

Bayesian Forecasting of Parts Demand

Phillip M. Yelland

Sun Microsystems Laboratories, 16 Network Circle, Menlo Park, CA 94025, U.S.A.

Abstract

As supply chains for high technology products increase in complexity, and as the performance expected of those supply chains also increases, forecasts of parts demand have become indispensable to effective operations management in these markets. Unfortunately, rapid technological change and an abundance of product configurations mean that demand for parts in high-tech is frequently volatile and hard to forecast. The paper describes a Bayesian statistical model developed to forecast parts demand for Sun Microsystems, Inc., a major vendor of network computer products. The model embodies a parametric description of the part life-cycle, allowing it to anticipate changes in demand over time. Furthermore, using hierarchical priors, the model is able to pool demand patterns for a collection of parts, smoothing out idiosyncratic variation and furnishing calibrated forecasts for new parts with little or no demand history. The paper discusses the problem addressed by the model, the model itself and a procedure for calibrating it, and compares its forecast performance with that of “off-the-shelf” alternatives.

Key words: Bayesian methods, demand forecasting, forecasting practice, state space models, supply chain

1 Introduction

1.1 Background

Manufacturing modern high technology products like computers is an exacting business: Competition is intense, product lifecycles fleeting, components are frequently expensive and prone to rapid obsolescence, and supply chains often span the globe. As Lapede (2006) observes, demand forecasts have become increasingly central to supply chain management for participants in these markets. This paper focuses on one such market participant—Sun Microsystems Inc., a vendor of enterprise computing products—and on forecasts of the demand for the manufacturing parts used in its supply chain.

Email address: phillip.yelland@sun.com (Phillip M. Yelland).

