}(\tilde{t}_k)=C_1\} =$$
$$= \prod_{l\in M_3}\frac{p_{l0}+\sum_{j\notin d_b}p_{lj}\exp\{-\gamma_{lj}(t-s_l)\}+\sum_{j\in d_b}p_{lj}\exp\{-\gamma_{lj}(t_{k+1}(j)-s_l)_+\}}{p_{l0}+\sum_{j\notin d_b}p_{lj}\exp\{-\gamma_{lj}(t-s_l)\}+\sum_{j\in d_b}p_{lj}\exp\{-\gamma_{lj}(t_k(j)-s_l)_+\}} \quad (3.30)$$

Here p_{lj}, $l\in M_3, j\in d_b$, are the probabilities that customer l switches to station j and γ_{lj} are the corresponding transition rates. The proof of (3.30) will be given in the end of the subsection.

Finally,

$$\mathrm{P}\{B_4/B_1,B_2,B_3,\tilde{A}(\tilde{t}_k)=C_1\} = \mathrm{P}\{B_4^1/\tilde{A}(\tilde{t}_k)=C_1\}\times\mathrm{P}\{B_4^2/B_2,\tilde{A}(\tilde{t}_k)=C_1\} =$$
$$= \prod_{j\in d_b}\prod_{l\in M_j}(1-\exp\{-\gamma_{lj}(t_{k+1}(j)-\max(t_k(j),s_l))_+\}) \times$$
$$\times \frac{p_{lj}\exp\{-\gamma_{lj}(t_k(j)-s_l)_+\}}{p_{l0}+\sum_{j\notin d_b}p_{lj}\exp\{-\gamma_{lj}(t-s_l)\}+\sum_{m\in d_b}p_{lm}\exp\{-\gamma_{lj}(t_k(m)-s_l)_+\}} \times$$
$$\times \mathrm{P}\{\hat{A}_1^j>A_{m_j}^j,\hat{A}_2^j>A_{m_j}^j,\ldots,\hat{A}_{|M_4^j|}^j>A_{m_j}^j/A_1^j<A_2^j<\ldots<A_{m_j}^j\}, \quad (3.31)$$

where $\hat{A}_i^j$ are arrivals of invisible customers to station j, $|M_4^j|$ is a number of customers in M_4^j and A_i^j, $,1\le i\le m_j$, are arrivals of visible customers to station j. Formula (3.31) will be discussed in the end of the subsection. Summarizing, the transition probabilities of the Markov chain are calculated using (3.26), (3.27), (3.29), (3.30) and (3.31).

Algorithm.

We can use the general algorithm for HMM problem that was introduced in Section 1 (see Algorithm 1.1 on page 28 and Algorithm 1.2 on page 32). This algorithm solves also the General Interpolation problem since estimates of cumulative arrivals number are derived for intermediate time points between starts of the current busy periods and t. Queue estimates can be derived from estimates of cumulative arrivals number.

The complexity of the algorithm depends on the number of stations, lengths of busy periods and number of invisible customers. It increases very rapidly when the number of invisible customers increases and, therefore, it is not practical to apply to "large" systems, and the need arises for approximations. We leave this topic for a possible future research.

Another problem is that in the case of exponential transitions the number of invisible customers increases with time since we never can be absolutely sure that a customer left the network. Hence, in practice, we need to truncate the set of invisible customers, removing customers that have not registered for a long time.

Comment on formulae. First we derive formula (3.30). Let D_l^1 and D_l^2, $l\in M_3$, stand for events "invisible customer l has not appeared on any station until $\tilde{t}_{k+1}$" and "invisible customer l has not appeared on any station until $\tilde{t}_k$" correspondingly.

$$\mathrm{P}\{B_3/\tilde{A}(\tilde{t}_k)=C_1\} = \prod_{l\in M_3}\mathrm{P}\{D_l^1/D_l^2\} =$$

$$= \prod_{l\in M_3} \sum_{j\in\mathcal{D}_0} \mathrm{P}\{D_l^1/D_l^2, R_l = j\}\mathrm{P}\{R_l = j/D_l^2\} \;=$$

$$= \prod_{l\in M_3} \frac{\sum_{j\in\mathcal{D}_0} \mathrm{P}\{D_l^1/D_l^2, R_l = j\}\mathrm{P}\{D_l^2/R_l = j\}p_{lj}}{\sum_{j\in\mathcal{D}_0} \mathrm{P}\{D_l^2/R_l = j\}p_{lj}} \;= \tag{3.32}$$

$$= \prod_{l\in M_3} \frac{p_{l0} + \sum_{j\notin d_b} p_{lj}\exp\{-\gamma_{lj}(t - s_l)\} + \sum_{j\in d_b} p_{lj}\exp\{-\gamma_{lj}(t_{k+1}(j) - s_l)_+\}}{p_{l0} + \sum_{j\notin d_b} p_{lj}\exp\{-\gamma_{lj}(t - s_l)\} + \sum_{j\in d_b} p_{lj}\exp\{-\gamma_{lj}(t_k(j) - s_l)_+\}}.$$

Here (3.32) follows from Bayes formula.

Now consider formula (3.31). Its first part (expression for $\mathrm{P}\{B_4^1/\tilde{A}(\tilde{t}_k) = C_1\}$) can be proved in the same way as (3.30). As for the second part,

$$\mathrm{P}\{\hat{A}_1^j > A_{m_j}^j, \hat{A}_2^j > A_{m_j}^j, \ldots, \hat{A}_{|M_4^j|}^j > A_{m_j}^j / A_1^j < A_2^j < \ldots < A_{m_j}^j\} \;=$$

$$= \frac{\mathrm{P}\{\hat{A}_1^j > A_{m_j}^j, \hat{A}_2^j > A_{m_j}^j, \ldots, \hat{A}_{|M_4^j|}^j > A_{m_j}^j; A_1^j < A_2^j < \ldots < A_{m_j}^j\}}{\mathrm{P}\{A_1^j < A_2^j < \ldots < A_{m_j}^j\}}, \tag{3.33}$$

where all random variables in consideration are truncated exponentials and the denominator can be calculated using technique from Subsection 1.5. The numerator of (3.33) is equal to

$$\int_0^t f_1(y_1)\int_{y_1}^t f_2(y_2)\ldots\int_{y_{m_j-1}}^t f_{m_j}(y_{m_j})\int_{y_{m_j}}^t g_1(z_1)\int_{y_{m_j}}^t g_2(z_2)\ldots\int_{y_{m_j}}^t g_{|M_4^j|}(z_{|M_4^j|}),$$

where f_i and g_i are densities of truncated exponential random variables (f_i of arrival times of visible customers and g_i of invisible customers). Some approximation technique must be used for calculation of this integral.

3.5 Immediate Transitions, Unknown External Arrivals.

Notation and Definitions. Suppose that a customer starts service at t_{ki}^j. If he arrived from station 0 he will be called *external customer*, otherwise, he will be called *internal customer*.

Let $d_b^0(t) \subseteq d_b(t)$ denote a subset of the set of the busy stations: station $j \in d_b^0(t)$ if the last service that starts at station j before t corresponds to an external customer.

As in section 3.3, we introduce a d-dimensional stochastic process $C = \{C(t),\ t \geq 0\}$ but the definition is a little different. For an idle station j we put $C_j(t) = t$. If a station is busy but no internal customer started service during the current busy period, $C_j(t) = t$ again. Otherwise, $C_j(t)$ stands for arrival time of the last internal customer, which started service on station j.

The **available stations process** for IR-times t_{ml}^n

$$\mathcal{M}_t(t_{ml}^n) = \{\{j : C_j(t) < t_{ml}^n\} \bigcup \{0\}\}.$$

is defined identically to Section 3.3.

The conception of the invisible customers is applied only to internal customers. Naturally, all internals customers, which occupy the queues at time t, are invisible. Therefore, we can represent the queue at station j as

$$Q_j(t) = Q_j^{\alpha}(t) + Q_j^{\beta}(t), \tag{3.34}$$

where $Q_j^{\alpha}(t)$ includes external customers and $Q_j^{\beta}(t)$ includes invisible customers.

Representation 5 of $\mathbf{E_t}$ via Stochastic Components.
Immediate Transitions, Unknown External Arrivals.

5.1(S) Services that terminated up to t are known.
$\xi^{\mathbf{j}}_{\mathbf{ki}} = \mathbf{t^j_{k,i+1}} - \mathbf{t^j_{ki}}$, $j \in \mathcal{D}$, $1 \le k \le b_j(t)$,
$0 \le i \le m_j(t) - 1$, for $j \in d_b(t)$, $k = b_j(t)$,
$0 \le i \le n_k^j$, otherwise.

5.2(S) Last service at a busy station have not terminated until t.
For $j \in d_b(t)$ $\xi^{\mathbf{j}}_{\mathbf{b_j(t),m_j(t)}} > \mathbf{t} - \mathbf{t^j_{b_j(t),m_j(t)}}$.

5.3(A) The start of a busy period coincides with the first arrival.
For $(j,k,0) \ni R_S^{-1}(j,k,0) = 0$,
$\mathbf{A^j_{k0}} = \mathbf{t^j_{k0}}$.
(A^j_{k0} is an external arrival.)

5.4a(A) Truncation rule for complete busy periods.
For $(j,k,i) \ni i \ge 1$, $R_S^{-1}(j,k,0) = 0$,
$\mathbf{A^j_{ki}} \le \mathbf{t^j_{ki}}$.
$1 \le k \le b_j(t)$ for $j \notin d_b(t)$, and $1 \le k \le b_j(t) - 1$ for $j \in d_b(t)$.

5.4b(A) Truncation rule for incomplete busy periods.
For $(j, b_j(t), i) \ni j \in d_b(t)$, $i \ge 1$, $R_S^{-1}(j, b_j(t), i) = 0$,
$\mathbf{A^j_{b_j(t),i}} \le \mathbf{t^j_{b_j(t),i}}$,
$1 \le i \le m_j(t)$.

5.5a(A) FCFS discipline for complete busy periods.
$\mathbf{t^j_{k0}} \le \mathbf{A^j_{k1}} \le \ldots \le \mathbf{A^j_{k,n^j_k}}$,
$1 \le k \le b_j(t)$ for $j \notin d_b(t)$, and $1 \le k \le b_j(t) - 1$ for $j \in d_b(t)$,
where some of A^j_{ki} are unknown external arrivals and some are known internal ones.

5.5b(A) FCFS discipline for incomplete busy periods.
$\mathbf{t^j_{b_j(t),0}} \le \mathbf{A^j_{b_j(t),1}} \le \ldots \le \mathbf{A^j_{b_j(t),m_j(t)}}$, $\mathbf{j \in d_b(t)}$.

5.6(R) Complete information about the switches of the visible customers.

For $(n,m,l) \in V_R(t)$

$\mathbf{R^n_{ml} = j}$.

5.7(RA) Partial information about the invisible routing.

5.7a For $(n,m,l) \in I_R(t)$:

$\mathbf{R^n_{ml} = j \Rightarrow j \in \mathcal{M}_t(t^n_{ml})}$.

5.7b If $j \in d^0_b(t)$:

$\mathbf{R^n_{ml} = j \Rightarrow t^n_{ml} \geq A^j_{b_j(t),m_j(t)}}$.

Equivalence of Representations 2' and 5.

The equivalence of events 2.1'-2.6' to events 5.1-5.6 is straightforward. The equivalence of event 2.7'=2.7 and event 5.7 follows from the definition of $\mathcal{M}_t(t^n_{ml})$ and the equality $\bar{A}^n_{ml} = t^n_{ml}$ which prevails in the case of immediate transitions.

Queue estimation at time t. Relevant information. According to formula (3.34)

$$\mathrm{E}[Q_j(t)/E_t] = \mathrm{E}[Q^\alpha_j(t)/E_t] + \mathrm{E}[Q^\beta_j(t)/E_t]. \tag{3.35}$$

Consider the two terms of (3.35) separately. First,

$$\begin{aligned} \mathrm{E}[Q^\beta_j(t)/E_t] &= \sum_{t^n_{ml} \in I_R(t)} \mathrm{P}\{R^n_{ml} = j/E_t\} \\ &= \sum_{t^n_{ml} \in I_R(t)} \mathrm{P}\{R^n_{ml} = j/E^{7b}_t, E^{7a}_t, E^{5b}_t, E^{4b}_t\}, \end{aligned} \tag{3.36}$$

where $E^{7b}_t, E^{7a}_t, E^{5b}_t, E^{4b}_t$ correspond to events 5.7b, 5.7a, 5.5b and 5.4b respectively. In formula (3.36) we condition on events containing $A^j_{b_j(t),m_j(t)}$ and R^n_{ml}, $(n,m,l) \in I_R(t)$. Other events are irrelevant. In particular, we omit events 5.4a and 5.5a since external arrivals from different busy periods are independent.

As for the second term

$$\mathrm{E}[Q^\alpha_j(t)/E_t] = \tag{3.37}$$

$$= \mathrm{E}[Q^\alpha_j(t)/E^{7c}_t, E^{7b}_t, E^{7a}_t, E^{5b}_t, E^{4b}_t] \tag{3.38}$$

$$= \mathrm{E}[\#\text{ (Poisson arrivals in } [A^j_{b_j(t),m_j(t)}, t])\ /E^{7c}_t, E^{7b}_t, E^{7a}_t, E^{5b}_t, E^{4b}_t]. \tag{3.39}$$

Mathematical Formulation of the Estimation Problem. Our objective is to simplify notation (as in Section 3.4) and to reduce the relevant information also.

Note that the external arrival times which took place before $C_j(t)$, are conditionally independent of those which happened after $C_j(t)$ given $C_j(t)$ is known (and it is known).

Therefore, the parts of events 4.4b and 4.5b that correspond to arrivals before $C_j(t)$ are irrelevant.

Let t_0^j denote the time of the service start of an internal customer, which arrived at $C_j(t)$ (we again omit the time index whenever convenient). The following registrations at station j until t are denoted by $t_0^j < t_1^j < \ldots < t_{m_j}^j$ (as though the busy period starts at t_0^j).

The notation for K invisible customers is preserved from the previous section.

Set $A_1^j, \ldots, A_{m_j}^j$ stand for the successive external arrivals to station j, which correspond to the service starts $t_1^j, \ldots, t_{m_j}^j$. From the memoryless property of the Poisson process we can assume that they are the points of the Poisson process starting at t_0^j.

Then the relevant events are given by:

5.4b $A_1^j \le t_1^j, \ldots, A_{m_j}^j \le t_{m_j}^j, \quad j \in d_b^0$;

5.7a $R_k = j \Rightarrow j \in \mathcal{M}_k, \quad 1 \le k \le K$.

5.7b For $j \in d_b^0 \quad R_k = j \Rightarrow A_{m_j}^j \le s_k$.

Events 5.5b is "installed" in the definition of $A_1^j, \ldots, A_{m_j}^j$.

Queue estimation at time t. We start with some additional definitions, proceed with the description of the algorithm and, finally, prove its correctness and give some comments. Unfortunately, the algorithm is static (we assume that t is fixed) and there seems to be no way to recalculate the estimate for time $t+h$ with essential economy in the computational effort.

Let B denote the event in 5.4b. Then

$$B = \bigcap_{j \in d_b^0} B_j,$$

where $B_j = \{A_1^j \le t_1^j, \ldots, A_{m_j}^j \le t_{m_j}^j\}, \ j \in d_b^0$. We also use notation

$$B^j \stackrel{\text{def}}{=} B \setminus B_j.$$

Consider event 5.7b. It can be represented as

$$D = \bigcap_{k=1}^{K} \bigcap_{j \in d_b^0} D_{kj}, \tag{3.40}$$

where event

$$D_{kj} = \{R_k = j \Rightarrow A_{m_j}^j \le s_k\} = \{R_k \ne j\} \bigcap \{A_{m_j}^j \le s_k\}. \tag{3.41}$$

Note that D is the intersection of events, organized in the matrix form (the rows correspond to the invisible customers and the columns correspond to the stations). It is important that events, which do not belong to the same row or to the same column, are independent. We shall need also

$$D^{lm} \stackrel{\text{def}}{=} \bigcap_{k\neq l}\bigcap_{j\neq m} D_{kj}; \quad D^{l*} \stackrel{\text{def}}{=} \bigcap_{k\neq l}\bigcap_{j\in d_b^0} D_{kj}; \quad D^{*m} \stackrel{\text{def}}{=} \bigcap_{k=1}^{K}\bigcap_{j\neq m} D_{kj}.$$

Here event D^{lm} is event D after removal of the l-th row and the m-th column. If only the l-th row was removed we get D^{l*} and, finally, D^{*m} corresponds to removal of the m-th column.

In order to deal with event 5.7a we define the conditional routing matrix of the type "invisible customers → stations". It is denoted by $\tilde{R} = (\tilde{r}_{mj},\ 1 \le m \le K,\ j \in \mathcal{D}_0)$, where

$$\tilde{r}_{mj} = 1_{\{j\in\mathcal{M}_m\}}\frac{p_{N_m,j}}{\sum_{l\in\mathcal{M}_m} p_{N_m,l}} \tag{3.42}$$

(see also formula (3.20) from Section 3.3). The assumption that the routing of the invisible customers is governed by matrix $\tilde{R}$ (instead of the routing matrix P) is equivalent to conditioning on event 5.7a. Notation $P_{\tilde{R}}$ is used for the conditional probabilities of this type.

Then we define

$$\omega_{kj} = \mathrm{P}\{A^j_{m_j} \le s_k / B_j\}, \tag{3.43}$$

and, finally, introduce events F_{kj}, $1 \le k \le K$, $j \in \mathcal{D}$, which mean "the invisible customer number k was the first invisible customer that switched to station j". Event F_{0j}, $, j \in \mathcal{D}$, denotes event "no invisible customer switched to station j".

