Bayesian hierarchical models in manufacturing bulk queues

Carmen Armero and David Conesa

Department of Statistics and Operations Research,

Universitat de València,

C/ Dr. Moliner 50,

46100 Burjassot (Valencia), SPAIN

Abstract

In this paper, Queueing Theory and Bayesian statistical tools are used to analyze the congestion of various manufacturing bulk service queues with the same characteristics that are working in equilibrium and independently one of each other. Hierarchical models are discussed in order to develop the whole inferential process for the parameters governing the system. Markov Chain Monte Carlo methods and numerical inversion of transforms are addressed to compute in practice the posterior predictive distributions of the usual measures of performance.

Key words and sentences: Hierarchical Bayesian analysis; Bulk service queues; Congestion in manufacturing systems; Markov Chain Monte Carlo integration; Numerical inversion of transforms; Simulation.

1 Introduction

Nowadays, manufacturers are forced to continuously improve the productivity so as to lower their manufacturing costs, while maintaining certain quality standards. In this context, they usually confront with two possible situations: some are interested in improving the quality and efficiency on a particular manufacturing system that is currently working, while others are interested in the possible behaviour of a new plant or facility to be installed having the same characteristics than the one already running. In both cases, in order to make strategic and tactical decisions, it becomes necessary to have some previous knowledge of the performance (productivity) of the former system.

In this paper, we use Queueing Theory and Bayesian statistical tools to analyze the congestion of a single-stage system with various independent bulk service queues (models in which the service facility can serve various customers simultaneously) with the same characteristics and running in equilibrium. This situation describes, for instance, a factory or a warehouse with various equal (or similar) machines (ovens, stamping presses machines, ...) with or without human operators, working independently (that is, generating their own lines) and, moreover, operating in a correct way for a long time (machines do not have to be necessarily in the same factory, they could also be in different geographical localizations).

There is a wide variety of bulk service models, depending on the specific arrival process of the customers (pieces, jobs, etc.), the service mechanism of the machine, the capacity of the whole system and the queue discipline (see Chaudhry and Templeton (1983) [12] for a detailed description of these models). Nevertheless, in this work, we will only concentrate on the most standard one, usually denoted as $M/M(a,b)/1$, where the customers arrive singly according to a Poisson process with arrival rate, λ; the service distribution is exponential with service rate, μ; the only server of the system operates according to the so-called *general bulk service rule* in which the batch service size is always at least the *quorum*, a, and can never exceed the maximum capacity of the server, that is b; the arrival and service processes are independent; the queue discipline for the groups is FIFO (first-in, first-out); and no restrictions on the capacity of the system are considered.

Bulk service queues are also very popular in many other practical situations such as transportation systems (buses, guided tours, medical evacuation systems, ...) and computing (information-transmission systems handling messages, ...). Two interesting references of the usefulness of these models are given by Griffiths (1995) [20] and Armero et al. (2001) [6]. In the first one, the author analyzes in depth the congestion of the Suez Canal by modeling the throughput of ships as a bulk service queue, while the second is a study of the behaviour of the waiting list for a renal transplant in the Comunitat Valenciana (Spain).

Our main purpose is to statistically analyze the congestion of a system formed by a set of K independent work stations generating, each one of them, a $M/M(a,b)/1$ queueing system in the steady-state (like, for instance, a system of elevators in a skyscraper, a set of furnaces for the curing process of the braking pads in the automotive friction industry, or even a service company offering tourist guides in order to enjoy a sightseeing tour around the city).

The remainder of the paper is organized as follows. In Section 2, we use Bayesian hierarchical models to develop the whole inferential process for the parameters of the system. In Section 3, we concentrate on the study and practical computation of the joint congestion of the system through the posterior predictive distribution of the most relevant measures of performance of each queue (number of customers or jobs in the queue, waiting times in the queue or in the system, busy and idle times, batch service size, etc.). Finally, in order to illustrate the obtained results, in Section 4, a numerical fictitious example that considers a plant with various machines is discussed.

2 The Bayesian hierarchical model for estimation.

In order to learn about the congestion of the system, we first make inference about the parameters governing the whole system, $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_K})$ with $\boldsymbol{\theta_i} = (\lambda_i, \mu_i)$, where λ_i and μ_i are, respectively, the arrival and the service rate of the i-th queue (i=1,...,K). Note also that the fact that all queues are consolidated and operating for a long time implies that the parameters can only move freely in the reduced parametric space $\Theta_e = \{(\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_K});\ \lambda_i < b\mu_i\,,\ i = 1, \ldots, K\}$ defined by the ergodic condition for each one of the K queues.

The problem of analyzing statistically a queueing system is not a new one, indeed, as Bhat et al. (1997) [9] stand, it conforms a very important field of research in Queueing Theory. Most of the works in this area are in the framework of the frequentist approach, but lately there has been an increasing interest in Bayesian methods (see Armero and Bayarri (1999) [4] and Conesa (2000) [13] for an exhaustive list of Bayesian references up to date).

It is well know that, in general, the two basic sources of information in Bayesian inference are the prior beliefs and the data. Then, and in order to estimate $\boldsymbol{\theta}$, our first step consists of assigning a prior distribution that suitably describes our prior knowledge about it.

The simplest approach to this problem could be to analyze separately each one of the queues assigning a prior distribution to the parameters of each queue independently from the other ones. On the contrary, we could carry out a joint analysis without any distinction among the different queues. Nevertheless, we will not pursue them here because, even though machines are identical, there is always variability in their performance (they are usually managed by human operators); and moreover, because it is reasonable to think that the arrival and service rates of all the queues of the system are related. Then, we are going to construct an specific model that allows us to include all these characteristics into the analysis.

We start assuming that the K pairs $\{\boldsymbol{\theta_i} = (\lambda_i, \mu_i);\ i = 1, \ldots, K\}$ corresponding to the K queues are a random sample from a common prior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ that depends on an unknown hyperparametric vector, $\boldsymbol{\phi}$. This kind of modelling requires a second level in which we must express our knowledge about $\boldsymbol{\phi}$ through another prior distribution, $p(\boldsymbol{\phi})$, usually known as hyperprior distribution. This resulting structure is usually known as Bayesian hierarchical model (for a detailed description of them, see for instance Berger (1985) [7], Gelman et al. (1995) [16] and more recently, Draper (2001) [14]).

Once we have specified all our prior knowledge of $(\boldsymbol{\theta}, \boldsymbol{\phi})$ through the joint prior distribution, $p(\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi})$, the following stage is to get experimental information. There are different designs to obtain sample information on queues (see Armero and Bayarri (1996) [3] for a revision on this topic). In this paper, we use an easy experiment that consists of observing, simultaneously or not, the arrival and the service process of each one of the queues. In particular, for the i-th queue, we record the interarrival times of n_{ai} consecutive customers and the times spent to serve n_{si} groups (where n_{ai} and n_{si} are fixed in advance). Then, in accordance to the general hypothesis of the $M/M(a,b)/1$ queue, the n_{ai} interarrival times, $\boldsymbol{Y_i} = (Y_{i\,1}, \ldots, Y_{i\,n_{ai}})$, are a random sample from $\mathrm{Ex}(\lambda_i)$ and the n_{si} recorded service times, $\boldsymbol{S_i} = (S_{i\,1}, \ldots, S_{i\,n_{si}})$, constitute a random sample from $\mathrm{Ex}(\mu_i)$.

The resulting likelihood function of $\boldsymbol{\theta}$ for all the data $\boldsymbol{z} = (\boldsymbol{z_1}, \ldots, \boldsymbol{z_K})$ (where $\boldsymbol{z_i} = (\boldsymbol{y_i}, \boldsymbol{s_i})$ represent the data of each queue) is:

$$\ell(\boldsymbol{\theta}) = \prod_{i=1}^{K} \ell_i(\lambda_i, \mu_i) \propto \prod_{i=1}^{K} \lambda_i^{n_{ai}} \mathrm{e}^{-\lambda_i t_{ai}} \mu_i^{n_{si}} \mathrm{e}^{-\mu_i t_{si}}\,, \quad \lambda_i < b\mu_i\,, \quad i = 1, \ldots, K\,, \tag{2.1}$$

where $t_{ai} = \sum_{j=1}^{n_{ai}} y_{ij}$ is the total time until the arrival of the n_{ai} customer considered in the i-th queue, and $t_{si} = \sum_{l=1}^{n_{si}} s_{il}$ is the total time taken by the server of the i-th queue to complete all the considered services. Note that the restrictions in the domain of the likelihood in (2.1) correspond to the ergodic condition of each queue.