BAYESIAN FORECASTING OF PARTS DEMAND

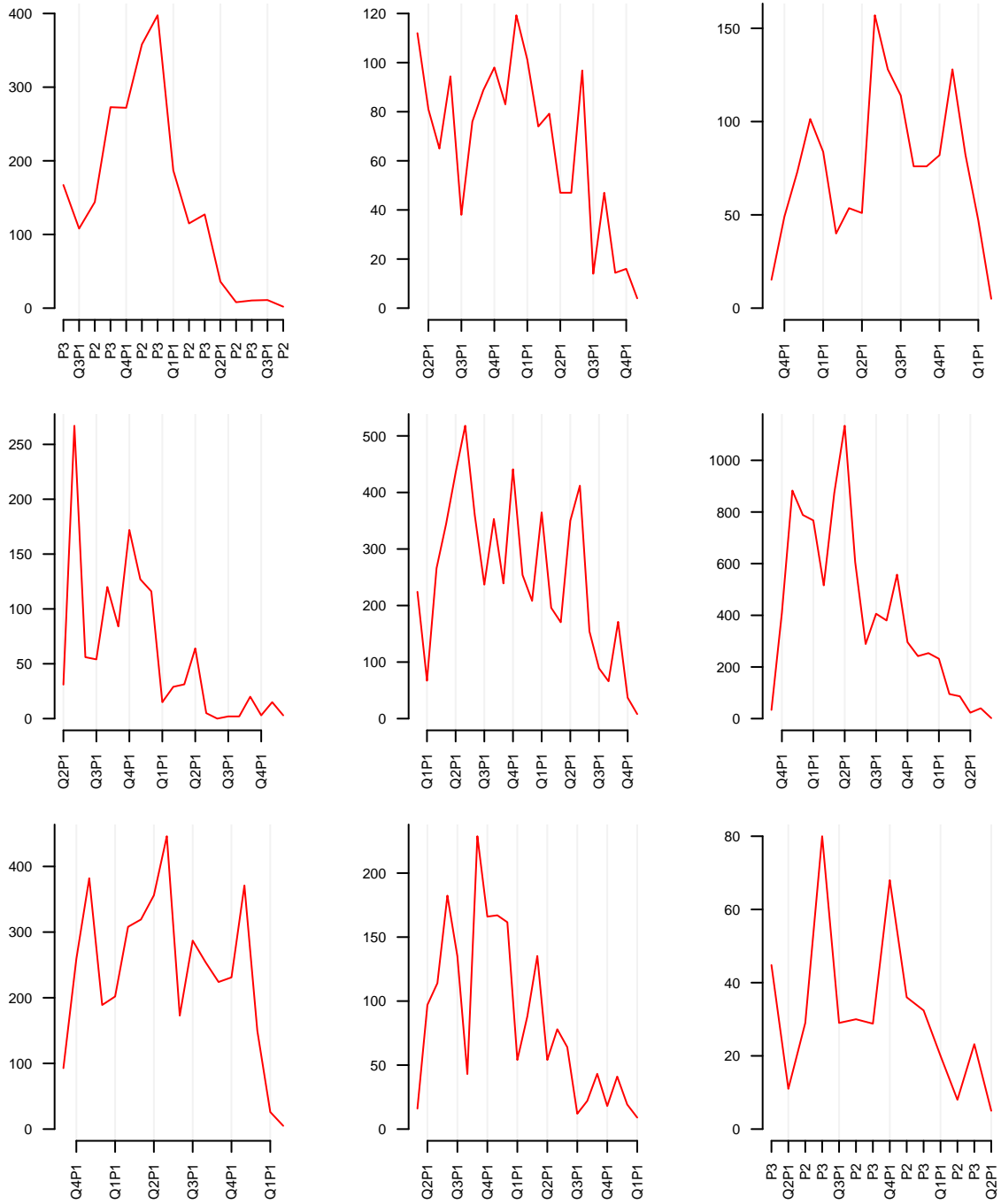


Figure 1. Sample part demands (units)

BAYESIAN FORECASTING OF PARTS DEMAND

	Minimum	1 st Quartile	Median	Mean	3 rd Quartile	Maximum
Total units	474	1,058	1,951	2,817	3,958	8,916
Total periods	13.00	17.00	21.00	20.78	24.00	31.00
Mean units/period	23.79	62.15	110.66	136.94	210.08	405.28
Coef. of variation	0.49	0.68	0.87	0.85	0.97	1.27

Figure 2. Sampled part demand series—summary statistics

As might be expected for a diversified IT vendor with approximately \$14 billion in annual sales,¹ Sun’s manufacturing operations consume a multitude of parts, ranging from semi-populated computer chassis and disk-drive assemblies to power cables and case fasteners. In fact tens of thousands of different parts are used in Sun’s product lines, and around 1,000 of these (the exact number fluctuates constantly) are sufficiently expensive or subject to sufficiently long lead times that they require forecast updates at least every month. Prior to the work described in this paper, the company produced forecasts for these latter parts using a heuristic calculation—detailed later in the paper—based on the product-level sales forecasts produced by its sales and marketing organizations (Yelland 2004). The study described in this paper was initiated with the aim of establishing whether such forecasts might be improved using statistical techniques.

Figure 1 displays the lifetime demands for a selection of the manufacturing parts at issue.² Here demand is represented by unit shipments in a *planning period*, which are of roughly on month’s duration. For the purposes of supply chain planning, the company’s fiscal quarters are divided into three such planning periods, of four, four and five weeks respectively; unit shipments in the figure (and throughout this paper) have been normalized *pro rata* to 20 trading days in each period. The parts in the figure are excerpted from a larger sample of some 45 parts that was used to guide the development of the model described in this paper. Characteristics of the series in this sample are summarized in Table 2. In selecting parts for study, we were careful to ensure that: (1) The per-unit cost of a part (which ranged into thousands of dollars for those in the study) was sufficient to justify the effort involved in developing the model and running it on a routine basis, and (2) Unit shipments per planning period were generally high enough that issues such as demand intermittency (Boylan 2005) and discreteness (McCabe and Martin 2005; Yelland 2008) might reasonably be put aside.

¹ \$13.88 billion in the financial year ending June 2008.

² To preserve commercial confidentiality, data in this and other figures in the paper have been mildly disguised.

These selection criteria notwithstanding, the Figure illustrates the challenge involved in producing forecasts for manufacturing parts at Sun. Demand for parts is normally much more variable than that of the products they comprise, since part demand depends on customer configuration choices, supplier sourcing arrangements, technological changes, the availability of substitutes, and so on. (This variability is roughly quantified by the coefficients of variation in Table 2, which are calculated by taking the ratio of the standard deviation of demand in a series to the series mean.) Technological development generally means that part lifecycles are also appreciably shorter than products', and the 14 to 24-period lifecycles in the figure are quite typical.