Algorithm 3.3. Real-Time Estimation.
Immediate Transitions, Unknown External Arrivals.

1°. Basic queue decomposition.

$$\mathrm{E}[Q_j(t)/E_t] = \mathrm{E}_{\tilde{R}}[Q^\alpha_j(t)/B; D] + \mathrm{E}_{\tilde{R}}[Q^\beta_j(t)/B; D]. \tag{3.44}$$

2°. Calculation of ω_{kj}.

$$\begin{aligned}\omega_{kj} &= \frac{\mathrm{P}\{A^j_{m_j} \le s_k; B_j\}}{\mathrm{P}\{B_j\}} = \\ &= \frac{\mathrm{P}\{A^j_1 \le \min(s_k, t^j_1); A^j_2 \le \min(s_k, t^j_2); \ldots; A^j_{m_j} \le \min(s_k, t^j_{m_j})\}}{\mathrm{P}\{B_j\}}.\end{aligned} \tag{3.45}$$

The numerator of (3.45) and $\mathrm{P}\{B_j\}$ are calculated using formulae (3.2)–(3.5) from Subsection 3.2.

3°. Calculation of $\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj}\}$.

$$\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{1j}\} = \tilde{r}_{1j};\ \ \mathrm{P}_{\tilde{\mathrm{R}}}\{F_{0j}\} = \prod_{l=1}^{K}(1-\tilde{r}_{lj});\ \ \mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj}\} = \tilde{r}_{kj}\prod_{l=1}^{k-1}(1-\tilde{r}_{lj}),\ \ 2 \le k \le K. \quad (3.46)$$

4°. Calculation of the internal queue.

$$\mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^{\beta}(t)/D;B] = \frac{\sum_{k=1}^{K}\mathrm{P}_{\tilde{\mathrm{R}}}\{R_k = j; D/B\}}{\mathrm{P}_{\tilde{\mathrm{R}}}\{D/B\}}. \quad (3.47)$$

The denominator of (3.47) is given by the recursive formula:

$$\mathrm{P}_{\tilde{\mathrm{R}}}\{D/B\} = \sum_{l\in d_b^0}\tilde{r}_{1l}\omega_{1l}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{1l}/B^l\} + \sum_{l\notin d_b^0}\tilde{r}_{1l}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{1*}/B\}. \quad (3.48)$$

For $k = 1$ formula (3.48) turns to

$$\mathrm{P}_{\tilde{\mathrm{R}}}\{D/B\} = \sum_{l\in d_b^0}\tilde{r}_{1l}\omega_{1l} + \sum_{l\notin d_b^0}\tilde{r}_{1l}. \quad (3.49)$$

Finally, the numerator of (3.47) is calculated using

$$\mathrm{P}_{\tilde{\mathrm{R}}}\{R_k = j; D/B\} = \tilde{r}_{kj}\omega_{kj}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{k*}/B^j; B_j\bigcap\{A_{m_j}^j \le s_k\}\}. \quad (3.50)$$

The last conditional probability can be computed recursively using formula (3.48).

5°. Calculation of the external queue.

For $j \in d_b \setminus d_b^0$

$$\mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^{\alpha}(t)/B;D] = \lambda_j(t - C_j(t)), \quad (3.51)$$

where $\lambda_j = \lambda p_{0j}$, λ is an external arrival rate and the conditional distribution of the external queue is Poisson with parameter $\lambda_j(t - C_j(t))$.

For $j \in d_b^0$

$$\begin{aligned}\mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^{\alpha}(t)/B;D] &= \frac{1}{\mathrm{P}_{\tilde{\mathrm{R}}}\{D/B\}}[\sum_{k=1}^{K}\mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^{\alpha}(t)/B_j; A_{m_j}^j \le s_k]\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj}\}\omega_{kj}\mathrm{P}_{\tilde{\mathrm{R}}_k}\{D^{kj}/B^j\} + \\ &+ \mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^{\alpha}(t)/B_j]\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{0j}\}\mathrm{P}_{\tilde{\mathrm{R}}_0}\{D^{*j}/B^j\}]. \quad (3.52)\end{aligned}$$

Here $\mathrm{E}_{\tilde{\mathrm{R}}}[Q_j^{\alpha}(t)/B_j; A_{m_j}^j \le s_k]$ is computed using Algorithm 3.1 and $\tilde{R}_k = \tilde{R}/F_{kj}$, i. e. in (3.42) we must replace $\mathcal{M}_m$ by $\mathcal{M}_m \setminus \{j\}$ for $m < k$, if $k \neq 0$, and for $1 \le m \le K$, if $k = 0$.

Proof of correctness and comments.

1°. Formula (3.44) is equivalent to (3.35).

2°. Formula (3.45) follows from the definition of ω_{kj}. Note that we can represent the numerator of (3.45) as

$$\mathrm{P}\{A_1^j \le t_1^j; A_2^j \le t_2^j; \ldots; A_i^j \le t_i^j; A_{i+1}^j \le s_k; \ldots; A_{m_j}^j \le s_k\},$$

where $i = \max\{l : t_l^j < s_k\}$. Such probabilities are computed using formulae (3.2)-(3.5) from Subsection 3.2. Expressions $g(t_1, \ldots, t_m)$ from this formulae are known from the calculation of $\mathrm{P}\{B_j\}$, therefore we save some computational effort.

3°. The following straightforward representation of events F_{kj} implies formula (3.46):

$$F_{1j} = \{R_1 = j\}; \quad F_{0j} = \bigcap_{l=1}^{K}\{R_l \neq j\}; \quad F_{kj} = \{\bigcap_{l=1}^{k-1}\{R_l \neq j\}\}\bigcap\{R_k = j\}, \quad 2 \leq k \leq K.$$

4°. Formula (3.47) follows from

$$Q_j^{\beta}(t) = \sum_{k=1}^{K} 1_{\{R_k = j\}}.$$

In order to derive formula (3.48), we condition on the switch of the first invisible customer (recall that $s_1 < s_2 < \ldots < s_K$) :

$$\mathrm{P}_{\tilde{\mathrm{R}}}\{D/B\} = \sum_{l \in \mathcal{D}_0} \mathrm{P}_{\tilde{\mathrm{R}}}\{D/B; R_1 = l\}\tilde{r}_{1l}.$$

The following calculation is typical for this section. For the first time, it is carried out in detail, and further similar calculations will be presented in short.

If $l \in d_b^0$

$$\begin{aligned}
\mathrm{P}_{\tilde{\mathrm{R}}}\{D/B; R_1 = l\} &= \mathrm{P}_{\tilde{\mathrm{R}}}\{\bigcap_{k=1}^{K}\bigcap_{j \in d_b^0}\{\{R_k \neq j\}\bigcup\{A_{m_j}^j \leq s_k\}\}/B; R_1 = l\} \\
&= \frac{\mathrm{P}_{\tilde{\mathrm{R}}}\{A_{m_l}^l \leq s_1; \bigcap\bigcap_{(k,j)\neq(1,l)}\{\{R_k \neq j\}\bigcup\{A_{m_j}^j \leq s_k\}\}; R_1 = l; B\}}{\mathrm{P}_{\tilde{\mathrm{R}}}\{R_1 = l; B\}} \\
&= \frac{\mathrm{P}_{\tilde{\mathrm{R}}}\{A_{m_l}^l \leq s_1; \bigcap_{k=2}^{K}\bigcap_{j \neq l} D_{kj}; R_1 = l; B^l; B_l\}}{\mathrm{P}_{\tilde{\mathrm{R}}}\{R_1 = l; B^l; B_l\}} \qquad (3.53) \\
&= \frac{\mathrm{P}_{\tilde{\mathrm{R}}}\{A_{m_l}^l \leq s_1; B_l\}\mathrm{P}_{\tilde{\mathrm{R}}}\{R_1 = l\}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{1l}; B^l\}}{\mathrm{P}_{\tilde{\mathrm{R}}}\{R_1 = l\}\mathrm{P}_{\tilde{\mathrm{R}}}\{B_l\}\mathrm{P}_{\tilde{\mathrm{R}}}\{B^l\}} \\
&= \mathrm{P}_{\tilde{\mathrm{R}}}\{A_{m_l}^l \leq s_1/B_l\}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{1l}/B^l\} \\
&= \omega_{1l}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{1l}/B^l\}.
\end{aligned}$$

Formula (3.53) prevails since event $\{R_1 = l\}$ implies $\{R_1 \neq j, \ j \neq l\}$ and event $A_{m_l}^l \leq s_1$ implies $A_{m_l}^l \leq s_k, \ 2 \leq k \leq K$.

Then, in the same way, we prove that for $j \in d_b \setminus d_b^0$

$$\begin{aligned}
\mathrm{P}_{\tilde{\mathrm{R}}}\{D/B; R_1 = l\} &= \mathrm{P}_{\tilde{\mathrm{R}}}\{\bigcap_{k=2}^{K}\bigcap_{j \in d_b^0} D_{kj}/B\} = \\
&= \mathrm{P}_{\tilde{\mathrm{R}}}\{D^{1*}/B\}.
\end{aligned}$$

As for formula (3.49), note that if $k = 1$

$$\begin{aligned} \mathrm{P}_{\tilde{\mathrm{R}}}\{D/B\} &= \mathrm{P}_{\tilde{\mathrm{R}}}\{\bigcap_{j\in d_b^0} D_{1j}/B\} = \\ &= \sum_{l\in d_b} \mathrm{P}_{\tilde{\mathrm{R}}}\{R_1 = l; \bigcap_{j\in d_b^0} D_{1j}/B\} = \\ &= \sum_{l\in d_b^0} \tilde{r}_{1l}\omega_{1l} + \sum_{l\neq d_b^0} \tilde{r}_{1l}. \end{aligned}$$

Note, that all these computations are in some way similar to the computation of a matrix determinant.

Finally, (3.50) is derived using the following calculation:

$$\begin{aligned} \mathrm{P}_{\tilde{\mathrm{R}}}\{R_k = j; D/B\} &= \mathrm{P}_{\tilde{\mathrm{R}}}\{R_k = j; \bigcap_{l\neq k}\bigcap_{j\in d_b^0} D_{kj}; A^j_{m_j} \le s_k/B\} \\ &= \tilde{r}_{kj}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{k*}; A^j_{m_j} \le s_k/B\} \\ &= \tilde{r}_{kj}\omega_{kj}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{k*}/A^j_{m_j} \le s_k; B\} \\ &= \tilde{r}_{kj}\omega_{kj}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{k*}/B^j; B_j \bigcap \{A^j_{m_j} \le s_k\}\}. \end{aligned}$$

5°. Formula (3.51) prevails since only external customers that arrived after the known time $C_j(t)$ stand in queue at time t at station j.

For $j \in d_b^0$ we condition on the first internal customer which arrived to station j:

$$\mathrm{E}_{\tilde{\mathrm{R}}}[Q^\alpha_j(t)/B; D] = \sum_{k=0}^{K} \mathrm{E}_{\tilde{\mathrm{R}}}\{Q^\alpha_j(t)/F_{kj}; B; D\}\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj}/B; D\}. \tag{3.54}$$

If $k \neq 0$,

$$\begin{aligned} \mathrm{E}_{\tilde{\mathrm{R}}}\{Q^\alpha_j(t)/F_{kj}; B; D\} &= \mathrm{E}_{\tilde{\mathrm{R}}}\{Q^\alpha_j(t)/R_k = j; \bigcap_{l<k} R_l \neq j; B; D\} \\ &= \mathrm{E}_{\tilde{\mathrm{R}}}\{Q^\alpha_j(t)/R_k = j; \bigcap_{l<k}\{R_l \neq j\}; B_j \bigcap B^j; A^j_{m_j} < s_k; D^{kj}\} \qquad (3.55) \\ &= \mathrm{E}_{\tilde{\mathrm{R}}}\{Q^\alpha_j(t)/B_j; A^j_{m_j} < s_k\}. \qquad (3.56) \end{aligned}$$

Formula (3.55) prevails since $\{R_l \neq j\}$ implies D_{lj}, $l < k$ and $\{A^j_{m_j} < s_k\}$ implies D_{lj}, $l > k$. Formula (3.56) is true since $Q^\alpha_j(t)$ is the number of Poisson arrivals in $[A^j_{m_j}, t]$, which depends only on the events that are left in the condition.

If $k = 0$,

$$\begin{aligned} \mathrm{E}_{\tilde{\mathrm{R}}}\{Q^\alpha_j(t)/F_{0j}; B; D\} &= \mathrm{E}_{\tilde{\mathrm{R}}}\{Q^\alpha_j(t)/\bigcap_{l\in d_b^0}\{R_l \neq j\}; B; D\} = \\ &= \mathrm{E}_{\tilde{\mathrm{R}}}\{Q^\alpha_j(t)/\bigcap_{l\in d_b^0}\{R_l \neq j\}; B; D^{*j}\} = \\ &= \mathrm{E}_{\tilde{\mathrm{R}}}\{Q^\alpha_j(t)/B_j\}. \end{aligned}$$

Now consider the second term of (3.54):

$$\begin{aligned} \mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj}; D/B\} &= \mathrm{P}_{\tilde{\mathrm{R}}}\{D/F_{kj}; B\}\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj}/B\} = \\ &= \mathrm{P}_{\tilde{\mathrm{R}}}\{D/F_{kj}; B\}\mathrm{P}_{\tilde{\mathrm{R}}}\{F_{kj}\}, \end{aligned}$$

since events F_{kj} and B are independent. Then if $k \neq 0$:

$$\begin{aligned} \mathrm{P}_{\tilde{\mathrm{R}}}\{D/F_{kj}; B\} &= \mathrm{P}_{\tilde{\mathrm{R}}}\{D/\bigcap_{l<k}\{R_l \neq j\}; R_k = j; B\} = \\ &= \mathrm{P}_{\tilde{\mathrm{R}}}\{A^j_{m_j} \leq s_k; D^{kj}/\bigcap_{l<k}\{R_l \neq j\}; R_k = j; B_j \bigcap B^j\} = \\ &= \omega_{kj}\mathrm{P}_{\tilde{\mathrm{R}}}\{D^{kj}/\bigcap_{l<k}\{R_l \neq j\}; B^j\} = \\ &= \omega_{kj}\mathrm{P}_{\tilde{\mathrm{R}}_k}\{D^{kj}/B^j\}. \end{aligned}$$

Similarly, for $k = 0$:

$$\begin{aligned} \mathrm{P}_{\tilde{\mathrm{R}}}\{D/F_{0j}; B\} &= \mathrm{P}_{\tilde{\mathrm{R}}}\{D^{*j}/F_{0j}; B\} \\ &= \mathrm{P}_{\tilde{\mathrm{R}}_0}\{D^{*j}/B^j\}. \end{aligned}$$

General Interpolation. Here we give an outline of the solution of the General Interpolation problem. Suppose we need to calculate $\mathrm{E}[Q_j(s)/E_t]$, $s < t$. Two special cases must be considered.

If $s \leq C_j(t)$ all internal arrivals to station j until s are known. The external arrivals can be estimated using HMM approach from Subsection 1.2.

Otherwise, if $s > C_j(t)$, we have to estimate arrivals in $[C_j(s), C_j(t)]$ and in $[C_j(t), s]$ separately. Arrivals that took place during the first interval can be estimated using HMM approach again. As for the second interval, the technique of this subsection must be used. External arrivals are estimated using formulae (3.51) and (3.52), where we replace t by s. Internal queue is estimated using (3.47) where only those internal customers whose transition-start times $s_k \leq s$ are included in the sum.

3.6 Exponential transitions. Unknown External Arrivals.

Hidden Markov Model in the spirit of Subsection 3.4 can be constructed in this case. The state space of HMM and, consequently, notation will be very cumbersome. Indeed, the state space will be infinite, due to an unrestricted number of external invisible customers. Any practical implementation of the algorithm must involve a truncation of the state space. We leave this to a possible future research.

Chapter III

Applications.

1 Data and Tools.

1.1 The Service System.

The service system analyzed is a typical branch of a commercial bank. Measurements from 14 days of work (2 weeks + 2 days) were collected. The data covered service stations, servers' positions, ID's of customers, arrival times, times of service starts and terminations etc.

As an example, consider the data in Table 1. Each line corresponds to a specific service task. During a single service, several service tasks can be performed. Note customer 1021 that goes through two operations (tasks) at service position 5. Service position corresponds to a specific server. At a multi-server station there are several service positions.

Times of service starts, service terminations and external arrivals to the bank entrance (station 0) are recorded directly by the measurement system but internal arrival times are not. The system assumes that an arrival time of a customer to the first internal station on his route is equal to his external arrival time. Each successive arrival time to a station is assumed to be equal to the preceding service termination time. In essence, the assumption of immediate transitions between stations is implicitly used by the measurement system. (This assumption seems to be appropriate for our bank, where walking distances are short.)

Now consider the other assumptions, presented in Subsection 3.1 of the Introduction. FCFS queueing discipline applies in general but there can be exceptions. It is reasonable to assume that arrivals to the system constitute a non-homogeneous Poisson process, whose rate is approximately constant during short intervals, for example half an hour. As a first approximation, we may assume a Markovian switching mechanism. (Although the routing probabilities may depend on the path of a customer, hence a multi-type models might be more appropriate.)

Most problems in our data analysis have been connected with violations of the work-

conserving principle. Indeed, a human server often stops working for some periods of time even when there are customers waiting for service.

This data set is rather convenient for our applications, since it covers the features of our simplest model: transitions between stations are immediate and external arrivals are known. Then we can suppose that, for example, external arrivals are unknown, run our algorithms on incomplete data and, finally, compare the queue estimate with the exact queue.

The main problems of the data are violations of work-conserving, FCFS discipline and also some kinds of measurement errors (see Section 2.2 for examples).