Next, we consider the problem of assigning a prior distribution to the unknown parameters of the model. As mentioned above, we specify the joint prior distribution of $(\boldsymbol{\theta}, \boldsymbol{\phi})$ in two stages. In the first one, we assign a prior distribution, $p(\boldsymbol{\theta}|\boldsymbol{\phi})$, for $\boldsymbol{\theta}$ given the hyperparameters. Then, we select an hyperprior distribution $p(\boldsymbol{\phi})$ for those hyperparameters.

For simplicity, we select the conditional prior distribution for each pair $\boldsymbol{\theta_i} = (\lambda_i, \mu_i)$, as the product of two independent Gamma distributions (a member of the common natural conjugate prior distribution that would result from an individual analysis of each queue):

$$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = \prod_{i=1}^{K} p(\lambda_i, \mu_i|\boldsymbol{\phi}) = \mathrm{C}^K \prod_{i=1}^{K} \mathrm{Ga}(\lambda_i|\alpha_0, \beta_0)\, \mathrm{Ga}(\mu_i|a_0, b_0)\,, \ \ \lambda_i < b\,\mu_i\,, \ \ i = 1, \ldots, K\,, \tag{2.2}$$

where the hyperparameters are $\phi = (\alpha_0, \beta_0), a_0, b_0)$ and the common proportionality constant corresponding to each queue is:

$$\mathrm{C}^{-1} = \frac{\Gamma(a_0 + \alpha_0)}{\Gamma(\alpha_0)\Gamma(a_0)} \frac{b_0^{a_0} (b\beta_0)^{\alpha_0}}{\alpha_0 (b\beta_0 + b_0)^{(a_0+\alpha_0)}} \, F\left(1, \ a_0 + \alpha_0; \ \alpha_0 + 1; \ \frac{b\beta_0}{b\beta_0 + b_0}\right)\,,$$

being $F(a, b; c; z)$ the Gauss's hypergeometric function (Gradshteyn and Ryzhik (1965) [19]).

Now, we have to choose an hyperprior distribution, $p(\boldsymbol{\phi})$. In general, the specification of these type of distributions is a very challenging problem in Bayesian statistics (see, for instance, Berger (1985) [7]). In spite of this, it is worth to note that in situations when the prior knowledge about $\boldsymbol{\phi}$ is vague, we can always use non informative priors (never forgetting that when using improper prior densities, we must always check if the corresponding posterior distribution is proper). Nevertheless, in most real problems, one should always have enough knowledge about the hyperparameters (maybe from historical data) at least to constrain them into a finite region, if not to assign a not vague hyperprior. In both cases, note that sometimes it will be easier to assign a prior distribution for that reparameterization of the hyperparameters that better describes our knowledge about them (see García-Donato (2001) [15] and Gelman et al. (1995) [16] for examples of this situation).

An alternative way of estimating the hyperparameters is to use the so-called empirical Bayes (EB) methods, that basically consist of using the experimental information to produce an estimate, $\hat{\boldsymbol{\phi}}$, of the unknown hyperparameters, $\boldsymbol{\phi}$, and then proceed in the usual Bayesian way, but working with $p(\boldsymbol{\theta}|\hat{\boldsymbol{\phi}})$ as the prior distribution of the parametric vector $\boldsymbol{\theta}$. Lehozcky (1990) [22], Thiruvaiyaru and Basawa (1992) [31], Sohn (1996) [29] and Armero and Conesa (2000) [5] have applied EB methods to the statistical analysis of different queueing systems.

There has been a lot written about advantages and disadvantages of both procedures and so we will not pursuit here this question. Nevertheless, as Berger (1985) [7] stands, EB methods work fine when the number of queues is high because the $\boldsymbol{\theta_i}'s$ are a random sample from the general population $p(\boldsymbol{\theta}|\boldsymbol{\phi})$, and so, the bigger the sample, the better the information about $\boldsymbol{\phi}$. For small samples, however, these methods do not reflect the errors introduced in the hyperparameters estimation, and only in those particular situations in which all the possible values of $\boldsymbol{\phi}$ are mostly concentrated in their estimation $\hat{\boldsymbol{\phi}}$, the error would be small leading to a good performance of the method. Robbins (1983) [28], Maritz and Lwin (1989) [23] and Carlin and Louis (2000) [10] are very good references in the subject.

Once the elicitation of $p(\boldsymbol{\phi})$ is done, the joint prior distribution that contains all our prior knowledge about the parameters and hyperparameters is:

$$\begin{aligned} p(\boldsymbol{\theta},\boldsymbol{\phi}) &= p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi}) = p(\boldsymbol{\phi})\prod_{i=1}^{K} p(\boldsymbol{\theta_i}|\boldsymbol{\phi}) = \\ &= \mathrm{C}^K p(\alpha_0,\beta_0,a_0,b_0)\prod_{i=1}^{K}\mathrm{Ga}(\lambda_i|\alpha_0,\beta_0)\mathrm{Ga}(\mu_i|a_0,b_0)\,, \quad \lambda_i < b\,\mu_i\,,\ i=1,\ldots,K\,. \end{aligned} \tag{2.3}$$

Then, this prior information is updated via Bayes' theorem with the experimental one contained in the likelihood $\ell(\boldsymbol{\theta})$ in (2.1), to get the joint posterior distribution:

$$\begin{aligned} p(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z}) &\propto \ell(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi}) = \\ &= \mathrm{C}^K p(\alpha_0,\beta_0,a_0,b_0)\prod_{i=1}^{K}\lambda_i^{n_{ai}}\mathrm{e}^{-\lambda_i t_{ai}}\mu_i^{n_{si}}\mathrm{e}^{-\mu_i t_{si}}\mathrm{Ga}(\lambda_i|\alpha_0,\beta_0)\mathrm{Ga}(\mu_i|a_0,b_0)\,, \end{aligned} \tag{2.4}$$

being obviously $\lambda_i < b\,\mu_i$, for $i = 1,\ldots,K$.

All the elements and relationships of this hierarchical structure are represented by the *graphical model* in figure 1. Obviously, the usefulness of these models goes beyond the visual representation of our hierarchical Bayesian model. For a detailed discussion of definitions and properties of graphical models, see for instance Lauritzen et al. (1996) [21].

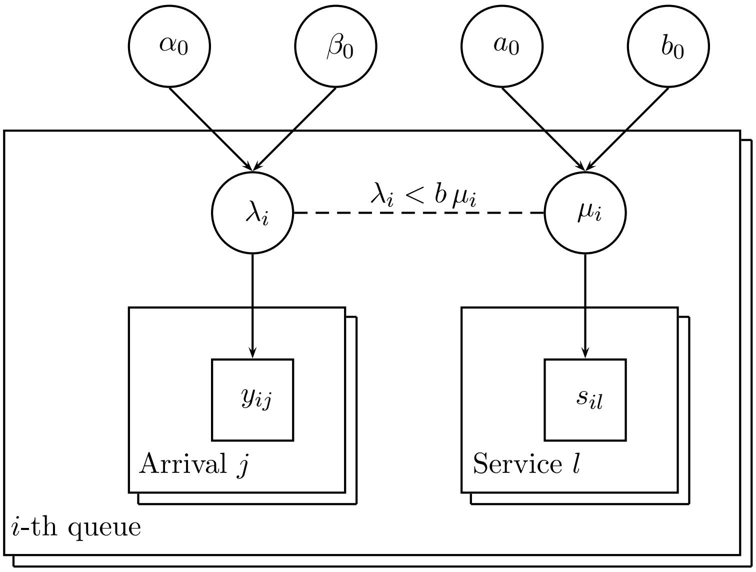

Figure 1: *Graphical model corresponding to the hierarchical Bayesian model for the system with K bulk queues $M/M(a,b)/1$ working in equilibrium.*

The distribution in (2.4) contains all our knowledge about the system, but it does not have a known analytical expression. Then, in order to provide a more intuitive information about $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, numerical methods will be needed to evaluate in practice the main characteristics of this posterior distribution. Among others, one feasible (indeed the most popular) possibility is to use Markov chain Monte Carlo (MCMC) methods, that draw samples from any intractable posterior by running for a long time a cleverly constructed Markov chain which stationary distribution is just the one we want to simulate from. Among the different ways of building these chains, the most popular are Gibbs sampling and Metropolis-Hastings algorithm (see Gilks et al. (1996) [17] for a detailed discussion of them).