1.2 Prior Work

The literature exploring operations management in the face of uncertain demand is vast, and dates back to the 1950's and beyond (Arrow, Karlin, and Scarf 1958). By and large however, researchers have concentrated on investigating policy responses to demand that is generated by a known stochastic process. For example, early works such as (Veinott 1965) assume that demand conforms to a known set of independent probability distributions, while more recent exercises assume a known member of a class of processes, such as ARIMA (Gilbert 2005) or linear/Gaussian state-space (Aviv 2003). Specific prescriptions for using formal quantitative methods to forecast demand for supply chain management are far less numerous. (Harvey and Snyder 1990; Snyder, Koehler, and Ord 2002; Gardner 1990) are amongst the few examples describing generally applicable methods, while Boylan (2005) and Shenstone and Hyndman (2005) concentrate on specialized forecasting techniques applicable to items such as spare parts with intermittent demands. Documented instances of statistical forecasting models actually employed in commercial supply chain management are rarer still; while commercially-available software packages for supply chain management include forecasting techniques such as ARIMA, linear regression and spectral-decomposition-based smoothing (Yurkiewicz 2006), both Sanders and Manrodt (1994) and Lapide (2006) note the continued heavy reliance on judgmental and ad-hoc methods in most companies. Our own experience at Sun attested to the difficulties involved in forecasting for operations using statistical models—we found that such approaches were frequently stymied by little or no historical data (which rendered techniques such as ARIMA and spectral smoothing inapplicable) or lack of suitable predictors (ruling out linear regression, for example).

The quantitative forecasting technique that perhaps most appeals to both academics and practitioners is one of the most venerable: *exponential smoothing*. For practical applications, exponential smoothing is technically straightforward to implement, requires little or no historical data for calibration, and no predictors or regressors. Armstrong and Green (2005) aver that “exponential smoothing is the most popular and cost effective of the statistical extrapolation methods”, and Gardner (1990) documents the use of exponential smoothing in operations management. On the theoretical side, as summarized by

Gardner (1985, 2006), the technique has enjoyed almost 40 years of continuous development since its invention by Brown (1959), and has recently engendered the development of a new class of structural time series models (Hyndman, Koehler, Ord, and Snyder 2008).

Though exponential smoothing appeared a natural choice as a generic approach to the parts forecasting problem, experience with the product-level sales model documented in (Yelland 2004), suggested that a specially-designed statistical model developed might well yield superior results, and also pointed to the effectiveness of Bayesian techniques for extrapolating short time series like part demand histories. Unfortunately, the model in (Yelland 2004) relies on informative Bayesian priors deduced from judgmental forecasts for product sales, and these priors are an unreliable guide to demand for parts within those products, for the reasons described in the previous section. On the other hand, the use of diffuse or non-informative priors is ruled out by the necessity of producing forecasts early in the life of a part, before sufficient observations accrue to produce proper forecast distributions. Therefore, though the model presented here is also Bayesian, it uses a *hierarchical* prior (Gelman and Hill 2006) to produce initial parameter estimates from sales records of established parts. This sort of “forecasting by analogy” echoes the work of Duncan, Gorr, and Szczypula (2001), who observe that Bayesian pooling also helps deal with time-series volatility. Früwirth-Schatter and Kaufmann (2008) use hierarchical Bayesian priors for time-series analysis, too, though their focus is on clustering, rather than forecasting.

2 A Model for Parts Demand at Sun

This section describes the Bayesian model developed for the parts forecasting problem. For reference purposes, notation, quantities and probability distributions used throughout the paper are documented in Section A and Section B. A summary exposition of the model is given in Figure 3, and the hierarchical structure of the model is illustrated by Figure 4, which captures its dependencies in the form of a directed graph of the sort described by Rossi, Allenby, and McCulloch (2005, pp. 67–81), for example.