1.2 Description of Software.

In order to apply our theoretical results to real data, several programs, written in C, were created. The main are:

- Program QIE1. Busy-period interpolation. Immediate transitions, known external arrivals.
- Program QIE2. Busy-period interpolation. Immediate transitions, unknown external arrivals.

Program QIE1. An input of the program has the form of Table 2. The data for each station is ordered according to service starts. Special programs are used to convert initial data files (Table 1 shows a part of such a file) to the format of Table 2. In particular, the convertion unites service tasks performed successively at one service position.

The example of an output of QIE1 is presented in Table 3. It includes 7 columns of numbers: the estimation time, the main estimate of the queue (further referenced as exact queue), the lower estimate of the queue, the cumulative number of arrivals, the cumulative number of departures from the queue, the cumulative number of departures from the station and the number of working servers.

The main estimate of the queue at time t is calculated as the number of arrivals minus number of departures from the queue. If the assumption of immediate transitions between stations really stands, this estimate is equal to the exact queue. But sometimes a customer can "disappear" for an hour and only then return to the next station. The lower estimate tries to treat such cases. However, it is more appropriate for large bureaucratic institutions or medical clinics where customers wait for hours.

The output of QIE1 includes also a file with simple statistics. It comprises arrival rates, average service times and average waiting times.

Program QIE2. Program QIE2 assumes a homogeneous arrival rate during busy periods.

The input of QIE2 has the same form as the input of QIE1 (see Table 2) but the third row (arrival times) is not used by the program. Table 4 illustrates the output. Queue estimate is calculated at the points of service starts and terminations (it is linear between these points). In addition, 90% quantile of the conditional distribution of the queue and σ(estimate) are calculated.

Table 1. Example of Data File.

DATE	CUST ID	ARRIVAL	SERVICE BEG	SERVICE END	SERVICE TIME	WAIT TIME	SERVICE POS	SERVER ID	TASK TYPE
040293	1005	8:03:11	8:03:46	8:14:17	10:31	0:35	10	10	27
040293	1006	8:05:23	8:11:33	8:12:45	1:12	6:09	12	12	29
040293	1007	8:07:01	8:07:16	8:09:47	2:32	0:14	12	12	11
040293	1008	8:07:10	8:09:34	8:15:51	6:18	2:24	6	6	20
040293	1009	8:07:19	8:09:47	8:11:33	1:45	2:28	12	12	11
040293	1006	8:12:45	8:19:32	8:19:44	0:12	6:48	5	5	29
040293	1010	8:14:04	8:14:17	8:15:51	1:34	0:13	10	10	27
040293	1008	8:15:51	8:15:51	8:15:55	0:04	0:00	10	10	20
040293	1006	8:19:44	8:19:44	8:44:25	24:41	0:00	5	5	12
040293	1012	8:26:55	8:28:37	8:30:53	2:16	1:42	12	12	11
040293	1013	8:27:37	8:33:31	8:34:16	0:45	5:55	10	10	27
040293	1014	8:29:05	8:30:14	8:31:33	1:19	1:10	6	6	20
040293	1014	8:31:33	8:35:36	8:35:39	0:03	4:03	12	12	20
040293	1016	8:33:52	8:34:16	8:44:49	10:33	0:24	10	10	27
040293	1014	8:35:39	8:35:39	8:41:12	5:33	0:00	12	12	11
040293	1018	8:39:01	8:43:29	8:50:31	7:02	4:28	12	12	11
040293	1019	8:39:57	8:41:12	8:43:29	2:18	1:15	12	12	11
040293	1021	8:43:20	8:50:31	8:53:07	2:36	7:11	12	12	11
040293	1022	8:47:21	8:53:07	8:58:18	5:11	5:46	12	12	11
040293	1023	8:47:24	8:49:06	8:54:32	5:26	1:42	6	6	20
040293	1024	8:50:07	8:51:54	8:54:32	2:38	1:48	10	10	27
040293	1025	8:50:58	8:58:18	8:59:29	1:11	7:20	12	12	11
040293	1021	8:53:07	8:54:52	8:55:05	0:13	1:45	5	5	11
040293	1027	8:53:11	8:59:29	9:12:33	13:04	6:18	12	12	11
040293	1023	8:54:32	8:54:32	8:54:35	0:03	0:00	10	10	20
040293	1023	8:54:35	8:54:35	9:03:02	8:27	0:00	10	10	27
040293	1021	8:55:05	8:55:05	9:02:49	7:44	0:00	5	5	12
040293	1028	9:00:57	9:03:02	9:09:38	6:36	2:05	10	10	27
040293	1029	9:01:56	9:06:27	9:39:48	33:21	4:30	6	6	21
040293	1030	9:08:51	9:10:02	9:12:41	2:39	1:11	2	2	26
040293	1031	9:08:54	9:09:38	9:13:35	3:56	0:44	10	10	27
040293	1032	9:09:06	9:12:33	9:16:29	3:56	3:27	12	12	11
040293	1030	9:12:41	9:15:18	9:21:50	6:32	2:37	5	5	26
040293	1020	9:13:35	9:13:35	9:13:37	0:03	0:00	10	10	12
040293	1020	9:13:38	9:13:38	9:20:39	7:01	0:00	10	10	27
040293	1034	9:17:07	9:21:50	9:27:39	5:48	4:43	5	5	12
040293	1036	9:19:43	9:27:39	9:32:04	4:25	7:55	5	5	12
040293	1037	9:20:07	9:20:39	9:22:37	1:57	0:32	10	10	27
040293	1030	9:21:50	9:22:37	9:22:39	0:03	0:47	10	10	26
040293	1030	9:22:39	9:22:39	9:34:05	11:25	0:00	10	10	27
040293	1036	9:32:04	9:32:11	9:32:12	0:02	0:07	12	12	12
040293	1036	9:32:12	9:32:12	9:46:30	14:17	0:00	12	12	11

Table 2. Input of Program QIE1.

STATION	CUST	ARRIVAL	SERVICE START	SERVICE END
1	3003	8.4475	8.4983	8.5614
1	3030	8.5611	8.5703	8.6319
1	3031	8.5700	8.6050	8.6675
1	3028	8.5464	8.6214	8.6664
1	3034	8.5947	8.6400	8.9836
1	3042	8.6794	8.7053	8.9200
1	3055	8.7092	8.7200	8.8939
1	3065	8.7694	8.7808	8.8686
1	3047	8.6900	8.8686	8.9069
1	3021	8.8381	8.8939	9.0064
1	3082	8.8586	8.9069	9.0300
1	3089	8.8836	8.9200	9.0511
1	3108	8.9661	8.9803	9.0228
1	3110	8.9744	8.9836	9.1797
1	3075	9.0108	9.0511	9.2503
1	3124	9.0814	9.0858	9.1194
1	3145	9.1694	9.1797	9.2978
1	3133	9.1111	9.1931	9.3111
1	3151	9.2183	9.2269	9.3125
1	3146	9.1828	9.2503	9.4394

Table 3. Output of Program QIE1.

TIME	Q1	Q2	CUM ARRIVAL	DEPART FROM QUEUE	DEPART FROM STATION	NUMBER OF SERVERS
8.60	2	1	12	10	6	4
8.62	3	2	13	10	6	4
8.64	3	2	13	10	6	4
8.66	4	2	14	10	6	4
8.68	5	5	16	11	7	4
8.70	6	6	20	14	10	4
8.72	9	8	23	14	11	3
8.74	8	7	23	15	12	3
8.76	10	9	25	15	12	3
8.78	9	8	26	17	13	4
8.80	8	7	27	19	15	4
8.82	9	5	30	21	17	4
8.84	9	7	33	24	20	4
8.86	9	6	34	25	22	3
8.88	11	5	37	26	23	3
8.90	14	4	41	27	24	3
8.92	13	0	44	31	27	4
8.94	13	12	46	33	30	3
8.96	10	9	46	36	32	4
8.98	12	12	49	37	33	4
9.00	10	10	49	39	35	4

TABLE 4. Output of Program QIE2.

TIME	QUEUE ESTIMATE	90%	SIGMA
8.59780	0.00000	0.00000	0.00000
8.62580	2.73606	4.00000	1.20692
8.62580	1.73606	3.00000	1.20692
8.63420	2.51524	4.00000	1.25980
8.63420	1.51524	3.00000	1.25980
8.65690	3.38062	5.00000	1.48414
8.65690	2.38062	4.00000	1.48414
8.68360	4.50375	7.00000	1.53885
8.68360	3.50375	6.00000	1.53885
8.68940	3.96356	6.00000	1.51206
8.68940	2.96356	5.00000	1.51206
8.69440	3.35504	5.00000	1.48633
8.69440	2.35504	4.00000	1.48633
8.72580	4.67561	6.00000	1.32130
8.72580	3.67561	5.00000	1.32130
8.72640	3.71994	5.00000	1.31323
8.72640	2.71994	4.00000	1.31323
8.73140	3.08358	5.00000	1.24997
8.73140	2.08358	4.00000	1.24997
8.74720	3.16511	5.00000	1.06240
8.74720	2.16511	4.00000	1.06240
8.75000	2.35407	4.00000	1.01052
8.75000	1.35407	3.00000	1.01052
8.75940	1.90143	3.00000	0.92213
8.75940	0.90143	2.00000	0.92213
8.80810	2.93798	4.00000	0.79331
8.80810	1.93798	3.00000	0.79331
8.81420	2.18983	3.00000	0.68428
8.81420	1.18983	2.00000	0.68428
8.82390	1.55173	2.00000	0.49732
8.82390	0.55173	1.00000	0.49732
8.84220	1.00000	1.00000	0.00000
8.84220	0.00000	0.00000	0.00000
8.86000	0.00000	0.00000	0.00000
8.86190	1.09087	1.00000	0.29370
8.86190	0.09087	0.00000	0.29370
8.87440	1.27454	2.00000	0.44628
8.87440	0.27454	1.00000	0.44628
8.89220	1.00000	1.00000	0.00000
8.89220	0.00000	0.00000	0.00000
8.90560	0.00000	0.00000	0.00000

2 Busy-Period Interpolation.

2.1 Immediate transitions. Known External Arrivals.

We consider a joint queue at a multi-server station. Several tellers (up to five) work simultaneously at the station.

Using QIE1 we calculate the exact queue for 12 days (complete two weeks) of observations.

Figure 1.1 shows several examples of queues. We immediately note that the queue size changes significantly for different days. There can be three possible causes of such queue variability:

- variability of arrival rate;
- variability of service times;
- variability of number of working servers.

When arrival rates are derived from the data we observe that the first explanation is the main one. Ten days (we do not analyze days 6 and 12 with different working ours and queue patterns) can be divided into three categories according to queue sizes, waiting times or arrival rates.

Day 7 is called "catastrophic day", days 8, 9 and 10 are "heavily loaded days" and the rest are "regular days". Figures 2.2 and 2.3 show average queues and average arrival rates for the categories.

The variability of the arrival rates has a natural explanation. On day 7 (it was Sunday) a payment of social insurance took place. Usually the payment is produced during two days but this time one of them fell on weekend. Therefore, the flow of customers on day 7 was enormous and three days afterwards the bank worked in heavily loaded conditions.

We established that the variability of arrivals is the main source of the queue variability. However, it is not the unique cause. Consider, for example, two heavily loaded days, day 8 and day 10. Figure 2.4 compares the arrival rates on these days and figure 2.5 compares the queues. The arrival rate on day 8 (just after the catastrophic day) was larger, but the queue and the average waiting time (12.2 minutes versus 8.2 minutes) were larger on day 10. The explanation lays in figure 2.6 which compares the number of working servers for two days. We observe that at the "critical periods" (10:00-11:00, 11:15-11:40) four servers only worked on day 10 versus five servers on day 8. The average service time plays some role also (3.35 minutes on day 8 versus 3.58 minutes on day 10).