When, as in our case, the distribution of interest is multivariate, these methods produce a random sample from the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{z})$ in (2.4). Then, we can use Monte Carlo integration methods to approximate the main features of the marginal posterior distribution of the arrival, service rates and, consequently, of the traffic intensity of each one of the K queues. These marginal distributions will provide us the first intuitive impression about the congestion and performance of the queues we were looking for. MCMC methods has also been successfully used by Rios Insua et al. (1998) [27] in order to deal with the posterior distribution of the parameters in a $M/H_k/1$ queueing system.

3 Predicting the whole congestion of the system

As we have just commented, MCMC allow us to extract mainly all the information contained in the joint posterior distribution in (2.4) about the parameters and hyperparameters, $(\boldsymbol{\theta}, \boldsymbol{\phi})$, managing this system. Nevertheless, it is well know that, in Queueing Theory, when a system is working in the steady-state (like the one we are dealing with), its congestion is better described through the so-called *measures of performance* of the queue (number of customers in the system, waiting times in the system and in the queue, idle and busy periods, etc.).

In particular, and depending on the manufacturers' objectives, we could be interested either in the behaviour of all the K queues observed or in the performance of a generic one not included in the experiment (this would be appropriate in the case of an identical machine that is yet working in the plant, or even, as part of a future study about the possible performance of a new factory that will operate with the same type of machines).

In general, the posterior predictive distribution of a congestion variable can be obtained through its steady-state distribution given the parameters of the model and the current information about them expressed in terms of their marginal posterior distribution. But, in our system, this general procedure gets very complicated in practice: the conditional distribution of these variables is usually expressed is terms of transforms without analytical inversion, and we have not a closed expression for the corresponding marginal posterior distribution of the parameters and we can only work with simulated samples from it.

In this Section, we will discuss some practical questions involved when computing the posterior predictive distribution of the measures most often used to describe congestion in bulk service queues, specially those presenting computational problems. Of course, all these measures are interconnected, and their utility in a real problem will depend on its particular characteristics and the objectives of the manufacturers who are analyzing (or paying for) it. This Section is organized in four subsections. The first three ones are devoted to the K sampled queues, while the last one deals with the study of the behaviour of a new queue not included at the beginning in the analysis.

3.1 The state of the system

Let (H_i, Nq_i) denote the *state* of the i-th queue where $H_i = 1$ means that the server is busy processing a batch of s items, $a \leq s \leq b$, and $H_i = 0$ indicates that the server is idle because the number of items in the system is less than the quorum a. In both cases, Nq_i is the number of customers waiting in queue. Given the arrival and service rate, its probability distribution is (Medhi (1991) [24]):

$$\begin{aligned}
p\Big((0,0)\Big|r_i\Big) &= \left[\frac{a}{1-r_i} + \frac{r_i^{a+1} - r_i^{b+1}}{(1-r_i)^2}\right]^{-1}, \\
p\Big((0,n)\Big|r_i\Big) &= \left(\frac{1-r_i^{n+1}}{1-r_i}\right) p\Big((0,0)\Big|r_i\Big), \qquad n = 1, \ldots, a-1 \\
p\Big((1,n)\Big|r_i\Big) &= \left(\frac{1-r_i^{b}}{1-r_i}\right) p\Big((0,0)\Big|r_i\Big)\, r_i^{n+1}, \quad n = 0, 1, \ldots
\end{aligned}$$

where, for simplicity, $p\Big((h_i, n_i)\Big|r_i\Big)$ denotes the probability $p\Big((H_i, Nq_i) = (h_i, n_i)\Big|r_i\Big)$, and r_i is the unique real root in $(0,1)$ of the equation:

$$\frac{\lambda_i}{\mu_i} = r_i + r_i^2 + \ldots + r_i^b \,. \tag{3.1}$$

Taking into account that, given the parameters of the system, the K queues work independently, the joint conditional distribution that describes the state of the whole system, denoted from now on as $p\Big((h_1, n_1), \ldots, (h_K, n_K)\Big|r_1, \ldots, r_K\Big)$, is the product of its correspondent marginal distributions, $\prod_{i=1}^K p\Big((h_i, n_i)\Big|r_i\Big)$.

Then, the joint posterior predictive distribution of the state of all the queues would be:

$$\begin{aligned}
p\Big((h_1, n_1), \ldots, (h_K, n_K)\Big|\boldsymbol{z}\Big) &= \mathrm{E}^{(r_1,\ldots,r_K|\boldsymbol{z})}\left[p\Big((h_1, n_1), \ldots, (h_K, n_K)\Big|r_1, \ldots, r_K\Big)\right] \\
&= \int \prod_{i=1}^{K} p\Big((h_i, n_i)\Big|r_i\Big) p(r_1, \ldots, r_K|\boldsymbol{z}) \mathrm{d}(r_1, \ldots, r_K)\,. \quad (3.2)
\end{aligned}$$

But, obviously, we do not have a closed expression for $p(r_1, \ldots, r_K|\boldsymbol{z})$ nor for the joint posterior predictive either. Then, we can approximate the probabilities of this distribution, by using the following procedure based on MCMC methods:

1. From the random sample of size m of the joint posterior distribution in (2.4) obtained using MCMC techniques, consider the components, $\{(\boldsymbol{\theta_1}^{(j)}, \ldots, \boldsymbol{\theta_K}^{(j)});\ j = 1, \ldots, m\}$, as a random sample from the marginal posterior distribution of the parameters, $p(\boldsymbol{\theta}|\boldsymbol{z})$;

2. Generate a sample from $p(r_1, \ldots, r_K|\boldsymbol{z})$, where each generated component $r_i^{(j)}$ is obtained by solving the equation (3.1) for each simulated pair $\boldsymbol{\theta_i}^{(j)} = (\lambda_i^{(j)}, \mu_i^{(j)})$, $\forall\, i$;

3. Approximate each posterior predictive probability as:

$$\begin{aligned} p\Big((h_1, n_1), \ldots, (h_K, n_K)\Big|\boldsymbol{z}\Big) &\approx \frac{1}{m}\sum_{j=1}^{m} p\Big((h_1, n_1), \ldots, (h_K, n_K)\Big| r_1^{(j)}, \ldots, r_K^{(j)}\Big) \\ &= \frac{1}{m}\sum_{j=1}^{m}\prod_{i=1}^{K} p\Big((h_i, n_i)\Big| r_i^{(j)}\Big) . \end{aligned} \tag{3.3}$$

This procedure is similar to the one used in Wiper (1998) [33] to obtain the posterior predictive distribution of the number of customers in $E_r/M/1$ queueing systems.

From (3.3) it is easy to compute the posterior distribution of the situation of idleness or busyness of all the K machines, $p(h_1, \ldots, h_K|\boldsymbol{z})$. Note also that (3.3) does not provide an intuitive visualization of the congestion of each queue. To do so, it would be more suitable to analyze each one of them separately through their corresponding marginal distribution, $p\Big((h_i, n_i)\Big|\boldsymbol{z}\Big)$, that could be computed directly marginalizing (3.3), or maybe using a similar procedure to the one above presented, also based in MCMC methods:

$$p\Big((h_i, n_i)\Big|\boldsymbol{z}\Big) \approx \frac{1}{m}\sum_{j=1}^{m} p\Big((h_i, n_i)\Big| r_i^{(j)}\Big) , \tag{3.4}$$

where $\{r_i^{(j)};\ j = 1, \ldots, m\}$ is a random sample from the marginal distribution $p(r_i|\boldsymbol{z})$.

It would be very useful for manufacturing purposes (such as learning about the performance of the system, improving the system behaviour, or even for planning and building integrated decision projects) to increase our information about the system. Two ways of doing this will be to study the *service batch size* (a big value of its expected size would indicate that machines always work in their maximum capacity), or/and the *waiting time in the queue* for an arriving job (of special importance when dealing with pieces that can be damaged by a long wait). Obviously these special features could make the managers to think about the possibility of increasing the quorum of the machines, to buy new bigger machines, etc. For both congestion measures, their conditional distribution depends on the root of the equation (3.1), and so, in order to approximate their corresponding posterior predictive distribution, we could just reproduce similar procedures to the ones above presented.