2.1 Lifecycle Curve

Motivation

Like the forecasting model in (Yelland 2004), the model in this paper is based on a stylized representation of a product’s life-cycle demand. The representation used in this model is derived from the *Weibull distribution*, following the work of Moe and Fader (2002), who in turn draw on the body of research into new product diffusion modeling summarized by Mahajan, Muller, and Wind (2000). According to this representation, the total number of units of a part i demanded by period t after its initial availability is determined by (a

Period sales

$$y_{it} \sim \mathbf{N}(\gamma_i[\Delta\mathbb{W}(t|\alpha_i, \delta_i) + x_{it}], [\gamma_i(1 + 3\zeta_{it})\sigma_y]^2), \quad p(\sigma_y) \propto \mathcal{I}(\sigma_y > 0)$$

Outliers

$$\zeta_{it} \sim \text{Bern}(0.05)$$

Total Demand

$$\begin{aligned} \gamma_i &\sim \mathbf{N}_{[0, \infty)}(g_{\text{prod}(i)}, \sigma_\gamma^2), \quad p(\sigma_\gamma) \propto \mathcal{I}(\sigma_\gamma > 0) \\ g_j &\sim \mathbf{N}(\mu_g, \sigma_g^2), \quad p(\mu_g) \propto 1, \quad p(\sigma_g) \propto \mathcal{I}(\sigma_g > 0) \\ s_i &\sim \mathbf{N}(\gamma_i, [0.2\gamma_i]^2) \end{aligned}$$

Life-cycle parameters

$$\begin{aligned} \alpha_i &\sim \mathbf{N}_{[0, \infty)}(a_{\text{prod}(i)}, \sigma_\alpha^2), \quad p(\sigma_\alpha) \propto \mathcal{I}(\sigma_\alpha > 0) \\ \delta_i &\sim \mathbf{N}_{[0, \infty)}(d_{\text{prod}(i)}, \sigma_\delta^2), \quad p(\sigma_\delta) \propto \mathcal{I}(\sigma_\delta > 0) \\ a_j &\sim \mathbf{N}(\mu_a, \sigma_a^2), \quad p(\mu_a) \propto 1, \quad p(\sigma_a) \propto \mathcal{I}(\sigma_a > 0) \\ d_j &\sim \mathbf{N}(\mu_d, \sigma_d^2), \quad p(\mu_d) \propto 1, \quad p(\sigma_d) \propto \mathcal{I}(\sigma_d > 0) \end{aligned}$$

Latent autoregression

$$\begin{aligned} x_{it} &\sim \mathbf{N}(\lambda_{i1}x_{i,t-1} + \lambda_{i2}x_{i,t-2}, \sigma_x^2), \quad \sigma_x = 0.8\sigma_y \\ (\lambda_{i1}, \lambda_{i2})^\top &\sim \mathbf{N}(\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda), \quad p(\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda) \propto |\boldsymbol{\Sigma}_\lambda|^2 \\ (x_{i0}, x_{i,-1})^\top &\sim \mathbf{N}(\boldsymbol{\mu}_{x_0}, \boldsymbol{\Sigma}_{x_0}), \quad \boldsymbol{\mu}_{x_0} = (0, 0)^\top, \quad \boldsymbol{\Sigma}_{x_0} = \text{diag}(2, 2) \end{aligned}$$

Figure 3. Model summary

multiple of) the cumulative distribution function (CDF) of a Weibull distribution:

$$\mathbb{W}(t|\lambda_i, k_i) = 1 - e^{-(t/k_i)^{\lambda_i}}.$$

The (strictly positive) parameters λ_i and k_i are respectively referred to as the *shape* and *scale* parameters of the distribution. Since demand is only observed over planning periods, Moe and Fader follow the germinal work of the Schmittlein and Mahajan (1982) in equating the expected number of units shipped during a period to a multiple of the difference in the value of the CDF over that period. Using the (backward) difference operator Δ (Enders 1995, p. 7, for example), such that $\Delta\mathbb{W}(t|\lambda_i, k_i) = \mathbb{W}(t|\lambda_i, k_i) - \mathbb{W}(t-1|\lambda_i, k_i)$:

$$\mathbb{E}(y_{it}) = \gamma_i \Delta\mathbb{W}(t|\lambda_i, k_i) = \gamma_i [e^{-((t-1)/k_i)^{\lambda_i}} - e^{-(t/k_i)^{\lambda_i}}]. \quad (1)$$

○ — Unobserved random variable ● — Observed random variable □ — Calculated quantity