Figure 2.1
Examples of queues

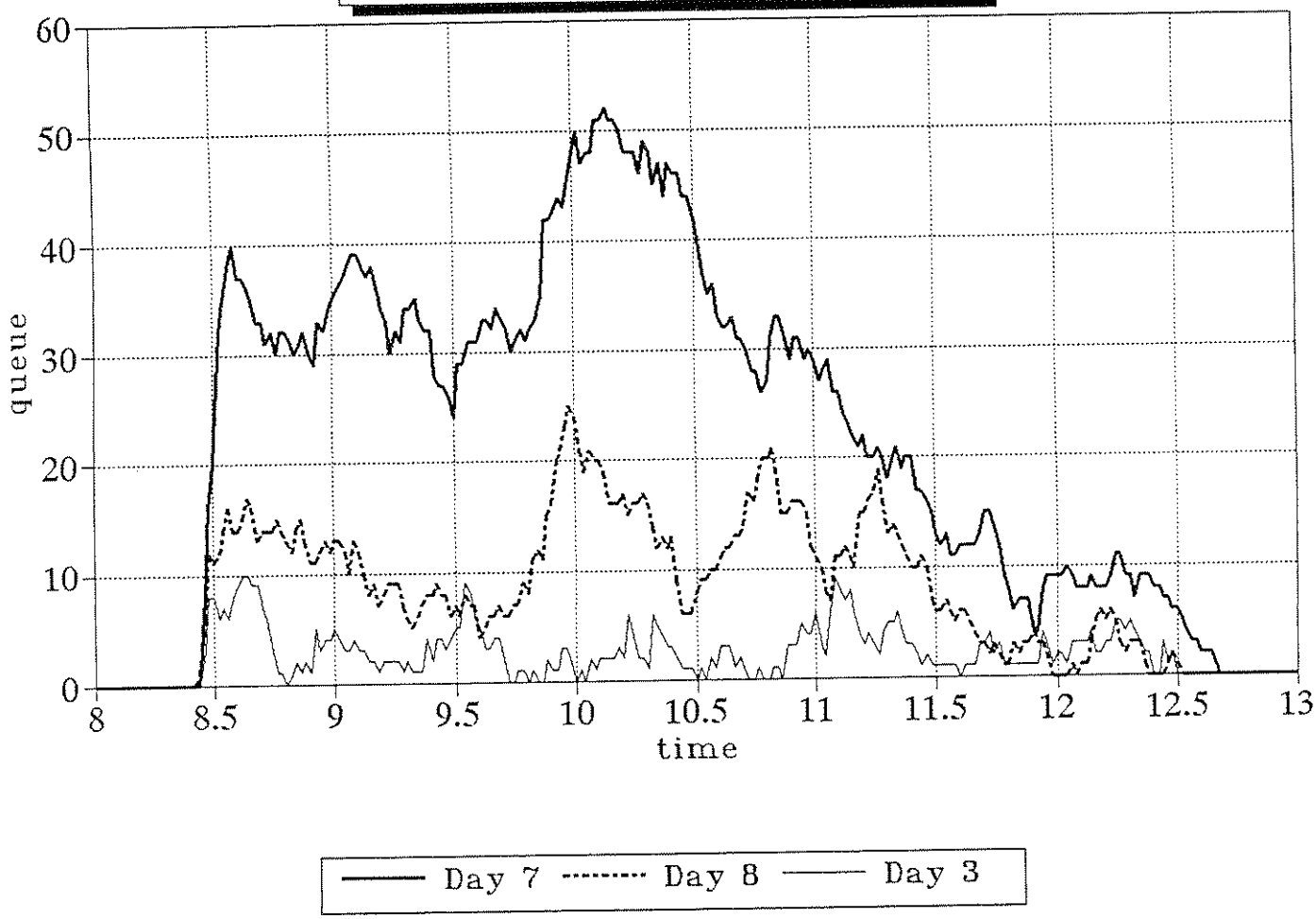


Figure 2.2
Three categories: average queues

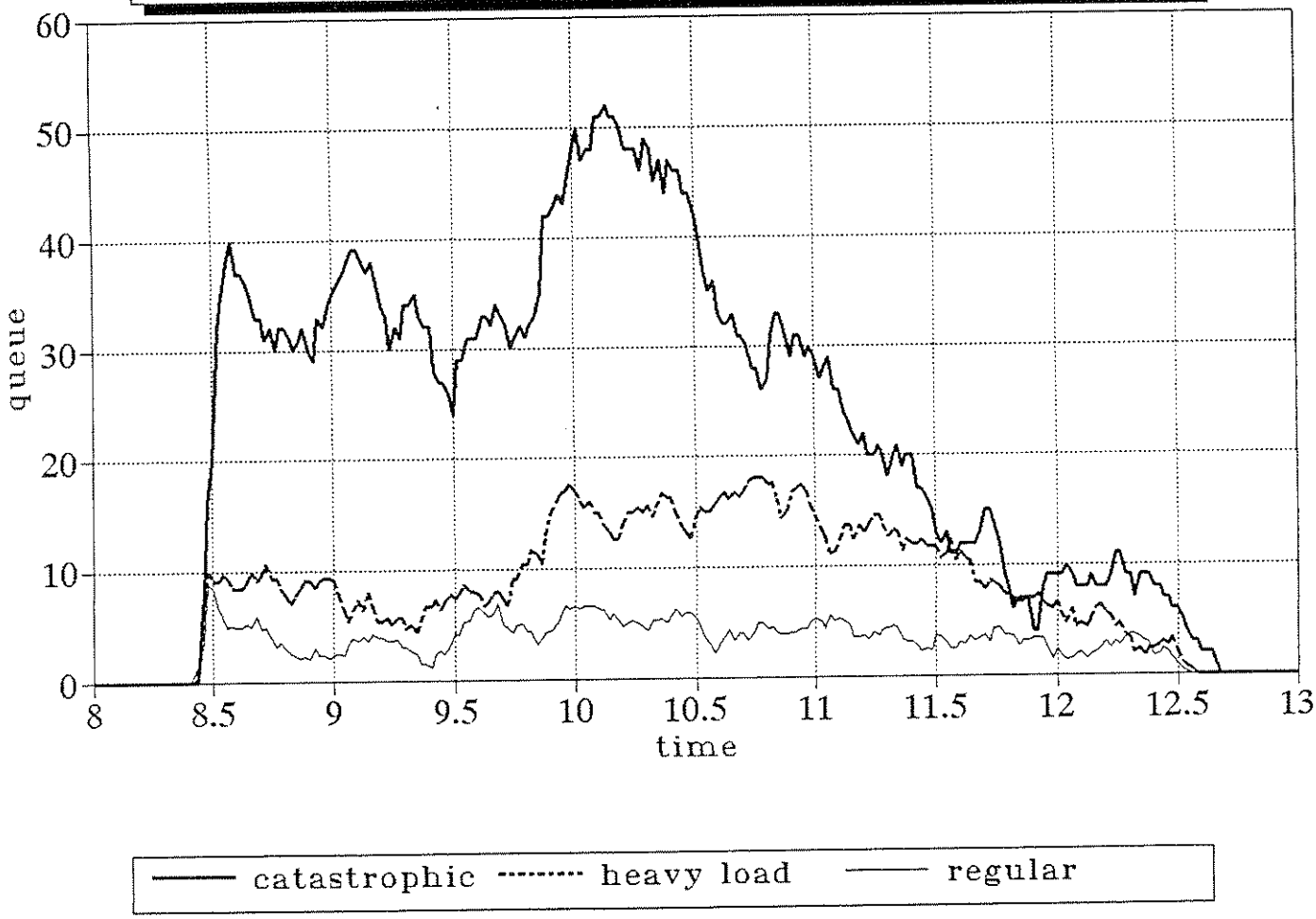


Figure 2.3
Three categories: arrivals

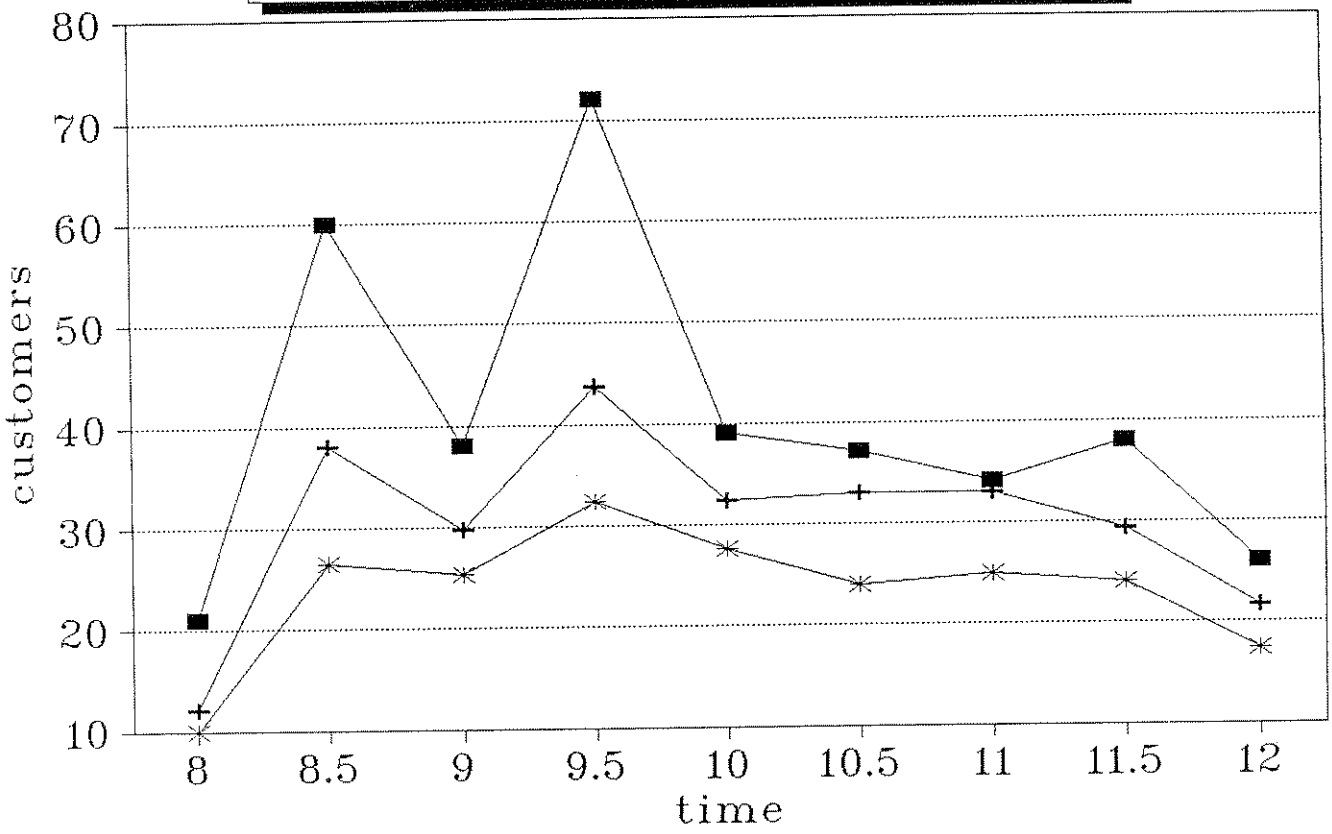


Figure 2.4
Arrivals: day 8 vs. day 10

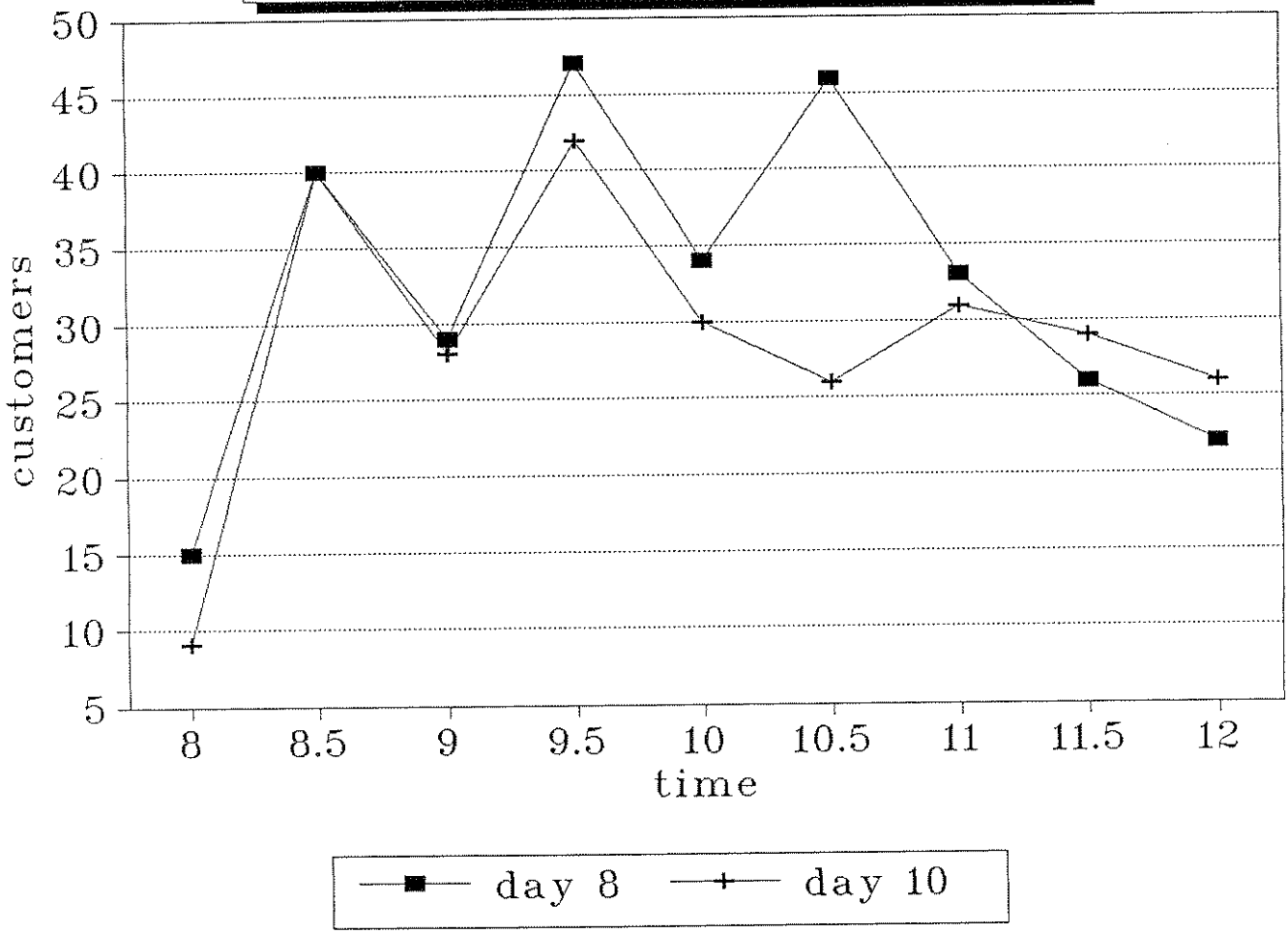


Figure 2.5
Queue: day 8 vs. day 10

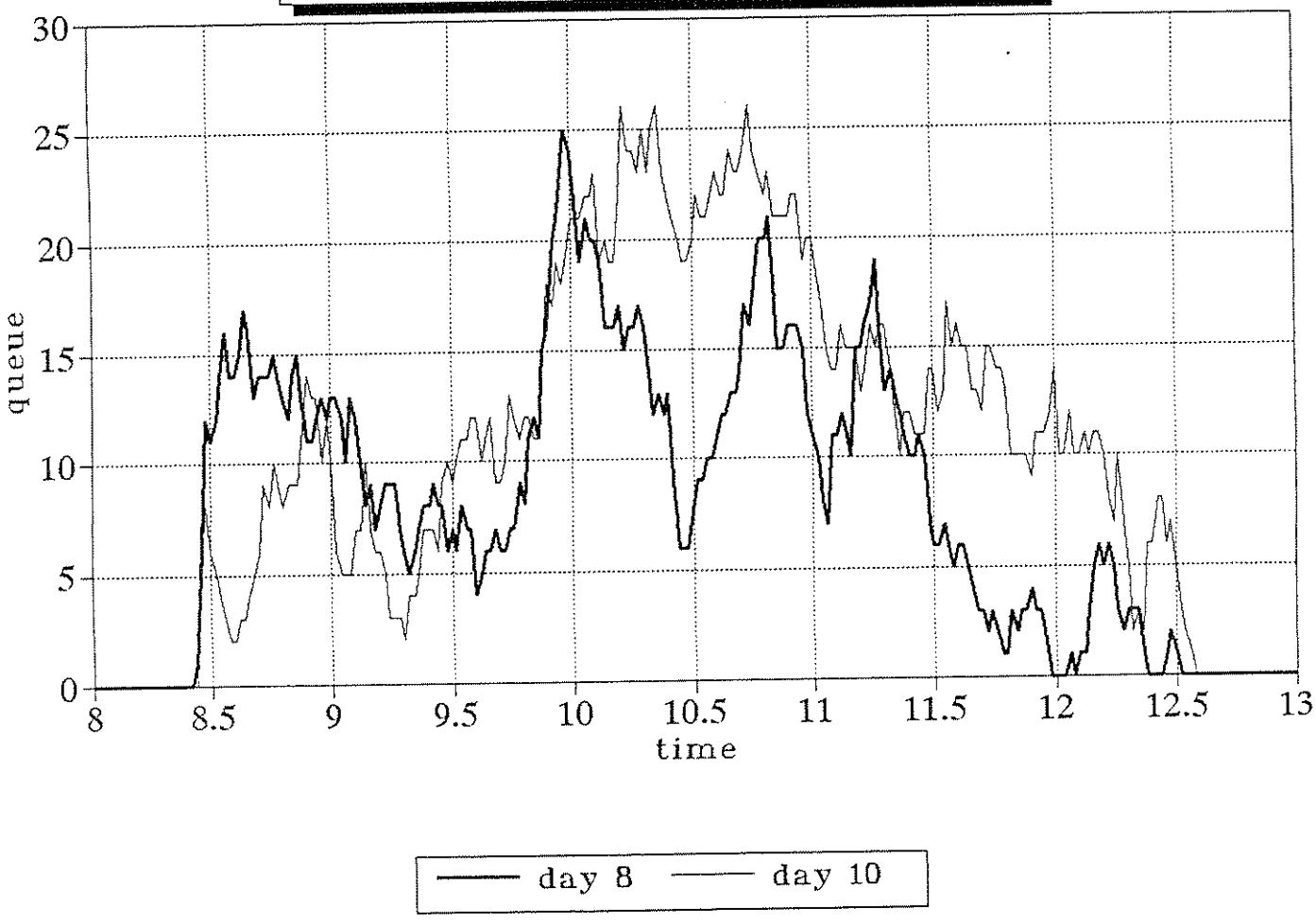
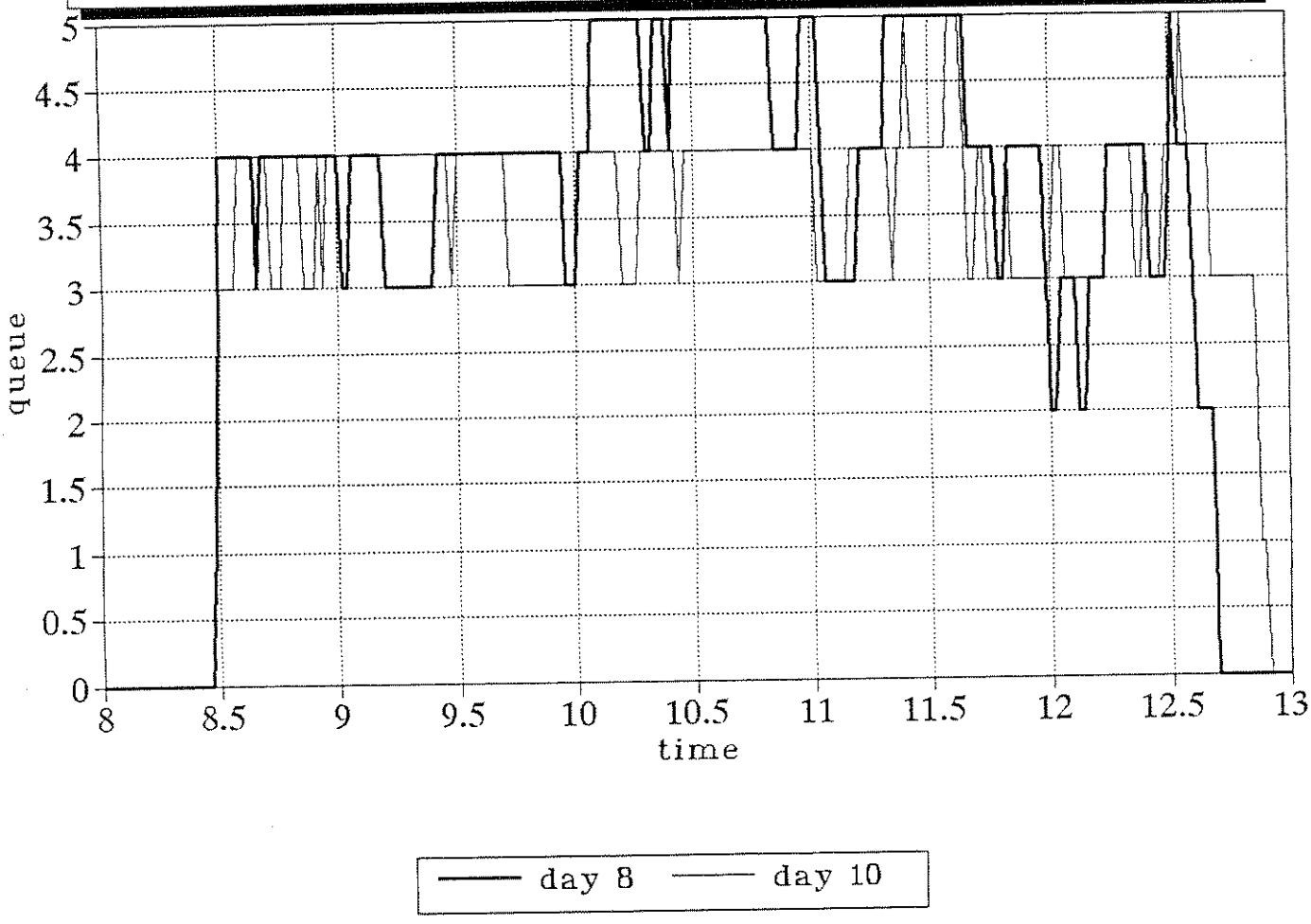


Figure 2.6
Number of servers: day 8 vs. day 10



2.2 Immediate transitions. Unknown External Arrivals.

Single-Server Case. A station with one teller was considered. The Program QIE2 was run for several days at the working periods 8:30-12:30. Some of the results are presented below.

Figures 2.7 and 2.8 illustrate the case of a good correspondence between the estimate computed by program QIE2 and the exact queue computed by program QIE1. When the estimate is computed under the assumption that all arrivals, external and internal, are unknown (using, in essence, the one-station algorithm of Larson), we get Figures 2.9 and 2.10. Figure 2.9 shows that the correspondence is worse in the three time intervals: after 9:00, after 10:00 and between 11:00 and 11:30. Note also the difference between 90% quantiles for two estimates. It is worth mentioning that in the largest busy period (see Figures 2.8 and 2.10) only four arrivals out of 20 are internal. However, their knowledge provides us with exact queue at these times (assuming FCFS stands) and helps to improve the estimate significantly.