3.2 Idle periods

As usually the interest of managers in charge of the system is to maximize its output (always with the last objective of maximizing benefits), it could be interesting to study how long the idle periods of the machines that made up the whole system are.

Due to the fact that, given all the parameters of the system, the K queues work independently, the joint probability density function (p.d.f.) of the duration D of an idle period in each queue will be the product of its correspondent marginal distributions,

$$d(t_1, \ldots, t_K | \lambda_1, \ldots, \lambda_K, r_1, \ldots, r_K) = \prod_{i=1}^{K} d_i(t_i | \lambda_i, r_i)\,,$$

where for the i-th queue, r_i is as in (3.1) and the conditional distribution of D_i is expressed in terms of the Laplace transform (L.T.) of their corresponding p.d.f. (Medhi (1991) [24]):

$$d_i^*(u|\lambda_i, r_i) = \int_0^\infty \mathrm{e}^{-ut_i} d_i(t_i|\lambda_i, r_i)\,\mathrm{d}t_i = \frac{\lambda_i(1-r_i)}{(1-r_i^a)} \frac{1}{(u+\lambda_i)^a} \frac{\lambda_i^a - [r_i(u+\lambda_i)]^a}{\lambda_i - r_i(u+\lambda_i)}\,. \tag{3.5}$$

This L.T. can not be analytically inverted, and so numerical inversion becomes necessary. This problem is not new in Queueing Theory. On the contrary, nowadays, computational analysis plays a fundamental role when analyzing practical congestion (Neuts (1981) [25], Abate and Whitt (1992) [1] and Tijms (1997) [32] are classical references in the subject).

Then, we could approximate the posterior predictive distribution of all the idle periods,

$$d(t_1, \ldots, t_K | \boldsymbol{z}) = \mathrm{E}^{((r_1,\lambda_1),\ldots,(r_K,\lambda_K)|\boldsymbol{z})} \left[\prod_{i=1}^{K} d_i(t_i|r_i, \lambda_i) \right], \tag{3.6}$$

using a procedure based, not only on MCMC methods, but also on algorithms for numerically inverting the involved L.T.'s:

1. Use the results in steps 1 and 2 of the general procedure given in Section 3.1 to compute (3.2), and consider $\{(\lambda_1^{(j)}, r_1^{(j)}), \ldots, (\lambda_K^{(j)}, r_K^{(j)}),\ j = 1, \ldots, m\}$ as a random sample from $p((\lambda_1, r_1), \ldots, (\lambda_K, r_K)|\boldsymbol{z})$.

2. For each simulated vector $(\lambda_i^{(j)}, r_i^{(j)})$, compute $d_i(t_i|\lambda_i^{(j)}, r_i^{(j)})$ by inverting numerically its L.T., $d_i^*(u|\lambda_i^{(j)}, r_i^{(j)})$, given in (3.5);

3. Approximate the p.d.f. of the joint posterior predictive distribution as:

$$d(t_1, \ldots, t_K|\boldsymbol{z}) \approx \frac{1}{m}\sum_{j=1}^{m}\prod_{i=1}^{K} d_i(t_i|\lambda_i^{(j)}, r_i^{(j)})\,, \tag{3.7}$$

It is clear that the corresponding marginal posterior predictive distribution for the idle period of each machine, more easily computed as

$$d_i(t_i|\boldsymbol{z}) \approx \frac{1}{m}\sum_{j=1}^{m} d_i(t_i|\lambda_i^{(j)}, r_i^{(j)})\,, \tag{3.8}$$

where $d_i(t_i|\lambda_i^{(j)}, r_i^{(j)})$ is obtained by numerically inverting $d_i^*(u|\lambda_i^{(j)}, r_i^{(j)})$ as in step 2 above, will provide a convenient visualization of the congestion of each machine.

Working in a similar way, we could also analyze the number of items in the queue when an idle period begins. Given all the parameters of the system, the joint distribution of the *number of items waiting to be processed by each machine when an idle period begins*, $p(Nd_1 = n_1, \ldots, Nd_K = n_K|r_1, \ldots, r_K)$, with $0 \leq n_i < a$, $i = 1, \ldots, K$, is the product of their corresponding marginal distributions, $p(Nd_i = n_i|r_i)$, now expressed in terms of their probability generating function (p.g.f.) (Medhi (1991) [24]):

$$G_{(Nd_i|r_i)}(u) = \sum_{n_i=0}^{\infty} p(Nd_i = n_i|r_i)\, u^{n_i} = \frac{(1-r_i)(1-(r_i u)^a)}{(1-r_i^a)(1-r_i u)}\,. \tag{3.9}$$

Then, in order to compute the corresponding joint posterior predictive distribution, we could use a similar procedure to the one presented in this Section to deal with idle periods in which, instead of inverting L.T.'s, we would need methods for numerically invert p.g.f.'s.

3.3 Busy periods

Another complementary way of describing the congestion of our system is analyzing the duration, B, of the busy periods. As usual, its joint conditional (on the parameters) distribution is the product of its corresponding marginal distributions, which p.d.f.'s are given in terms of their L.T.'s (Medhi (1991) [24]):

$$b_i^*(u|\mu_i, R_i) = \frac{\mu_i(1 - R_i^a)}{u + \mu_i - \mu_i R_i^a}\,, \tag{3.10}$$

where here R_i is the unique real root in $(0, 1)$ of the equation:

$$\mu_i R_i^{b+1} - (u + \lambda_i + \mu_i)R_i + \lambda_i = 0\,. \tag{3.11}$$

Then, in order to approximate the joint posterior predictive distribution of the duration of a busy period for each machine, $b(t_1, \ldots, t_K|\boldsymbol{z})$, we could use a similar procedure to the one presented in Section 3.2 for the idle periods:

1. From the generated sample of $p(\boldsymbol{\theta}, \boldsymbol{\phi})$ in (2.4) consider $\{(\boldsymbol{\theta_1}^{(j)}, \ldots, \boldsymbol{\theta_K}^{(j)});\, j = 1, \ldots, m\}$ as a random sample from the marginal posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{z})$;

2. For each simulated pair $\boldsymbol{\theta_i}^{(j)}$, compute $b_i(t_i|\mu_i^{(j)}, R_i^{(j)})$ by numerical inversion of its corresponding L.T. $b_i^*(u|\mu_i^{(j)}, R_i^{(j)})$, taking into account that the equation (3.11) must be solved at every evaluation of the L.T. necessary for its numerical inversion.

3. Approximate the p.d.f. of the joint posterior predictive distribution as:

$$\begin{aligned} b(t_1, \ldots, t_K|\boldsymbol{z}) &\approx \frac{1}{m}\sum_{j=1}^{m} b(t_1, \ldots, t_K|\mu_1^{(j)}, \ldots, \mu_K^{(j)}, R_1^{(j)}, \ldots, R_K^{(j)}) \\ &= \frac{1}{m}\sum_{j=1}^{m}\prod_{i=1}^{K} b_i(t_i|\mu_i^{(j)}, R_i^{(j)})\,. \end{aligned} \tag{3.12}$$

As in the previous measures, in order to have a more convenient visualization of the congestion of each queue, we can consider their corresponding marginal posterior predictive distribution, $b_i(t_i|\boldsymbol{z})$, that could be approximated as:

$$b_i(t_i|\boldsymbol{z}) \approx \frac{1}{m}\sum_{j=1}^{m} b_i(t_i|\mu_i^{(j)}, R_i^{(j)})\,, \tag{3.13}$$

where each $b_i(t_i|\mu_i^{(j)}, R_i^{(j)})$ is obtained by numerically inverting its L.T. $b_i^*(u|\mu_i^{(j)}, R_i^{(j)})$ as stated in stage 2 before introduced.

3.4 Performance of a new machine

In the previous subsections, we have been describing the congestion of the K queues considered in our system. Now, our interest (as it could be the manufacturers' one) will be in the performance of a new machine with the same characteristics that the K ones previously studied but not included in the analysis up till now. This situation would be the case of an existing machine that is already working in the factory, or even, a generic new one to be installed in other factory belonging to the same company. It is worth noting that the analysis here presented could be easily generalized for more than one new machines.