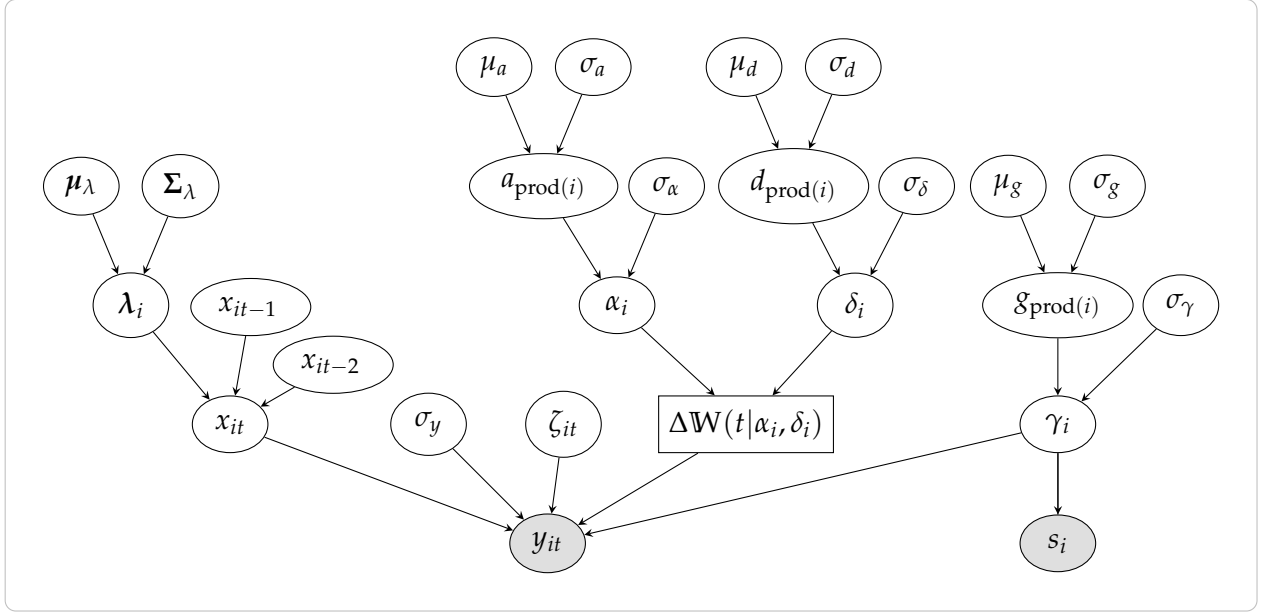


Figure 4. Model hierarchy

Here the multiplier γ_i may be interpreted as a notional “lifetime demand” for part i . Actual period demands are assumed to follow a multinomial distribution with probabilities determined by the first differences in the CDF:³

$$(y_{i1} \dots y_{iT_i})^\top \sim \text{Mult}(\gamma_i; \pi_{i1}, \dots, \pi_{iT_i}),$$

where $\pi_{it} = \Delta W(t|\lambda_i, k_i)$, for $t = 1, \dots, T_i$. (2)

Moe and Fader’s use of a Weibull curve to describe the adoption of a new product has theoretical appeal, given the Weibull distribution’s origins in the analysis of events that occur after a period of random duration. The use of the Weibull in this paper is rather more pragmatic—as Moe and Fader observe, it offers an appealing combination of parsimony and flexibility.⁴ An informal measure of how well the Weibull model captures the trend

³ Moe and Fader actually use a “rescaled” version of the multinomial formulation presented here, so as better to accommodate the right truncation of their data.

⁴ It could be said that our outlook conforms with the *technological* approach to modeling of Bernardo and Smith (1994, p.238), in that we are concerned less with the “‘true’ mechanisms of the phenomenon under study ... [than] simply with providing a reliable basis for practical action in predicting ... the phenomena of interest”. In fact an *ab initio* argument for the Weibull model might be made by postulating some form of “adoption” process for the parts themselves, along the lines of that presented by Norton and Bass (1987), for example. However, without detailed