Now consider the case of comparatively bad correspondence after 11:00. Table 5 presents the corresponding place of the data file. According to the record, customer 3468 received service more then 15 minutes (and the average service times is less then 3.5 minutes). Naturally, our algorithm "supposed" that several customers should arrive during such a long service. Probably, the server forgot to register the service termination of customer 3468 and, erroneously, it was registered only at the service start of customer 3524.

As was mentioned already, most problems with the real-data applications take place because the work-conserving principle does not always stand. See, for example, Figures 2.11 and 2.12. Most of the time, we have a reasonable correspondence between the exact queue and the estimate but between 11.5 and 12.5 the estimate turns to be very far from reality. Let us try to understand the cause looking at the data in Table 6. We observe that at 12.0700 and 12.2931 there are customers standing in queue but, anyway, the server interrupts service for a short time. Our algorithm interprets this fact as the end of busy period. If we prolong service times of customers 3638 and 3659 we get rather different result (see Figure 2.13).

In order to avoid the problems of this kind, a server should register intervals when he interrupts his work and there are customers in queue. Then we can suppose in the algorithm that the service before the interruption is prolonged until the start of the next service.

Consider the queue estimate on day 11 (Figures 2.14 and 2.15). The correspondence is normal everywhere except the time interval between 10.5 and 10.8. Table 7 contains the corresponding registrations. The estimate inaccuracy can be due to the violation of the assumption of immediate transitions (see customer 3356, for example).

Multi-Server Case. We consider a service station with 5 servers. Figure 2.16 shows the the exact queue versus the estimate on day 8 and Figure 2.17 shows 90% quantile on this day. Note that the estimate follows, in large, the pattern of the exact queue. However, there seems to be underestimation for the large queue peaks and at the beginning of the day. Apparently, this is due to the inhomogeneous arrival rate during the day.

TABLE 5. Day 8. Part of Data File.

Station	ID	Arrival	Service start	Service termination
4	3458	10.9594	10.9939	11.0322
4	3468	11.0456	11.0628	11.3219
4	3524	11.3011	11.3219	11.3358
4	3517	11.2622	11.3358	11.4075

Table 6. Day 10. Part of Data File.

Station	ID	Arrival	Service start	Service termination
4	3609	11.5517	11.9347	11.9783
4	3616	11.5781	11.9783	11.9953
4	3638	11.7042	11.9953	12.0700
4	3646	11.7189	12.1014	12.1353
4	3670	11.8286	12.1353	12.2317
4	3659	11.7814	12.2317	12.2931
4	3688	11.9175	12.3125	12.4094
4	3740	12.1894	12.4094	12.5467

Table 7. Day 11. Part of Data File.

Station	ID	Arrival	Service start	Service termination
4	3281	10.2150	10.4344	10.4778
4	3354	10.4722	10.4778	10.7283
4	3380	10.6556	10.7744	10.8311
4	3376	10.6258	10.8311	11.0192
4	3400	10.8011	11.0192	11.0433
4	3356	10.4806	11.0433	11.1022