With the aim of predicting the performance and congestion of this new queue (denoted as the $(K+1)$-th queue), our first interest is to make inference on its parametric vector, $\boldsymbol{\theta_{K+1}} = (\lambda_{K+1}, \mu_{K+1})$, formed by its arrival and service rate. If we assume that this machine works (or would work) like the previous K queues already studied, we can benefit from all the hierarchical structure built before, and consider that $\boldsymbol{\theta_{K+1}}$ would have to behave as its predecessors $\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_K}$, in the form described by their conditional distribution:

$$p(\boldsymbol{\theta_{K+1}}|\boldsymbol{\phi}) \propto \mathrm{Ga}(\lambda_{K+1}|\alpha_0, \beta_0)\mathrm{Ga}(\mu_{K+1}|a_0, b_0)\,, \qquad \lambda_{K+1} < b\,\mu_{K+1}\,, \tag{3.14}$$

where now our knowledge of $\boldsymbol{\phi}$, has been updated by the information contained in the K queues and is expressed through the marginal posterior distribution $p(\boldsymbol{\phi}|\boldsymbol{z})$.

Then, our whole knowledge about this new queue is expressed by:

$$p(\boldsymbol{\theta_{K+1}}, \boldsymbol{\phi}|\boldsymbol{z}) = p(\boldsymbol{\theta_{K+1}}|\boldsymbol{\phi})p(\boldsymbol{\phi}|\boldsymbol{z})\,, \tag{3.15}$$

and note that samples from this distribution can be easily obtained by firstly drawing a value from the marginal posterior distribution $p(\boldsymbol{\phi}|\boldsymbol{z})$, and then, given this posterior draw, simulate from the common conditional distribution in (3.14). Obviously, we could give a more intuitive prediction of the behaviour of that new machine through the posterior predictive distribution of the different measures of performance before presented. We could compute them following the same procedures that the ones presented in previous Sections for each one of the predictive distributions corresponding to a single machine but considering now the information about $(\boldsymbol{\theta_{K+1}}, \boldsymbol{\phi})$ as the one expressed in (3.15).

4 A numerical example with machines and facilities.

In this final Section, with the main objective of illustrating the previously presented results, we analyze the congestion in a manufacturing scenario formed by $K = 6$ independent machines that work processing items in packets according to a $M/M(9,20)/1$ queue model.

For each machine i of the system, $i = 1, \ldots, 6$, table 1 displays the results of a simulated experiment like the one proposed in Section 2. In particular, it contains the interarrival time of $n_{ai} = 20$ consecutive items, the time needed to process $n_{si} = 20$ packets of pieces and the maximum likelihood estimate (m.l.e.) of the traffic intensity,

$$\hat{\rho}_i = \frac{\hat{\lambda}_i}{20\hat{\mu}_i} = \frac{n_{ai}/t_{ai}}{20\,(n_{si}/t_{si})} = \frac{t_{si}}{20\,t_{ai}}\,.$$

Machine i	1	2	3	4	5	6
t_{ai}	1.39	0.94	1.16	0.84	1.04	0.92
t_{si}	15.88	18.67	13.59	19.55	30.24	16.99
$\hat{\rho}_i$	0.56	0.99	0.58	1.16	1.44	0.92

Table 1: *Sufficient statistics for the arrival and service rate on each machine corresponding to the total time (t_{ai}) to achieve the arrival of 20 consecutive pieces and the processing time (t_{si}) for 20 packets of pieces, respectively, and the m.l.e. ($\hat{\rho}_i$) of the traffic intensity for each machine.*

Note that, although we are considering that all the machines are similar and work in equilibrium, results are quite different between them. In particular, we observe a great congestion in the performance of machines 4 and 5 (in both cases, and in spite of the fact that in each machine the true value of the traffic intensity is less than one, the corresponding m.l.e.'s, $\hat{\rho}_4$ and $\hat{\rho}_5$, are greater than this critical value), while machines 1 and 3 seem to have a very good performance ($\hat{\rho}_1 = 0.56$ and $\hat{\rho}_3 = 0.58$). It is important to point out that if we did not distinguish among the different machines and considered that all the data come from an unique machine, the m.l.e. of the common ρ would be $\hat{\rho} = 0.91$, an intermediate level between the lower values for machines 1 and 3 and the upper ones of machines 4 and 5.

With the unique idea to serve as a reference framework, we start the statistical analysis of the congestion of this system by considering the two most extreme possible situations:

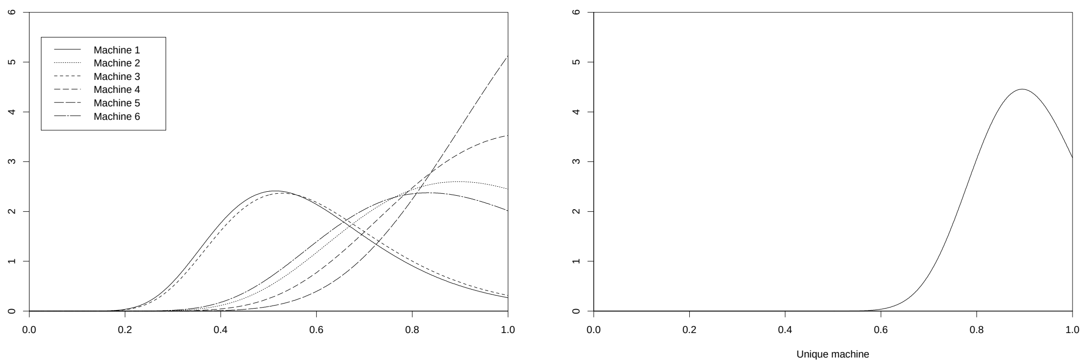

Figure 2: *Marginal posterior posterior distribution of the traffic intensity of each machine when they are analyzed : (a) separately; (b) together without any distinction among them.*

each machine is analyzed separately without sharing any common information with the other, or, all machines are studied together as if they were only one machine.

Firstly, we present the unpooled analysis. Figure 2(a) shows the marginal posterior distribution of the traffic intensity for each machine. In particular, the posterior one corresponding to machine i has been obtained assigning a Jeffrey's non-informative prior for the corresponding parametric vector $\boldsymbol{\theta_i}$ and combining, via Bayes' theorem, the data collected only on that machine. In case of analyzing all the machines together, figure 2(b) displays the posterior distribution of the traffic intensity computed from the posterior distribution of the common arrival and service rate computed using a Jeffrey's non informative prior and the data collected in all the machines (a total of 120 interarrival times and 120 service times).

In both graphics the assumption of equilibrium makes the right tail of all the densities not to approach to zero when the traffic intensity goes to the critical value 1 (this special feature is discussed in depth in Armero and Bayarri (1994b) [2]). Note also, that the different sample sizes considered when computing both posterior distributions are evident in the smaller variability of the posterior distribution of traffic intensity in the pooled analysis.

After these two extreme analysis, we will analyze the congestion of our manufacturing system by using a full Bayesian hierarchical model that allows us to maintain the particular features of each machine (and/or its human operator) and also incorporates the common ones that share all of them. To do so, we assume that the parameters $\{\boldsymbol{\theta_i} = (\lambda_i, \mu_i);\, i = 1, \ldots, 6\}$ of each machine are a random sample from the common population distribution in (2.2):

$$p(\lambda_i, \mu_i|\boldsymbol{\phi}) \propto \mathrm{Ga}(\lambda_i|\alpha_0, \beta_0)\mathrm{Ga}(\mu_i|a_0, b_0)\,, \quad \lambda_i < 20\,\mu_i\,, \tag{4.1}$$

where $\boldsymbol{\phi} = (\alpha_0, \beta_0, a_0, b_0)$ represents the vector of unknown hyperparameters.

Then, we express our previous knowledge about $\boldsymbol{\phi}$. At this stage, with the only objective of better understanding how Bayesian hierarchical models work in queues, we consider two possible situations: a first one in which our knowledge about $\boldsymbol{\phi}$ could be considered as scarce, being our selection for the hyperprior distribution, $p(\boldsymbol{\phi})$, the product of four (one for each hyperparameter) exponential distributions with mean 1/0.01, Ex(0.01), and a second one, in which we express a larger knowledge about $\boldsymbol{\phi}$ with the product of four independent Ex(0.1). From now, we will refer to this first situation as *LIH* (*L*east *I*nformative *H*yperprior), and to the second one as *MIH* (*M*ore *I*nformative *H*yperprior). In both cases, we have assumed independence among the hyperparameters, and also chosen proper distributions in order to obtain proper joint posterior distributions.