BAYESIAN FORECASTING OF PARTS DEMAND

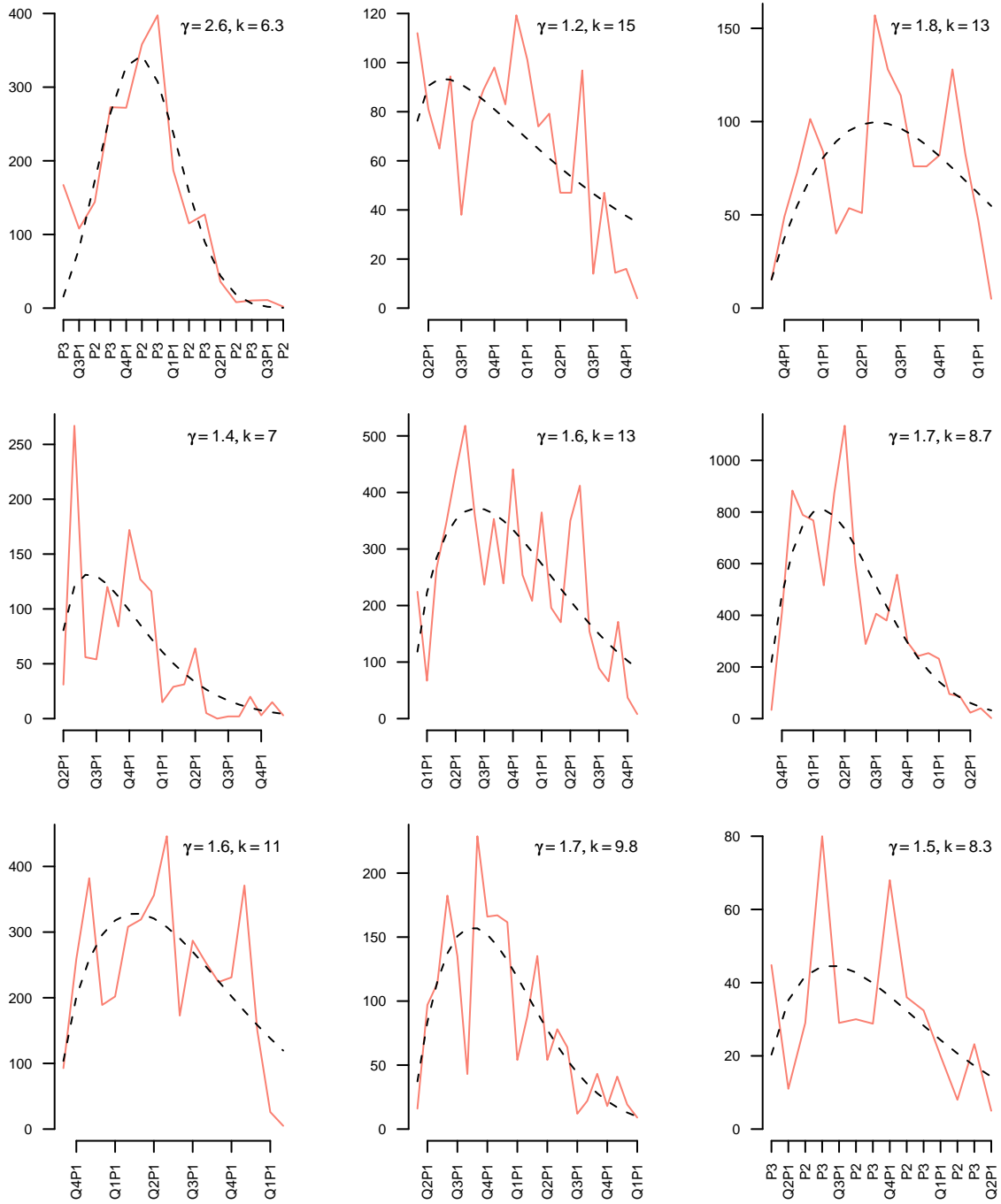


Figure 5. Weibull curve fits

in the Sun’s part demands with only two parameters can be gauged from Figure 5, where the series in Figure 1 have been fitted to (first differences in) Weibull curves.⁵

Alternate Parameterization

In fact as an aid to interpretability, and to help the convergence of the Gibbs sampler (see below), we actually use an alternate parameterization of the Weibull in the model, indexed by α_i and δ_i , which are respectively the 20th percentile of the distribution and the difference between its 95th and 20th percentiles.⁶ Informally, we would therefore expect 20 percent of the total demand for part i to occur by time α_i , and 95 percent by $\alpha_i + \delta_i$. Restating equation (1):

$$E(y_{it}) = \gamma_i \Delta W(t|\alpha_i, \delta_i). \tag{3}$$

The correspondence between the two parameterizations is given by:

$$\lambda_i = \frac{2.6}{\log(\alpha_i + \delta_i) - \log(\alpha_i)}, \quad k_i = \frac{\alpha_i}{\sqrt[2.6]{0.22}}.$$