Figure 2.7
Day 8. Single-server station

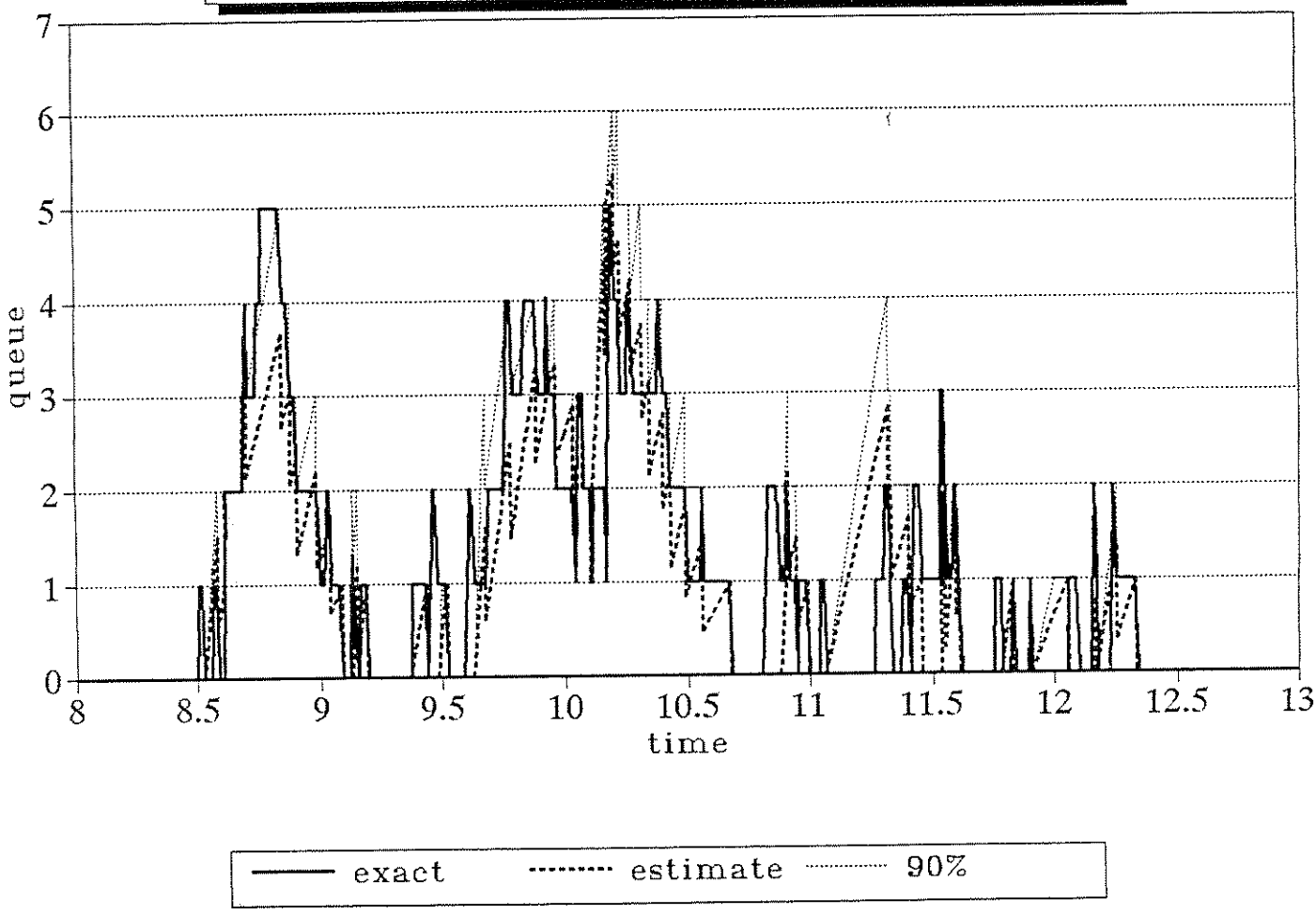


Figure 2.8
Day 8. Longest busy period

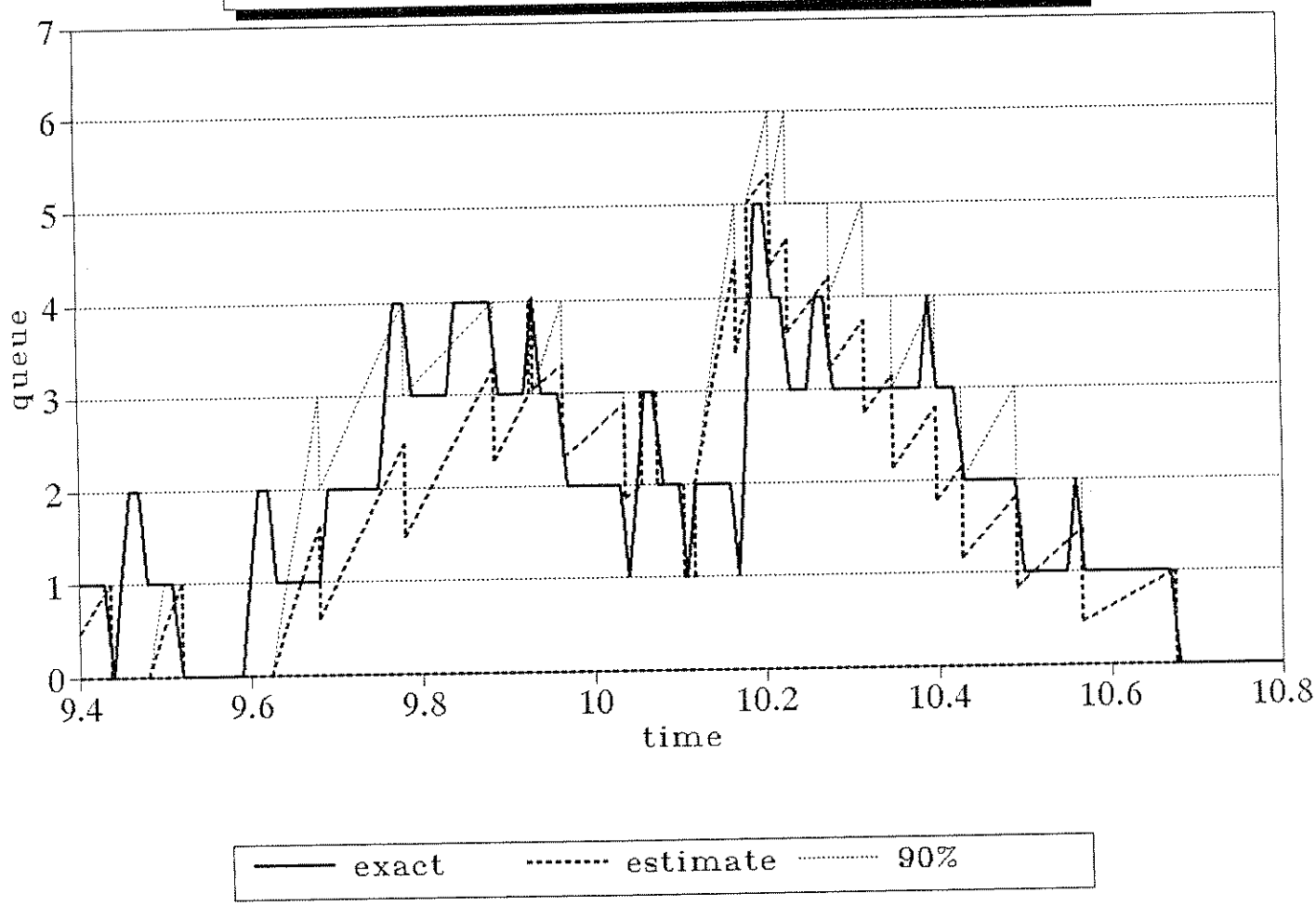


Figure 2.9
Day 8. All arrivals unknown

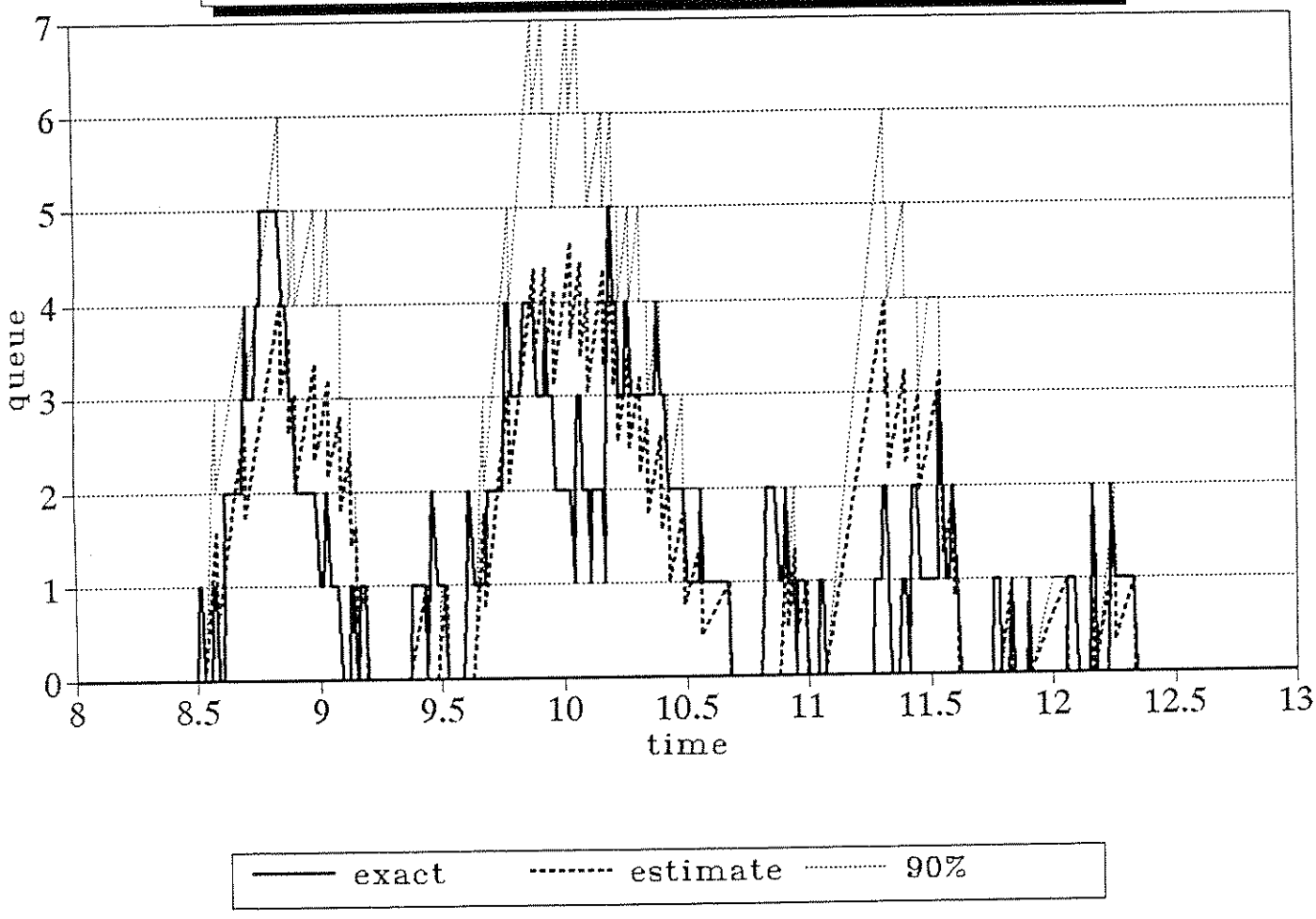


Figure 2.10
Day 8. All arrivals unknown. Longest BP

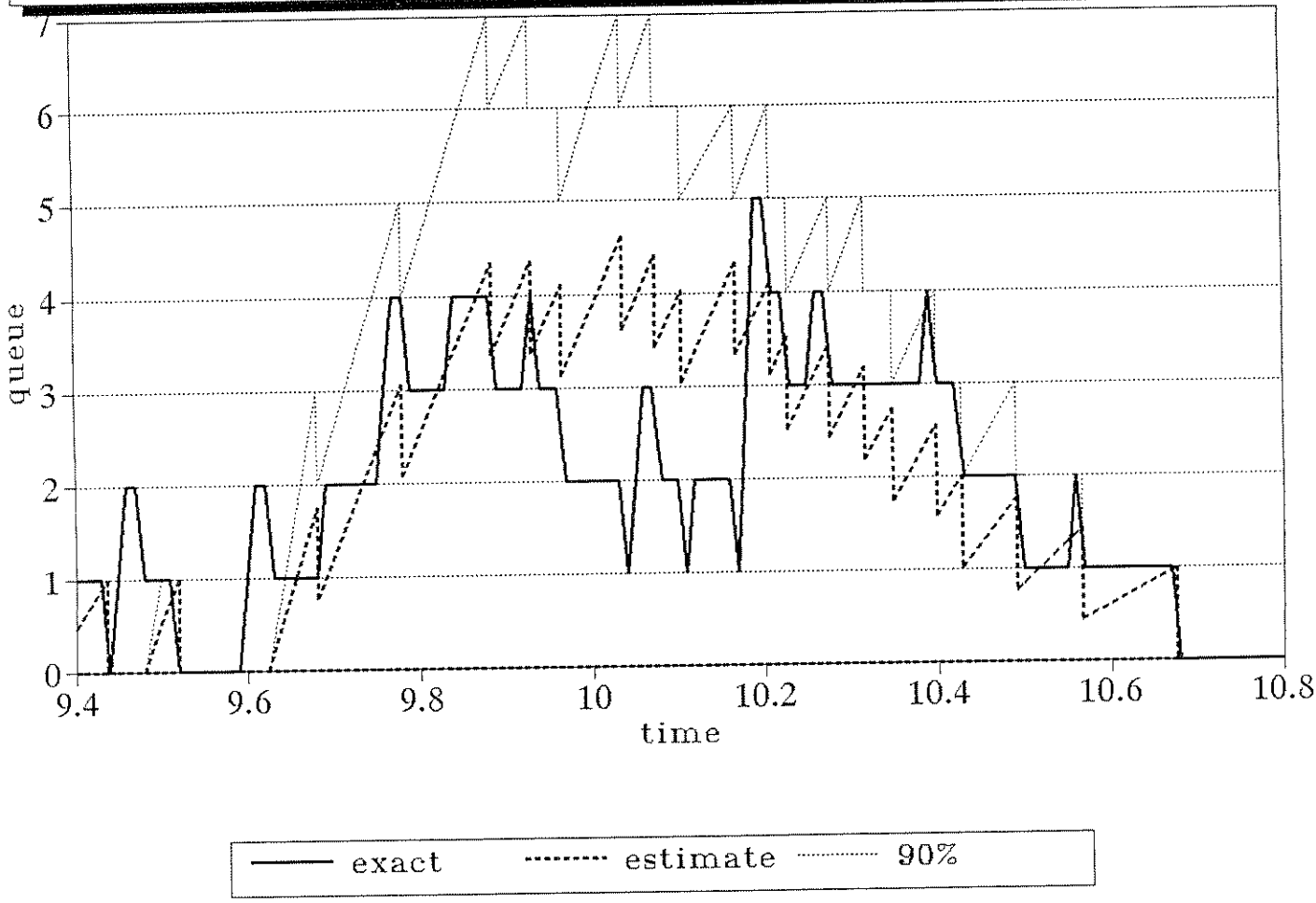


Figure 2.11
Day 10. Single-server station

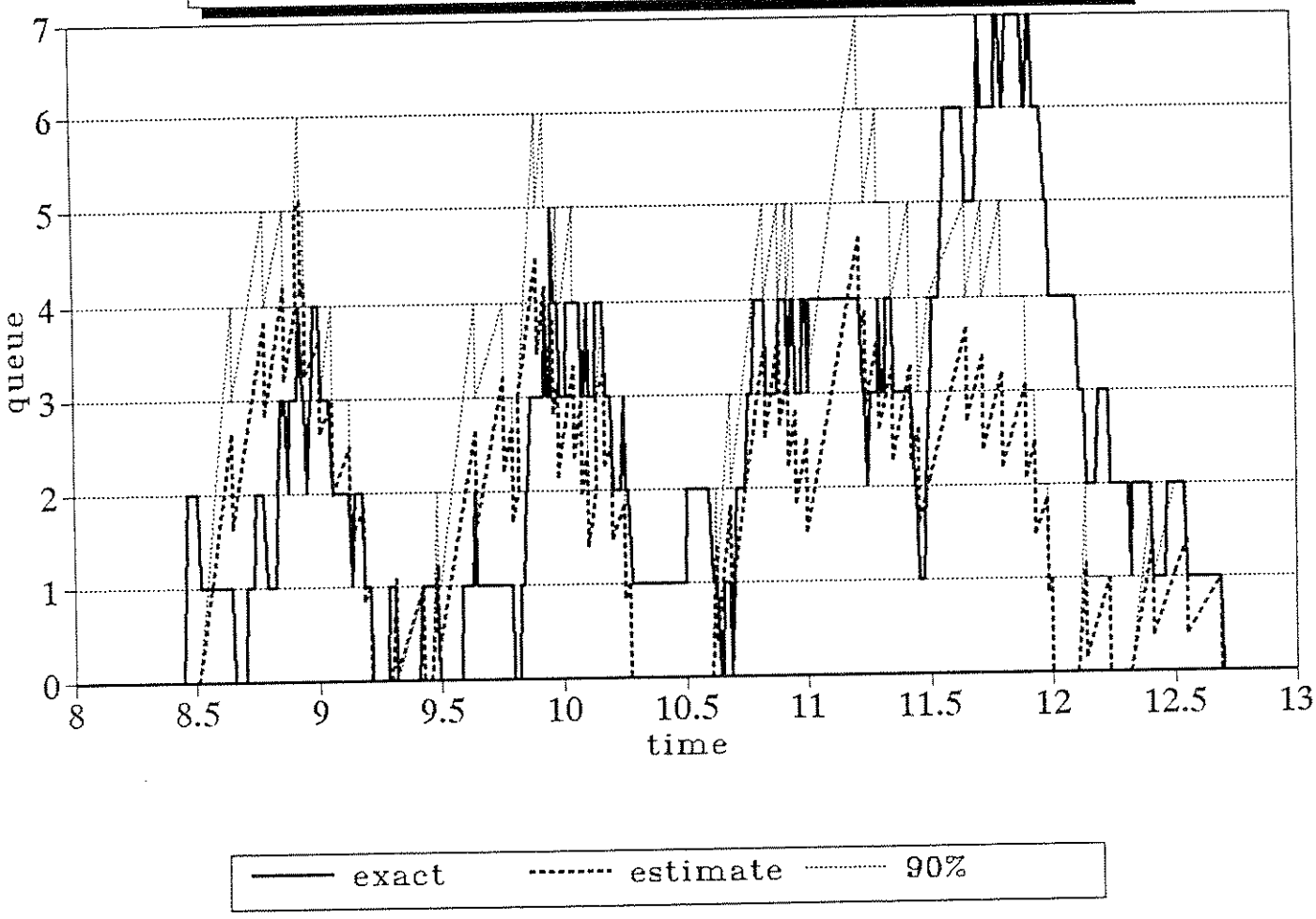


Figure 2.12
Day 10. Longest busy period

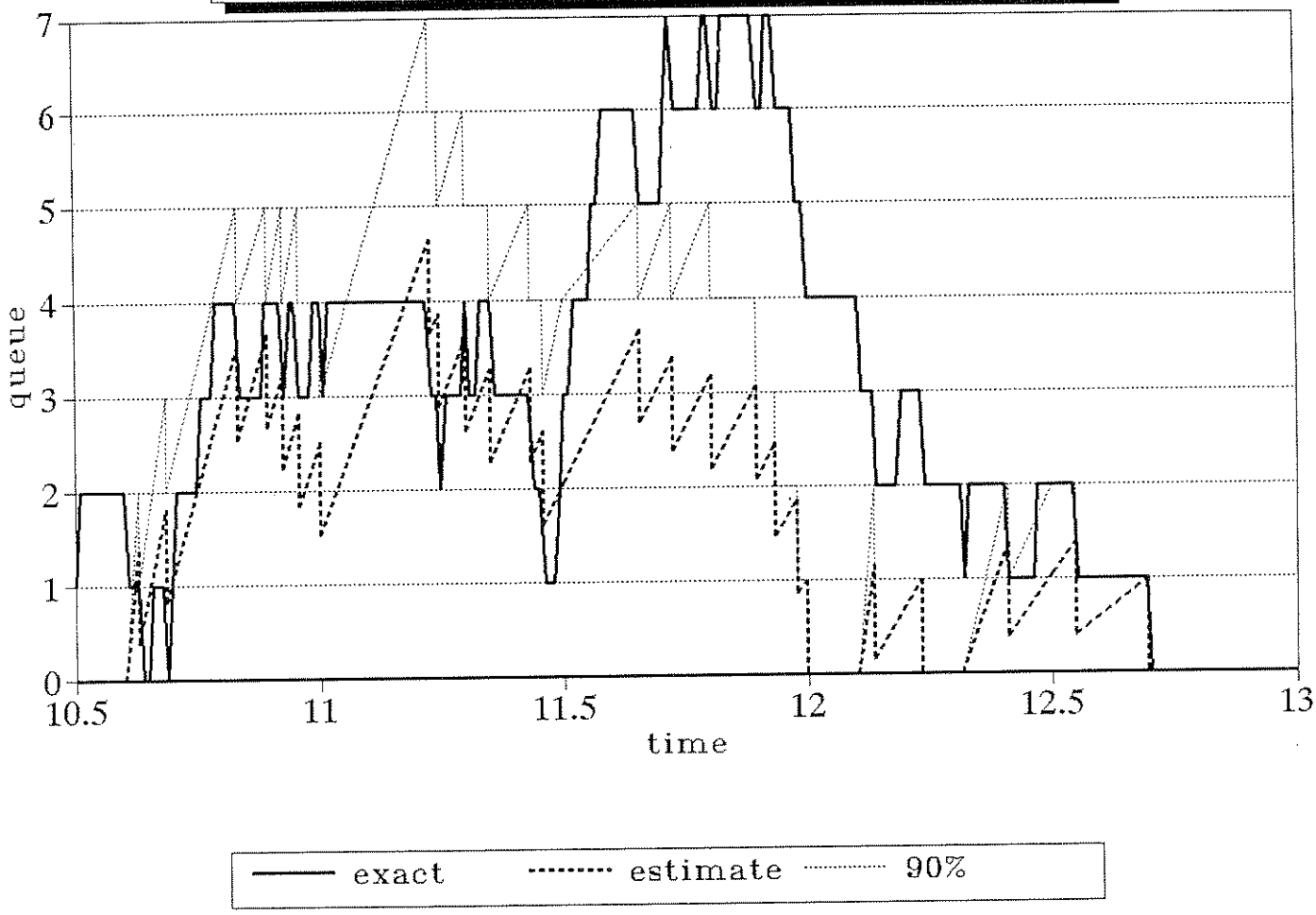


Figure 2.13
Day 10. Correction of estimate

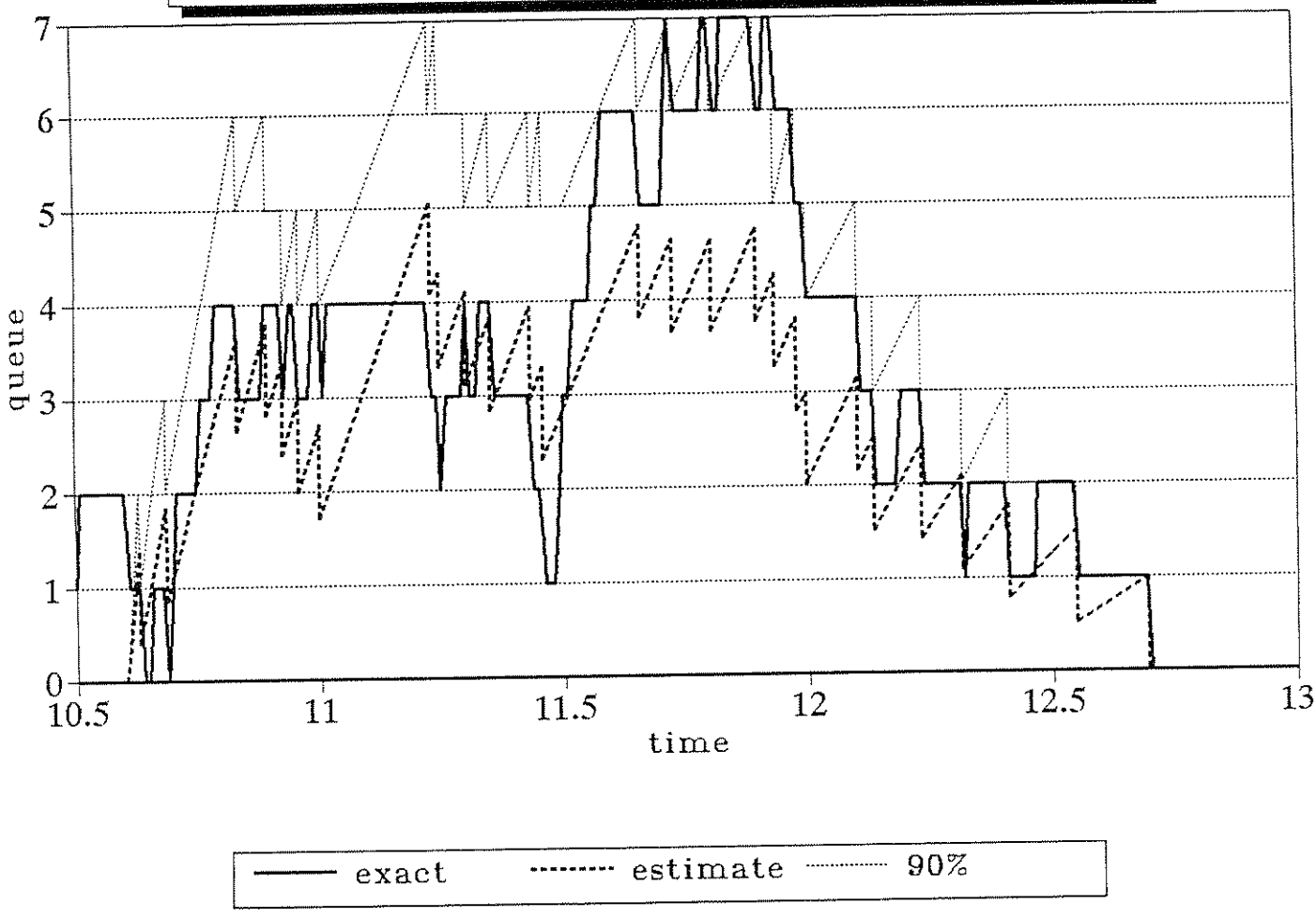


Figure 2.14
Day 11. Single-server station

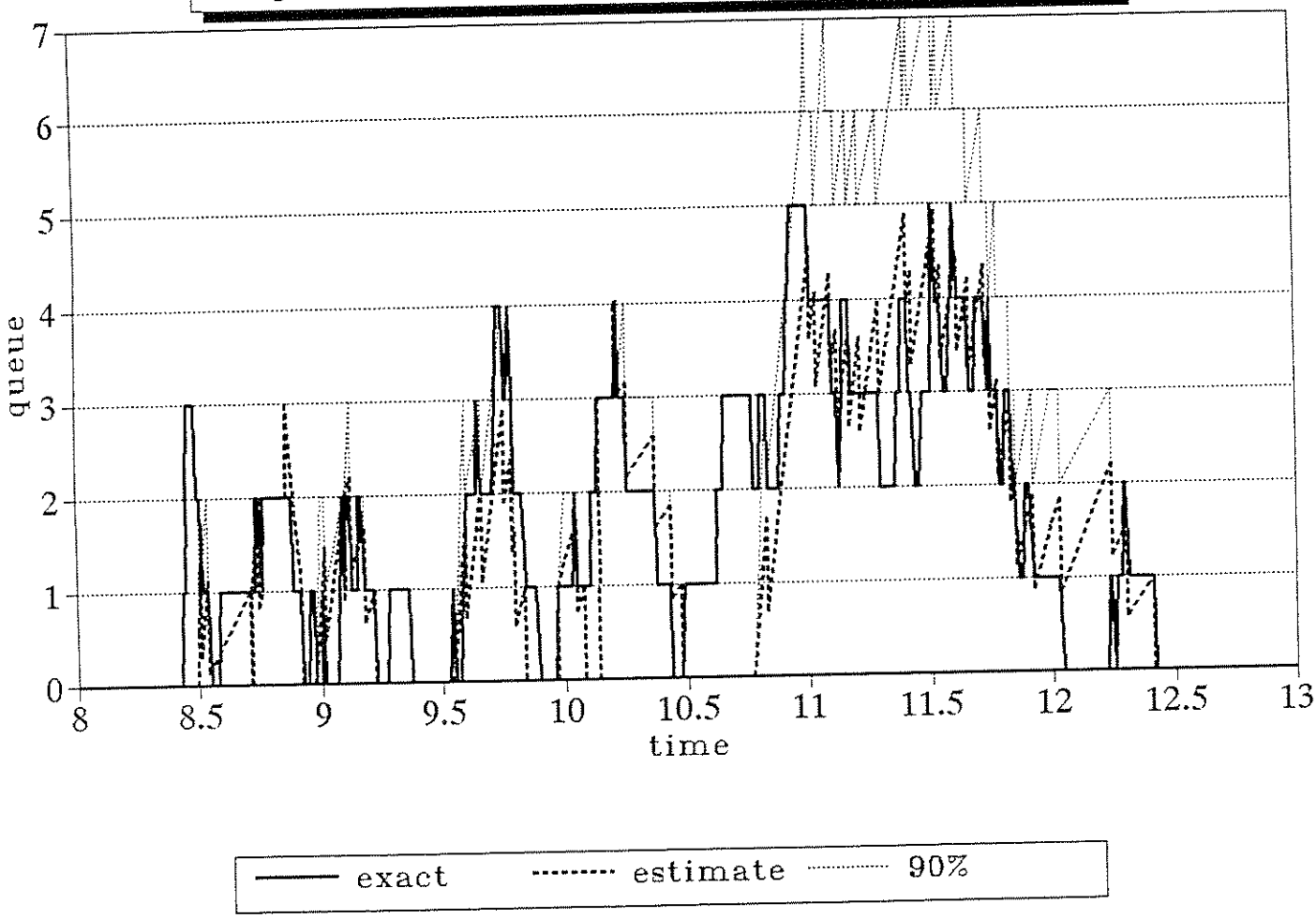


Figure 2.15
Day 11. Longest busy period

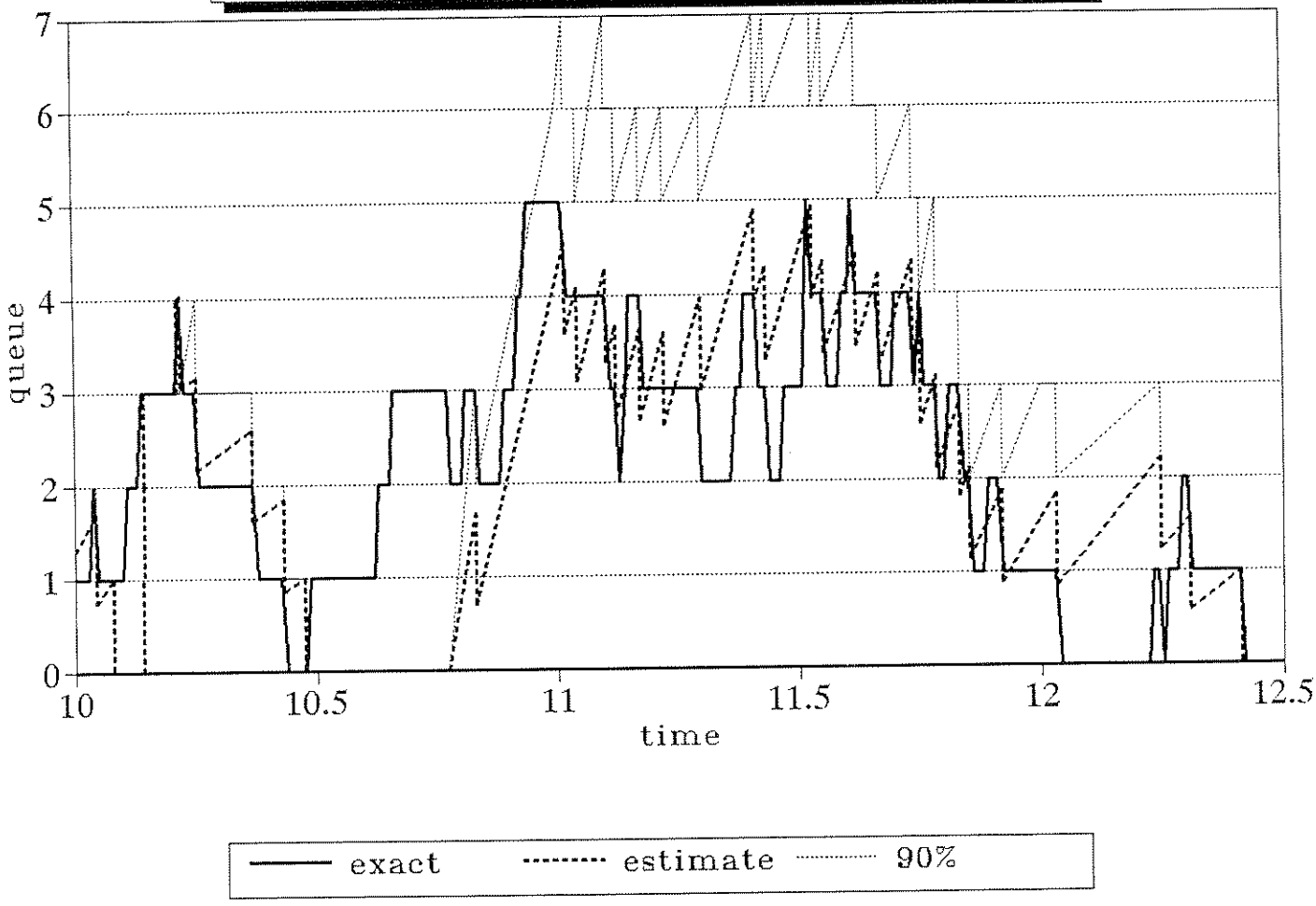


Figure 2.16
Day 8. Multi-server station

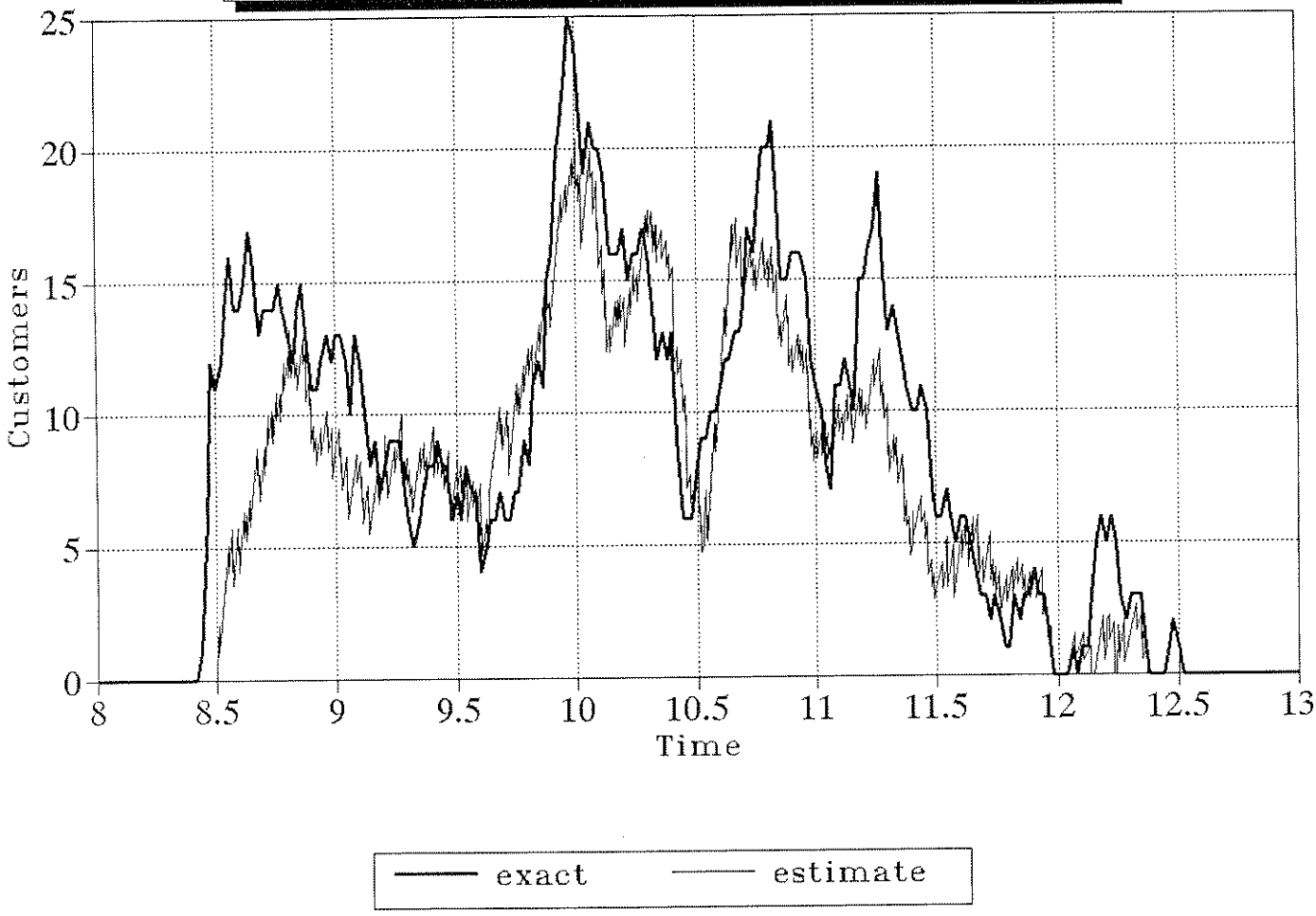
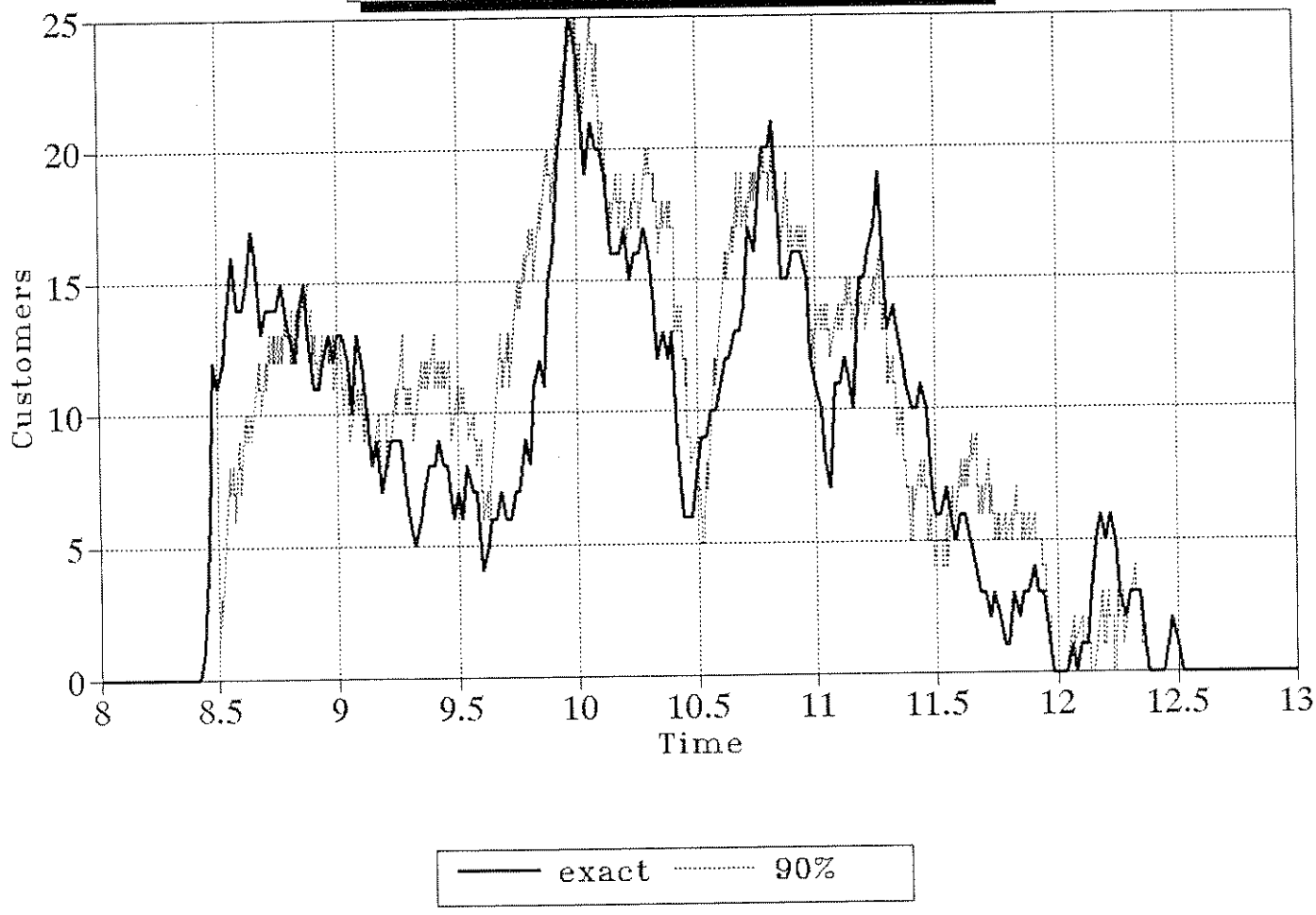


Figure 2.17
Day 8. Quantile 90%



Chapter IV

Possible Future Research.

We conclude with a list of possible research directions, that would build upon the work presented here.

Theoretical Research. It seems that, within the models presented here, we have exhausted most possibilities for exact analysis. Interesting extensions are possible in the following directions:

- Approximations. An appropriate mathematical framework is the theory of conditional weak convergence ([10],[15]) and excursion theory (for example, [1]). Queue length during a busy period, after proper normalization, can converge to a Brownian excursion when the number of services is large. Daley and Servi began working in this direction in [5].

- Cover other types of available information. This can include either more or less than transactional data. For example, more data could include receiving a signal when the queue length exceeds some threshold, as in Hall [8]). Less data could exclude ID's. Even in simple cases, exact solution of the model without ID's requires heavy combinatorial calculations and is, in general, non-computable. Therefore, approximations are desirable here.

- Add new modeling features, such as general (non- Poisson) arrival pattern. The exact solution for the single-station case was derived in Bertsimas and Servi [2] but it is computationally hard to apply to real data. Therefore, the need for approximations arises again. Other useful modeling features include buffers, abandonment, reneging, multi-type customers and violations of FCFS (for example, queues with priorities).

- Analyze new performance measures, such as waiting times distribution during a busy period.

Applications.

- Software programs for different special cases of the problem could be developed. The following cases seem to be both important and amenable for computations:
  - Busy-Period Interpolation. Exponential Transitions, Known External Arrivals.
  - Real-Time Estimation. Immediate Transitions, Known and Unknown External Arrivals.
  - Known non-homogeneous arrival rates for different special cases.
- Development of robust algorithms that are less sensitive to measurement errors and violations of model assumptions. In particular, being able to cope with violations of the work-conserving principle is very important. In principle, we may try to filter out long service times that seem to include periods of server's idleness.
- In the cases of Real-Time Estimation, exact calculations of the queue estimates seem to be extremely time-consuming (except for the case of immediate transitions and known external arrivals). Therefore, control in real-time would require suitable heuristics.
- It is important to search for real data, on which our observations could be tested. Possible sources are communication and transportation networks.

Bibliography

[1] R.M.Blumenthal "Excursions of Markov Processes", Birkhauser, Boston, (1992).

[2] D.J.Bertsimas, L.D.Servi "Deducing queueing from transactional data: the Queue Inference Engine, revisited", *Operations Research*, 40 (1992) 217-228.

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

[4] D.J.Daley, L.D.Servi "A two-point Markov chain boundary value problem", *Advances of Applied Probability*, 25 (1993) 607-630.