In both situations, the joint posterior distribution of the parameters and hyperparameters, $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{z})$, although proper, is analytically intractable. Then, we will approximate these distributions by simulation using Markov Chain techniques. One way for doing this is via the *Gibbs sampler* (see, for instance Casella and George (1992) [11], for a detailed description of this method), successfully implemented by Gilks et al. (1994) [18] in a program called BUGS © (*Bayesian inference Using Gibbs Sampling*). In particular, we work with its Windows © version, WinBUGS, developed by Spiegelhalter et al. (1999) [30] (see `http://www.mrc-bsu.cam.ac.uk/bugs/` for more information about this project).

To do so, we first choose a vector of starting values (for instance, $\lambda_i^{(0)} = 17$, $\mu_i^{(0)} = 1$, $i = 1, \ldots, 6$; $\alpha_0^{(0)} = a_0^{(0)} = \beta_0^{(0)} = b_0^{(0)} = 10$). Then, we decide the duration of the burn-in period (moment when we think that convergence to the stationary distribution has been reached) to be the first 5000 simulations and to work with the next 10000 ones. The obtained results with WinBUGS are a random sample (each element is a vector with 16 components, 2x6 parameters + 4 hyperparameters) from the joint posterior distribution

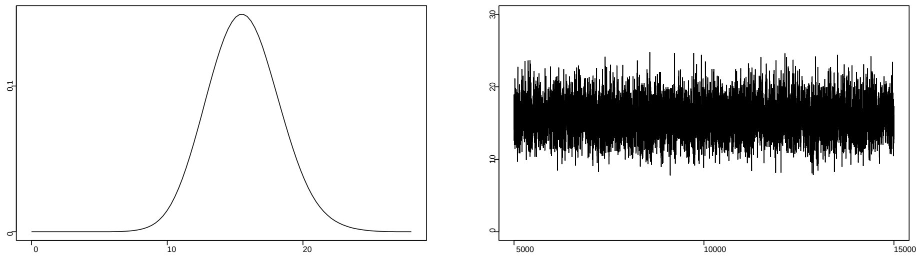

Figure 3: *Kernel estimation (on the left), based on the 10000 simulated values, of the marginal posterior distribution of* λ_1 *and the trace (on the right) of these 10000 values in the MIH case of the Bayesian hierarchical model application.*

$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{z})$.

As the way of presenting the results is similar for all the parameters, we only display here the ones for the arrival rate of pieces in the first machine, λ_1, for the *MIH* case. Figure 3 shows a kernel estimation of its marginal posterior distribution, $p(\lambda_1|\boldsymbol{z})$, based on the 10000 simulated values, and the trace of these 10000 values. Table 2 presents an estimation via Monte Carlo integration of its main characteristics.

Mean	Standard Deviation	Error M.C.	Quantile 2.5	Median	Quantile 97.5
15.63	2.455	0.03931	11.08	15.58	20.67

Table 2: *Summary statistics of the marginal posterior distribution of the arrival rate corresponding to the first machine,* $p(\lambda_1|\boldsymbol{z})$.

Convergence diagnosis of the resulting Markov chain is a main part of these kind of simulation techniques. To do so, we use CODA, a menu-driven set of S-Plus © functions which serve as an output processor for the BUGS © software (Best et al., 1995 [8]). Following the authors' recommendation, we have used a combination of diagnostic methods plus visual inspection of the trace plot and summary statistics to asses the convergence of the random samples generated by the WinBUGS software for all the parameters of this model with a reasonable degree of confidence.

From the obtained random sample of the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{z})$, we get

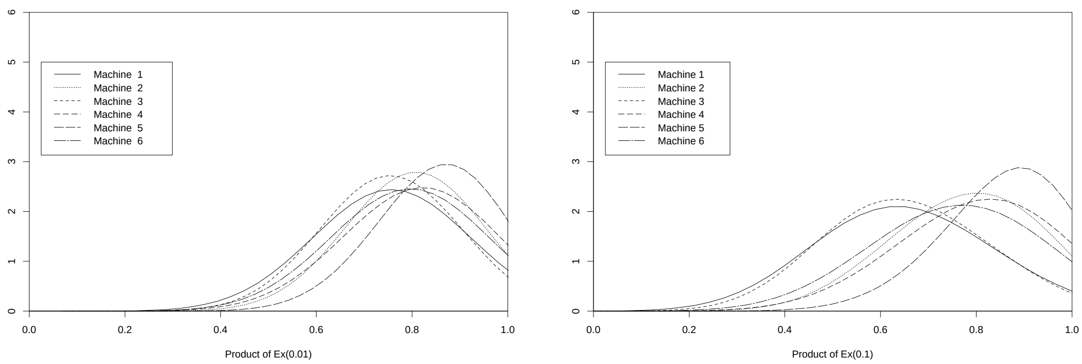

Figure 4: *Estimated marginal posterior distribution of the traffic intensity of each machine in the LIH case (on the left) and in the MIH (on the right).*

a random sample $\{\rho_i^{(j)}, j = 1, \ldots, 10000\}$ from the marginal posterior distribution of the traffic intensity for machine i, $p(\rho_i|\boldsymbol{z})$, $i = 1, \ldots, 6$. Figure 4 shows a kernel estimation of each one of these marginal distributions in both situations of prior knowledge about $\boldsymbol{\phi}$ considered. Note that, in the *LIH* case, the resulting densities are quite similar, indicating that the hierarchical structure gives more value to the common characteristics of the six machines than the particular differences among them. Nevertheless, when our knowledge about $\boldsymbol{\phi}$ is more accurate (the *MIH* case), the corresponding densities look more different because now the hierarchical model recognize and distinguish with more intensity the particular performance of each machine. However, if we compare the results in figure 2, where all the machines were analyzed separately, with the ones in the *MIH* situation, we observe how the model takes into account the common characteristics of the machines by smoothing a lot the high level of congestion on machine 5 and displacing the values of the traffic intensity on machines 1 and 3 toward more congested regions.

We continue with our analysis of the congestion by considering some of the measures of performance studied in Section 3. For simplicity, and in order not to extend unnecessarily the paper, we only deal here with the marginal posterior distribution of the state of the system, the waiting time in the queue and the duration of the busy periods. All them are analyzed in the unpooled case and also in the two applications of the Bayesian hierarchical model considered.

In these two latter situations, the marginal posterior distributions have been approximated using the random sample referred before of the joint posterior distribution obtained with WinBUGS and also, when needed, using methods for numerically invert the involved transforms. In particular, we work with Lattice-Poisson algorithm for inverting probability generating functions and the Euler algorithm to deal with Laplace transforms. Both methods are based on the Fourier series method and have been extensively discussed by Abate and Whitt (1992) [1]. Following these authors' recommendation, we also always use the Gaver-Stehfest algorithm (based on the Post-Widder formula) to check results given by Euler algorithm. As an exception, as Laplace transform of the conditioned distribution of the busy period is given in terms of a real root of an equation, we only use Gaver-Stefesth algorithm (which works with real variable) instead of Euler algorithm (which works with complex variable) to invert this particular transform.

In the unpooled analysis of the machines, the calculus of the relevant predictive distribution for each machine has been done using similar procedures to the ones in last Section based on Monte Carlo integration. Nevertheless, the expression of the joint posterior distribution of the arrival and service rate has usually a nice form (for instance, as a result of a conjugate analysis) from which it is easy to simulate without the need of MCMC.

In relation to the methods to find roots of the equations involved in these procedures, good results with moderate values of the quorum b motivated us to choose an hybrid algorithm based on a combination of Bisection and Newton-Raphson algorithms (see Press et al. (1986) [26] for a description of this routine). Obviously, this is our particular election here but other alternatives could have been used.

Figure 5 displays, for all the three situations of prior knowledge considered, the posterior predictive distribution of the state of each machine. The bars in each graphic are the probabilities of the number, N_q, of pieces waiting in queue to be processed when the machine remains idle (probabilities on the left, computed from 0 to 8 pieces, one unit less than the quorum $a = 9$) and when the machine is busy (probabilities on the right, computed for simplicity until the value N_q=14 pieces). The probability that the machine is idle is also presented inside the frame on the left corner.

Note that graphics in Figure 5 show the same general behaviour and relationships that the ones previously observed when dealing with the traffic intensity. In the unpooled

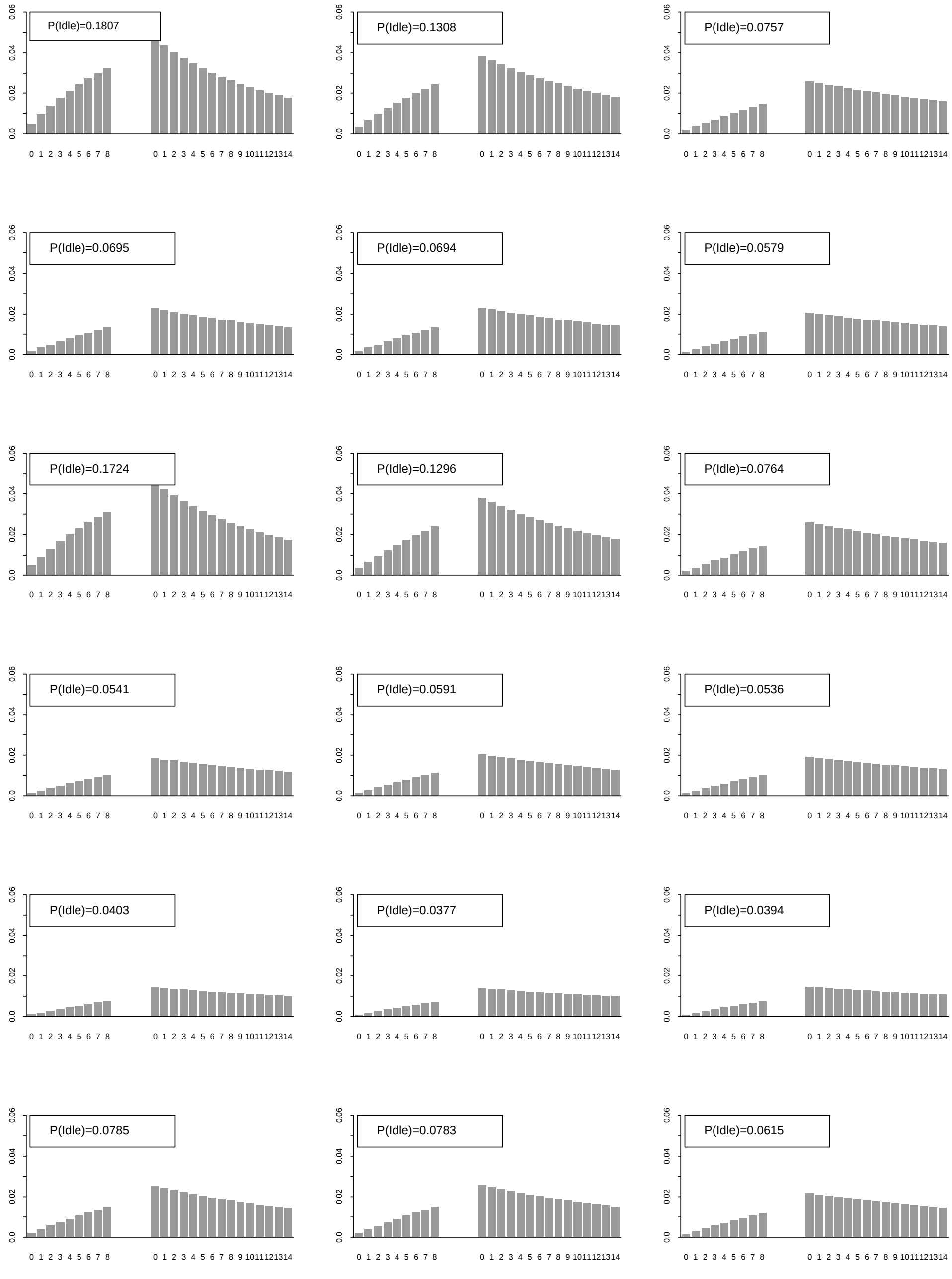

Figure 5: *Marginal posterior distribution of the state of every one of the six machines working in the facility. The bar charts in the column on the left correspond to the separate analysis; in the central column to the MIH Bayesian hierarchical model case; and the column on the right to the LIH Bayesian hierarchical model model case.*

analysis (on the left column), machines 1 and 3 seem to be the *less congested* (about 18.07% and 17.24% of the time are idle, respectively), while machines 4 and 5 are the *most congested* (with approximate values of idleness of 5.41% and 4.03%, respectively). Obviously, if this was a real problem these results would make people in charge of this facility to investigate about the particular characteristics of machines 1 and 3 and think over some future actions to increase their efficiency.

The partial pooling situation that incorporates the common characteristics among the machines through the hierarchical Bayes modelling is also presented in figure 5: bar charts on the central column correspond to the *MIH* case, and the ones on the right column correspond to the less informative hyperprior (*LIH*) distribution. If we compare them with the results of the separate analysis, we observe that the hierarchical model tends to make the performance of the machines uniform. In the *LIH* case, this effect is more intense in machines 1 and 3. They are idle in about 7.57% and 7.64% of the available time, quantities that are only a bit bigger than machines 4 and 5, the two most congested ones (with approximate percentages of 5.36% and 3.94%, respectively).

In the same way, if we compare the behaviour of each machine in the three situations considered, the differences in machines 1 and 3 are spectacular. For instance, the predictive probability that machine 1 is idle is 0.1807 in the separate analysis, 0.1308 in the *MIH* hierarchical model and 0.0757 in the *LIH*. But note that we can hardly observe any differences among the rest of the machines. The two most congested ones differ in less than one percentual unit the probability of being idle. For instance, in the most congested machine 5, in the unpooled analysis the probability of being idle is 0.0403, while in the *LIH* analysis is 0.0377 and 0.0394 in the *MIH*.

For all cases and machines, figure 6 displays the marginal posterior distribution (expressed through its cumulative distribution function) of the waiting time in the queue of a piece before to be processed. In these graphics, we can observe in general the same type of results and relationships than the ones previously commented in relation to the state of the system. Nevertheless, and observing basically the graphics corresponding to machines 1, 3 and 5, it seems that the waiting time in the queue, as a measure of performance of a queueing system, is less sensitive to the attempts of making the machines uniform of the Bayesian hierarchical model (in the *LIH* when the information about the hyperparameters

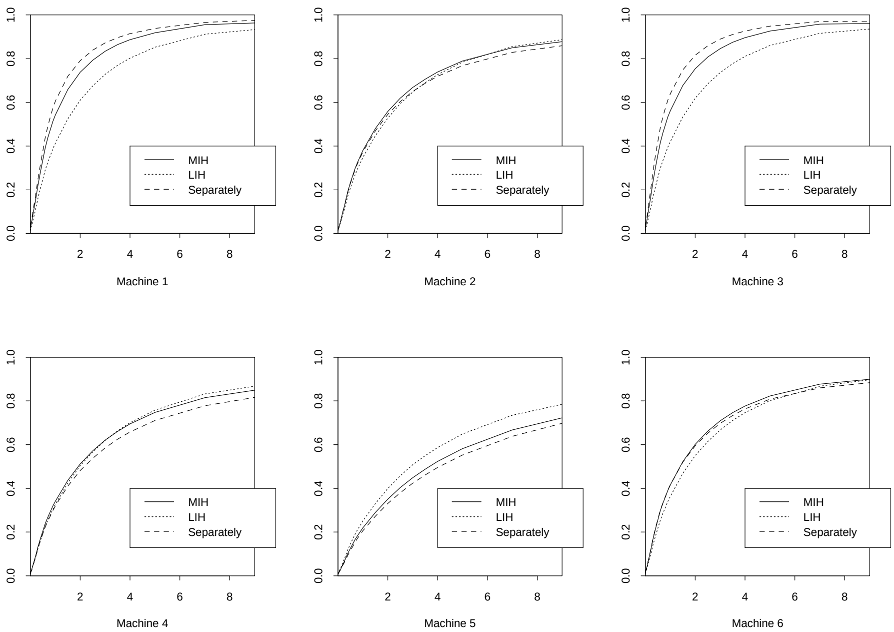

Figure 6: *Posterior cumulative distribution function of the waiting time in the queue for a piece in each one of the six machines in the separate analysis and in the LIH and MIH applications of the Bayesian hierarchical model considered.*