[5] D.J.Daley, L.D.Servi "Approximating last exit probabilities of a random walk, with application to conditional queue length moments within busy periods of $M/GI/1$ queues", *Journal of Applied Probability*, 31A (1994) 251-267.

[6] S.G.Eick, W.A.Massey, W.Whitt "The physics of the $M_t/G/\infty$ queue", *Operations Research*, 41 (1993) 731-742.

[7] R.Gawlick "Estimating disperse queueing networks: the Queue Inference Engine", *Computer Communication Review*, 20 (1990) 111-118.

[8] S.A.Hall "New directions in queue inference for management implementations", Massachusetts Institute of Technology, Ph.D. Thesis, (1992).

[9] R.W.Hall "Queueing Networks", Prentice Hall, (1991).

[10] Iglehart D.L. "Conditioned limit theorems for random walks", in M.Puri (Ed.) "Stochastic Processes and Related Topics", vol.1, Academic Press, New York.

[11] L.K.Jones, R.C.Larson "Efficient computation of probabilities of events described by order statistics and applications to queue inference", preprint (1992).

[12] R.C.Larson "The Queue Inference Engine: deducing queue statistics from transactional data", *Management Science*, 36 (1990) 586-601.

[13] R.C.Larson "The Queue Inference Engine: deducing queue statistics from transactional data, addendum", *Management Science*, 36 (1990) 1062.

[14] L.R.Rabiner "A tutorial on Hidden Markov Models and selected applications in speech recognition", *Proceedings of the IEEE*, 77 (1989) 257-286.

[15] Takacs, L "Queueing methods in the theory of random graphs", in J.H.Dshalow (Ed.) "Advances in Queueing: Models, Methods and Problems", CRC Press 1995, 45-78.

[16] J.Walrand "Introduction to Queueing Networks", Prentice Hall, (1988).

אמידת מאפיינים של רשתות תורים המבוססת על נתוני טרנסאקציות

חבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת תואר
מגיסטר למדעים בסטטיסטיקה

סרגיי זלטין

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

טבת, תשנ"ו חיפה ינואר 1996

המחקר נעשה בהנחיית פרופסור אבישי מנדלבאום
בפקולטה להנדסת תעשיה וניהול

אני מודה לטכניון – מכון טכנולוגי לישראל
ולמרים ואהרון גוטוירט
על התמיכה הכספית הנדיבה בהשתלמותי

תודה מיוחדת לפרופסור אבישי מנדלבאום על האנרגיה, ההשראה
והזמן שהשקיע בעבודה זו, וכן על התמיכה והעידוד המתמיד

תודתי העמוקה לאשתי אולגה על האהבה והאמונה בי

תוכן עניינים

רשימת ציורים

רשימת טבלאות

תקציר מורחב

1 הקדמה

נניח שרוצים למדוד תור המצטבר לפני מכשיר "כספומט". עקרונית, ניתן למדוד את התור במדויק: להתקין מצלמה או להעמיד עובד כדי לרשום מופעים ועזיבות. אבל אלו דרכים יקרות לקבלת מידע.

גישה אלטרנטיבית מבוססת על מדידה של התחלות וסיומי שרות (זמני טרנסאקציות, transaction times) בעזרתו של מכשיר הכספומט עצמו. במקרה זה לא מודדים הגעות של צרכנים, לכן לא ניתן למדוד את התור בצורה מדויקת ונאלצים לאמוד אותו על סמך מידע חלקי.

קיימות מערכות תורים רבות הדומות למערכת בדוגמה הנ"ל. הן בעלות שתי התכונות הבאות:

- תורים נסתרים (invisible queues), ז"א תורים שבלתי אפשרי או יקר למדוד אותם.
- ניתן להשיג את הזמני הטרנסאקציות בצורה פשוטה יחסית.

דוגמאות כוללות, למשל:

- שירות טלפוני. מספר הקווים במרכזיות טלפון מוגבל. זמני הטרנסאקציות מתאימים להתחלות וסיומי שיחות. אם צרכנים שקיבלו "תפוס" מנסים להתקשר שוב בהסתברות p, הם יוצרים retrial queue, שהוא תור נסתר. (בדרך כלל, לא ניתן להשיג זמני הגעות של צרכנים חסומים בתוך מרכזיה וחייבים לפנות ל"בזק" למטרה הזו.)
- מערכת פלאפונים (כאן באופן עקרוני לא ניתן למדוד זמני "תפוס").
- תורי מכוניות בצמתים. טרנסאקציות כאן מתאימות לזמנים שמכוניות חוצות כבל מדידה, המותקן בכביש.
- שרות פנים-על-פנים בבנקים, משרדי ממשלה, קופות חולים וכו'. גם כאן יכול להיות פשוט יותר למדוד זמני טרנסאקציות ולא זמני הגעות.

שני המקרים הראשונים הם דוגמאות של תורים בתחנת שירות בודדת. באשר למודלים של תחבורה ושרות פנים-על-פנים, ניתן להציגם כרשתות תורים.

2 מחקר קודם

המחקר הקודם מוקדש כולו למקרה של תחנה בודדת. Larson פתח את הנושא ב-1990 וטבע את המונח Queueing Inference Engine (QIE) כדי לתאר את אלגוריתם האומד תור על סמך זמני טרנסאקציות.

Larson הניח ש s שרתים עובדים במקביל בתחנת שירות, זמני הגעות ניתנים לתאור כתהליך פואסון וזמני שירות הם בעלי התפלגות כללית (תור $M/G/s$). קצב ההגעה λ יכול להיות לא ידוע. הוא השתמש בעובדה, שבקירוב, ניתן לזהות תקופות תעסוקה (תקופות בהן כל השרתים עסוקים) בעזרת זמני טרנסאקציות. ואכן, אם תחילת שרות חדש מתרחש מייד אחרי סיום השרות הקודם, אזי תקופת התעסוקה בפועל.