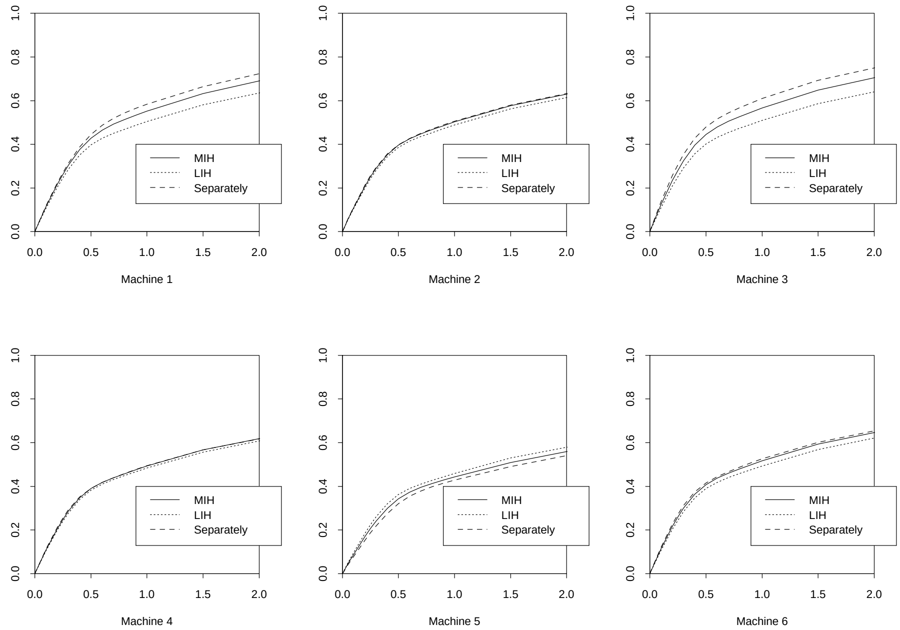

Figure 7: *Marginal posterior distribution of the duration of a busy period on every one of the six machines in the separate analysis and in the LIH and MIH applications of the Bayesian hierarchical model considered.*

is scarce) than it was in the case of the state of the system before discussed.

Another relevant thing to point out is that the posterior probability of a piece being processed immediately without waiting in the queue is so small that in all the graphics it seems to be zero. This predictive probability is the posterior probability that the machine is idle and, at the same time, there are 8 (the quorum of the machine minus one) pieces waiting in the queue for the arrival of a new piece that allows the machine to process a packet with nine pieces, that is $p\Big((0,8)\Big|\boldsymbol{z}\Big)$.

We finish our predictive analysis of this system with the study of the duration of the busy periods of each machine. Figure 7 displays the cumulative distribution function of the marginal posterior distribution of this variable in all the situations considered. We can observe small differences among the six machines, although now the machine 5 looks a bit different to the others ones. Also, it is worth noting that the separate analysis and the two applications of the Bayesian hierarchical model considered provide very similar results in the different machines of the manufacturing system. These features could indicate that the duration of a busy period is a robust measure of performance in relation to the use of an unpooled or a partial pooled analysis as the Bayesian hierarchical model here explored.

References

[1] J. Abate and W. Whitt, The Fourier-series method for inverting transforms of probability distributions, Queueing Systems 10 (1992) 5–87.

[2] C. Armero and M. J. Bayarri, Prior Assessments for prediction in queues, The Statistician 43 (1994), 139–153.

[3] C. Armero and M. J. Bayarri, Bayesian questions and answers in queues, in: *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Oxford University Press, 1996, pp. 3–23.

[4] C. Armero and M. J. Bayarri, Dealing with uncertainties in queues and networks of queues: a Bayesian approach., in: *Multivariate, Design and Sampling*, ed. S. Ghosh, New York: Marcel Dekker, 1998, in press.

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

[6] C. Armero, D. Conesa, J. Pérez, J.J. Abellán, M.J. García, M.A. Martínez, H. Vanaclocha and O. Zurriaga, Statistical analysis of the waiting list for renal transplants in the Comunitat Valenciana, in Abstracts of the EURO 2001 (European Operational Research Conference), p. 95.

[7] J. O. Berger, *Statistical decision theory and Bayesian analysis*, Springer Verlag, New York, 1985.

[8] N. G. Best, M. K. Cowles and S. K. Vines, CODA Convergence Diagnosis and Output Analysis software for Gibbs Sampler output: Version 0.3 (1995).

[9] U. N. Bhat, G. K. Miller and S. S. Rao, Statistical Analysis of Queueing Systems, in: *Frontiers in Queueing*, ed. Jewgeni H. Dshalalow, CRC Press, 1997, chapter 13.

[10] B. P. Carlin and T. A. Louis, *Bayes and empirical Bayes methods for data analysis*, Chapman and Hall/CRC Press, London, 2000.

[11] G. Casella and E. I. George, Explaining the Gibbs sampler, The American Statistician, 46(3) (1992), 167–174.

[12] M. L. Chaudhry and J. G. C. Templeton, *A first Course in Bulk Queues*, Wiley, New York, 1983.

[13] D. Conesa, Inferencia y predicción en colas con ingresos o llegadas en grupos, PHDThesis, Universitat de València, 2000 (in spanish).

[14] D. Draper, *Bayesian Hierarchical Modelling*, Preliminary draft avalaible in the url address `http://www.bath.ac.uk/~masdd`", 2001.

[15] G. García-Donato, Control de procesos. Alternativas Empírico-Bayes y Bayesianas a los métodos clásicos, Master Degree work, Universitat de València, 2001 (in spanish).

[16] A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin, *Bayesian Data Analysis*, Chapman and Hall, 1995.

[17] W. R. Gilks, S. Richardson and D. J. Spiegelhalter, *Markov Chain Monte Carlo in practice*, Chapman and Hall, 1996.

[18] W. R. Gilks, A. Thomas and D. J. Spiegelhalter, A language and program for complex Bayesian modelling, The Statistician, 43 (1994), 169–178.

[19] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, Academic Press, Inc., Boston.

[20] J. D. Griffiths, Queueing at the Suez Canal, J. Opl. Res. Soc. 46 (1995) 1299–1309.

[21] S. Lauritzen, *Graphical models*, Oxford University Press, 1996.

[22] J. Lehoczky, Statistical Methods, in: *Handbooks in O.R. and M.S.*, eds. Heyman, D. P. and Sobel, M. J., Elsevier Science Publishers B. V. (North Holland), 1990, chapter 6.

[23] J. S. Maritz and T. Lwin, *Empirical Bayes methods*, Chapman and Hall, 1989.

[24] J. Medhi, *Stochastic Models in Queueing Theory*, Academic Press, Inc., Boston, 1991.

[25] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach*, The John Hopkins University Press, Baltimore, 1981.

[26] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Numerical recipes*, Cambridge University Press, 1986.

[27] D. Rios Insua, M. Wiper and F. Ruggieri, Bayesian Analysis of $M/E_r/1$ and $M/H_k/1$ queues. Queueing Systems, 30 (1998), 289–308.

[28] H. Robbins, Some thoughts on empirical Bayes estimation, Ann. Stat. (1983) 713–723.

[29] S. Young Sohn, Empirical Bayesian analysis for traffic intensity: $M/M/1$ queues with covariates, Queueing Systems 22 (1996) 383–401.

[30] D. J. Spiegelhalter, A. Thomas and N. G. Best, Winbugs version 1.2 user manual, MRC Biostatistics Unit, 1999.

[31] D. Thiruvaiyaru and I. V. Basawa, Empirical Bayes Estimation for queueing systems and Networks, Queueing Systems 11 (1992) 179–202.

[32] H. C. Tijms, Computational methods in queueing, in: *Frontiers in Queueing*, ed. Jewgeni H. Dshalalow, CRC Press, 1997, chapter 12.

[33] M. Wiper, Bayesian analysis of $E_r/M/1$ and $E_r/M/c$ queues, J. Stat. Planning and Inference, 69 (1998), 65–79.