Larson עסק בבעית אינטרפולציה בתום תקופת תעסוקה: הוא הניח שנמדדו כל הטרנסאקציות של התקופה ופיתח אלגוריתם האומד את המסלול של התור במשך התקופה. האלגוריתם מבוסס על תכונת סטטיסטי הסדר של תהליך פואסון. מספר החישובים הנדרשים שווה ל $O(n^3)$, כש n הוא מספר הטרנסאקציות במשך תקופת תעסוקה.

Daley ו-Servi, החל מ-1992, פיתחו גישה אלטרנטיבית בכמה מאמרים. בעזרת הטכניקה של הסתברויות טאבו (taboo probabilities) בשרשרות מרקוב הם פתרו את הבעיה של Larson, ובנוסף גם את הבעיות הבאות:

- $E_k/G/1$ (זמנים בין-מופעים מתפלגים ארלנג).
- $M/G/s/m$ (מספר צרכנים במערכת חסום על ידי M).
- Reneging (צרכן יכול לעזוב כאשר הוא מחכה זמן רב מידי לטעמו).
- Bernoulli feedback (צרכן המסיים שרות חוזר למערכת בהסתברות p).
- Balking (צרכן מגיע נוטש בהסתברות p אם הוא נתקל בתור בגודל M או יותר).

Bertsimas ו-Servi, גם הם ב-1992, תרמו תרומה חשובה למחקר הבעיה. הם עסקו במודלים הבאים:

- אמידת תור ב-real-time, תוך כדי תקופת תעסוקה.
- $M_t/G/s$ (הגעות מתוארות על ידי תהליך פואסון לא הומוגני בזמן).
- $GI/G/s$ (זמנים בין-מופעים בעלי התפלגות כללית). במקרה זה, הפתרון כולל אינטגרלים רב-מימדיים וקונוולוציות של פונקצית התפלגות. לכן קשה לבצע חישובים מפורשים.

3 ניסוח הבעיה

נתונה רשת תורים פתוחה עם d תחנות שרות המסומנות $\mathcal{D} = \{1, 2, \ldots, d\}$. אנחנו מניחים על הרשת את ההנחות הבאות:

- הגעות חיצוניות לרשת מהוות תהליך פואסון לא הומוגני בזמן, בעל קצב $\alpha = \{\alpha(t), t \geq 0\}$.
- מעברי צרכנים בין תחנות נקבעים על ידי מטריצת מעברים $P = [p_{jk},\ j, k \in \mathcal{D}]$.

- זמני שירות, $\xi_{ji},\ j \in \mathcal{D},\ 1 \le i < \infty$ הם משתנים מקריים בעלי התפלגות משותפת כללית (בפרט, הם יכולים להיות תלויים).
- המרכיבים הסטוכסטיים של הרשת, כלומר זמני הגעות חיצוניות, זמני שירות ומעברים בין תחנות, הם בלתי תלויים זה בזה.
- יכולים לעבוד כמה שרתים במקביל, בכל תחנת שירות.
- מדיניות שירות First Come First Served (FCFS) תקיפה בכל התחנות.
- עיקרון work-conserving מתקיים: שרת לא יכול להיות בטל אם לקוח ממתין לקבלת שרות.
- כל הצרכנים מקבלים מספרי זהות (Identification Numbers, ID's) כאשר הם נכנסים לרשת. ה-ID נשאר עם הצרכן עד יציאתו מהמערכת.

זמני המעבר בין תחנות יכולים להיות מיידיים או להתפלג אקספוננציאלי עם קצב $\eta_{jk},\ j,k \in \mathcal{D}$. אנו מניחים שצרכנים חיצוניים מגיעים לתחנה 0 ומכאן עוברים לתחנות אחרות לפי הסתברויות המעבר $p_{0j},\ j \in \mathcal{D}$.

המדידות הנתונות לנו כוללות זמני טרנסאקציות ובמקרה פרטי גם זמני הגעות חיצוניות.

במחקר נעסוק בשלושה סוגי בעיות:

אינטרפולציה בתום תקופת תעסוקה. נסמן ב-$[T_1, T_2]$ תקופת תעסוקה בתחנה j. נאמוד מסלול של תור $Q_j(s),\ s \in [T_1, T_2]$ על סמך נתוני מדידות עד זמן T_2.

אמידה ב-*real-time*. נאמוד ב-real-time $Q(t)$ על סמך מדידות עד זמן t.

אינטרפולציה כללית. נאמוד מסלול של תור $Q(s),\ 0 \le s \le t$, על סמך מדידות עד זמן t.

4 אינטרפולציה בתום תקופת תעסוקה

נסמן

$$t_0,\ t_1, \ldots, t_n, t_{n+1},$$

זמני טרנסאקציות בתקופת תעסוקה בודדת בתחנה j, כאשר

t_0 תחילת תקופת תעסוקה,

$t_1, \ldots, t_n$ התחלות שירות הזהות גם לסיומי שירות,

t_{n+1} סוף תקופת תעסוקה.

נסמן $A_0 = t_0, A_1, \ldots, A_n$ זמני הגעות לתחנה j של הצרכנים המתחילים שירות ב- $t_0,\ t_1, \ldots, t_n$. מספר

ההגעות המצטבר בקטע $(t_0, t]$ יסומן

$$A(t), \;\; t_0 \leq t \leq t_{n+1} \;\; (A(t_0) = 0, \; A(t_n) = A(t_{n+1}) = n).$$

4.1 זמני מעבר מיידיים, הגעות חיצוניות ידועות

נציג תור בתחנה j בזמן t כ-

$$Q_j(t) = Q_j^{\alpha}(t) + Q_j^{\beta}(t),$$

כאשר $Q_j^{\alpha}(t)$ מורכב מצרכנים חיצוניים (שהגיעו לתחנה ישירות מחוץ) ו-$Q_j^{\beta}(t)$ מורכב מצרכנים פנימיים.

גודל $Q_j^{\alpha}(t)$ ידוע משום שאנו רושמים הגעות חיצוניות. זמני הגעות של צרכנים פנימיים שווים לזמני סיומי שירות האחרונים שלהם (משום שהמעברים מיידיים). בעזרת ID-ים ניתן לחשב את הזמנים הללו ולכן, לחשב את התור במדויק.

4.2 זמני מעבר מיידיים, הגעות חיצוניות לא ידועות

אנו נעבור על מקרה זה בפרוט, ונסתפק בסקירה בלבד של המקרים הפרטיים האחרים.

ננית שהגעות חיצוניות מהוות תהליך פואסון הומוגני בזמן עם קצב λ. (קיימת הכללה פשוטה למקרה הלא הומוגני.) נסמן ב-s_i זמן סיום שירות אחרון של צרכן פנימי שמתחיל שירות חדש ב-t_i. ניתן לייצג את המידע הרלוונטי לאמידת התור כתיתוך של ארבע מאורעות:

E_1. $A_0 = t_0$.

E_2. $A_i = s_i$, אם A_i זמן הגעה פנימי;

E_3. $A_i \leq t_i$, אם A_i זמן הגעה חיצוני;

E_4. אם $A_k < A_l$ הן הגעות פנימיות עוקבות, אזי

$$s_k = A_k < A_{k+1} < A_{k+2} < \ldots < A_{l-1} < A_l = s_l.$$

אם A_l ההגעה הפנימית הראשונה, אזי

$$t_0 < A_1 < A_2 < \ldots < A_{l-1} < A_l = s_l,$$

אם A_k ההגעה הפנימית האחרונה, אזי

$$s_k = A_k < A_{k+1} < A_{k+2} < \ldots < A_n \leq t_n.$$

(מאורע E_4 נובע ממדיניות השירות FCFS.)

נתבונן בשתי הגעות פנימיות עוקבות $A_k = s_k$ ו-$A_l = s_l$. נאמוד את המספר ההגעות המצטבר $A(t)$ על הקטע $[s_k, s_l]$. (אמידה של $A(t)$ שקולה לאמידה של $Q(t)$.)

מקרה ראשון: $t_{k+1} > s_l$.

אזי מאורע E_3 נובע ממאורע E_4, ל-$i \in \{k+1, \ldots, l-1\}$. מאורע E_4 והתכונה הידועה של תהליך פואסון גוררים ש-$A_{k+1}, \ldots, A_{l-1}$ מתפלגים כסטטיסטי הסדר של התפלגות אחידה על $[s_k, s_l]$. לכן, אם $t \in [s_k, s_l)$, $0 \le i \le l-k-1$, אזי

$$\begin{aligned} \mathrm{P}\{A(t) = k+i/E_r\} &= \binom{i}{l-k-1} \frac{(t-s_k)^i (s_l-t)^{l-i-1}}{(s_l-s_k)^{l-k-1}}, \\ \hat{A}(t) = E[A(t)/E_r] &= k + (l-k-1)\frac{t-s_k}{s_l-s_k}, \\ Var[A(t)/E_r] &= (l-k-1)\frac{(t-s_k)(s_l-t)}{(s_l-s_k)^2}, \end{aligned}$$

כאשר E_r הוא המידע הרלוונטי. אם $t \in [t_i, t_{i+1})$, $0 \le i \le n-1$, אזי

$$\hat{Q}(t) = \hat{A}(t) - i.$$

מקרה שני: $t_{k+1} < \ldots < t_m < s_l < t_{m+1} < \ldots < t_l$.

אנו נבנה Hidden Markov Model(HMM) ברוח מאמר של Daley ו-Servi. נתבונן בשרשרת מרקוב:

$$\{A(s_k), A(t_{k+1}), \ldots, A(t_m), A(s_l).\}$$

תנאי השפה הם $A(s_k) = k$ ו-$A(s_l) = l$.

אם $k+1 \le i \le m$, אזי $A(t_i) \in \mathcal{G}_i$, כאשר

$$\mathcal{G}_i = \{i, i+1, \ldots, l-1\},$$

נקראות קבוצות אפשריות.

אם נסמן

$$z_k = s_k,\ z_{k+1} = t_{k+1}, \ldots, z_m = t_m,\ z_{m+1} = s_l,$$

אזי הסתברויות המעבר של השרשרת שלנו ניתנות לחישוב לפי:

$$\tilde{p}_i(u,v) = P\{A(z_i) = v/A(z_{i-1}) = u\} =$$

$$= e^{-\lambda(z_i - z_{i-1})} \frac{(\lambda(z_i - z_{i-1}))^{v-u}}{(v-u)!}, \quad k+1 \le i \le m,\ i \le u \le v \le l-1,$$

$$\tilde{p}_{m+1}(u,v) = P\{A(z_{m+1}) = l / A(z_m) = u\} =$$

$$= e^{-\lambda(z_l - z_m)} \frac{(\lambda(z_{m+1} - z_m))^{l-1-u}}{(l-1-u)!}, \quad m \le u \le l-1.$$

אלגוריתם אמידה.

1°. נחשב את מטריצות הסתברויות המעבר:

$$\tilde{P}_i = \{\tilde{p}_i(u,v),\ u \in \mathcal{G}_{i-1},\ v \in \mathcal{G}_i\}, \quad k+1 \le i \le m+1.$$

2°. נחשב את מטריצות הסתברויות הטאבו:

$$\begin{aligned} W^{k,k+1} &= \tilde{P}_{k+1}, \\ W^{k,i} &= W^{k,i-1}\tilde{P}_i,\ k+2 \le i \le m+1, \\ W^{m,m+1} &= \tilde{P}_{m+1}, \\ W^{i,m+1} &= \tilde{P}_{i+1}W^{i+1,m+1},\ k+1 \le i \le m-1. \end{aligned}$$

3°. נחשב את ההתפלגות המותנה של מספר ההגעות המצטבר:

$$P\{A(t_i) = u/E_r\} = P\{A(t_i) = u / A(s_k) = k; A(s_l) = l; A(t_r) \in \mathcal{G}_r,\ k+1 \le r \le m-1\} =$$
$$= \frac{w_{k,u}^{k,i} w_{u,l}^{i,m+1}}{W^{k,i} W^{i,m+1}}.$$

4°. נחשב את ההתפלגות המותנה של גודל התור בזמני הטרנסאקציות:

$$\hat{Q}(t_i) = \hat{A}(t_i) - i, \quad k+1 \le i \le m.$$

(אמד גודל התור לינארי בין זמני הטרנסאקציות.)

4.3 זמני מעבר אקספוננציאלים, הגעות חיצוניות ידועות

שוב נסמן ב- $t_0, t_1, \ldots, t_n, t_{n+1}$, זמני טרנסאקציות בתקופת תעסוקה בודדת. אם צרכן נרשם ב-t_i, אזי s_i יסמן זמן של הטרנסאקציה האחרונה שלו. ניתן לחשב את הזמן הזה בעזרת ID-ים. זמני הגעות נתונים על ידי:

$$A_i = s_i + \tau_i, \quad 1 \le i \le n,$$

כאשר $\tau_i \sim exp(\gamma_i)$, γ_i הם קצבי המעבר המתאימים ו $1 \leq i \leq n$, τ_i, בלתי תלויים זה בזה.

המידע הרלוונטי שווה לחיתוך של שני מאורעות:

E_1. $A_1 \leq t_1,\ A_2 \leq t_2, \ldots, A_n \leq t_n$;

E_2. $t_0 = A_0 < A_1 < A_2 < \ldots A_n \leq t_n$.

נתאר שרשרת מרקוב

$$\tilde{A} = \{\tilde{A}(t_0), \ldots, \tilde{A}(t_n)\}.$$

מצב השרשרת ב-t_i זיהה לקבוצת הצרכנים שהגיעו בתוך $(t_0, t_i]$. קבוצה זו מסודרת לפי זמני ההגעות.

תנאי השפה הם:

$$\tilde{A}(t_0) = \{\emptyset\};\ \tilde{A}(t_n) = \{1, \ldots, n\}.$$

קבוצת המצבים האפשריים בזמן t_i נתונה על ידי:

$$\mathcal{G}_i = \{\ \{1, 2, \ldots, i\},\ \{1, 2, \ldots, i, i+1\},\ \{1, 2, \ldots, n\}\ \}.$$

הסתברויות המעבר של השרשרת כוללות ביטוים מסוג:

$$P\{\zeta_1 < \zeta_2 < \ldots < \zeta_n\},$$

כאשר $\zeta_1, \ldots, \zeta_n$ משתנים מקריים אקספוננציאלים קטומים. פרק מיוחד של העבודה מוקדש לחישוב הסתברויות אלו.

המקרה של זמני מעבר אקספוננציאלים, והגעות חיצוניות לא ידועות דומה למקרה שבו טיפלנו.

5 אמידה ב-Real-Time

במקרה של אמידה ב-real-time, המבנה של המידע הרלוונטי באמידת גודל התור הוא יותר מסובך. לכן פרק נפרד מוקדש לפיתוח שיטות מדויקות להשגה של מידע זה. הניתוח נעשה בעזרת הצגות שונות של המאורעות הניצפים.

בפרק נפרד אחר אנו עוסקים בבעיות של אינטרפולציה כללית על תחנה בודדת.

5.1 זמני מעבר מיידיים, הגעות חיצוניות ידועות

נסמן זמני טרנסאקציות ב-t^n_{ml} כאשר n אינדקס של תחנה, m מספר תקופת תעסוקה בתחנה n ו-l מספר טרנסאקציה בתוך תקופת תעסוקה. סיום שירות נקרא Invisible-Routing Time (מסומן $t^n_{ml} \in I_R(t)$) אם צרכן שסיים שירות ב-t^n_{ml} לא נרשם בשום מקום במערכת עד זמן t.

$C_j(t)$ מציין זמן הגעה לתחנה j של צרכן המקבל שירות בזמן t בתחנה j.
עבור $t^n_{ml} \ni I_R(t)$ נגדיר קבוצת תחנות ברות-השגה על ידי

$$\mathcal{M}_t(t^n_{ml}) = \{\{j : C_j(t) < t^n_{ml}\} \bigcup \{0\}\}.$$

אנו מבצעים את אמידת גודל התור בעזרת הנוסחה הבאה:

$$E[Q_j(t)/E_t] = \sum_{t^n_{ml} \in I_R(t)} 1_{\{j \in \mathcal{M}_t(t^n_{ml})\}} \frac{p_{nj}}{\sum_{k \in \mathcal{M}_t(t^n_{ml})} p_{nk}}.$$

הוצג אלגוריתם שמעדכן את האמד הזה באופן דינמי.

5.2 מקרים אחרים ב-Real-Time

במקרה של זמני מעבר מיידיים והגעות חיצוניות לא ידועות אנו משתמשים במינוח דומה, אולם החישובים מורכבים הרבה יותר.

כאשר המעברים בין תחנות אקספוננציאליים, ניתן להשתמש ב-Hidden Markov Models. (ניתחנו בצורה מפורשת רק את המודל עם מעברים מיידיים.) במקרה זה, מרחב המצבים של השרשרת מתפוצץ ולכן קשה לפתח אלגוריתמים שיעבדו ב-real-time.

6 שימושים

ישמנו את התוצאות התאורתיות שפיתחנו לנתוני תיפעול של סניף בנק. הגעות חיצוניות נירשמו, וסביר להניח שזמני המעבר בין התחנות כמעט מיידיים. לכן, בעזרת ID-ים, ניתן לחשב תורים באופן מדויק. כניסוי, הנחנו שההגעות החיצוניות לא ידועות והשתמשנו באלגוריתם של Hidden Markov Model. ברוב הניסויים קיבלנו התאמה טובה בין האמד והמציאות, אף-על-פי שהנחות המודל מתקיימות בבנק רק באופן חלקי (למשל, מדיניות FCFS או work-conserving).

מחקר עתידי יכול לכלול קירובים וישום של מודלים תאורתיים נוספים.