Observed Lifetime Demands

Unfortunately, in practice, the likelihood induced by equation (3) determines γ_i with insufficient precision to produce good forecasts.⁷ Thus we found it necessary to constrain estimates of γ_i by the additional assertion that for superannuated parts whose entire life cycles have been observed, the actual observed total lifetime demand, s_i , must approximate γ_i (the difference arising from noisy demand observations):

$$s_i \sim N(\gamma_i, [0.2\gamma_i]^2) \tag{4}$$

This helps constrain the values of γ_i for current parts by virtue of the hierarchical prior for γ_i (discussed next). For current parts—whose life cycles have not completed—is deemed “missing”, and γ is unconstrained by (4).

information about end-user behavior to support it (very difficult to obtain in this context), such a construction would amount to little more than “armchair theorizing”.

⁵ The curves in the figure were actually fitted using the modification to distributional assumption (2) set out in equation (7), since the latter better accommodates our data.

⁶ We use the difference between the 20th and 95th percentiles rather than the 95th percentile itself because α_i and δ_i might reasonably be considered *a priori* independent, whereas two percentiles would most assuredly not be.

⁷ Roughly speaking, with only a few demands actually observed for part i , even with prior constraints on α_i and δ_i , equation (3) gives support to very large values of γ_i .

Parameter Priors

Parameters α , δ and γ are drawn from hierarchical priors that group part parameters by the products they constitute, and group the hyperparameters of all products into common population priors. Treating the case of α in detail: For part i , α_i —which is necessarily positive—is drawn from a normal distribution, truncated on the left at 0. Scale parameter σ_α of this truncated normal distribution is common to all parts, but the location parameter, $a_{\text{prod}(i)}$, is shared only with other parts for the same product. At the next level of the hierarchy, for all products j , a_j is drawn from a normal distribution with mean and variance common to all products. The mean μ_a of this latter distribution has a non-informative prior. The other parameters in equation (3) (viz. δ_i and γ_i) are specified similarly, using truncated normals at the part level and normal distributions for the product-specific location parameters. In symbols, continuing with α as an example:

$$\alpha_i \sim N_{[0,\infty)}(a_{\text{prod}(i)}, \sigma_\alpha^2), \quad a_j \sim N(\mu_a, \sigma_a^2), \quad p(\mu_a) \propto 1 \quad (5)$$

Variance Priors

We use non-informative priors for the variances of the distributions in (5), as we do for scale parameters of normal and truncated normal distributions throughout the model. Gelman’s (2006) paper demonstrates that producing a truly “non-informative” prior for variance parameters is a delicate business, particularly in hierarchical models such as this (where, for example, the popular “weakly informative” inverse-gamma prior of Spiegelhalter, Thomas, Best, Gilks, and Lunn (2003) can lead to degenerate posterior distributions for variances of group parameters). Here, we use a uniform density on the positive half-line as the prior for the standard deviations in (5), which as Gelman (2006) indicates is formally equivalent to an inverse- χ^2 density with -1 degrees of freedom for the corresponding variances. To complete the specification in (5), therefore:

$$p(\sigma_\alpha) \propto \mathcal{I}(\sigma_\alpha > 0), \quad p(\sigma_a) \propto \mathcal{I}(\sigma_a > 0) \quad (6)$$

2.2 Error Structure

Srinivasan and Mason (1986) point out that in many applications, distributional assumption (2) is likely to understate substantially the observed dispersion of sales, and they advocate the use of a nonlinear least squares model instead. For y_{it} , such a model may be produced by introducing an additive error term⁸ into equation (3):

$$y_{it} = \gamma_i \Delta W(t|\alpha_i, \delta_i) + \epsilon_{it}. \quad (7)$$

⁸ Srinivasan and Mason (1986) aver that an additive error term produces forecasts that are more accurate than those produced with a multiplicative error, which is borne out by our own experience.

We also follow Srinivasan and Mason in assuming that the error term ϵ_{it} is normally-distributed. This is a substantial technical convenience, but strictly speaking, it gives rise to forecast distributions that lack *coherence* in the sense of McCabe and Martin (2005), since they give support to negative demand values. In practice, however, we have found that point forecasts—the focus of interest for Sun’s supply chain—produced using the mean of the forecast distributions from the model (see section 4) are invariably positive.

Scaled Error Variance

If we assume that ϵ_{it} is normally-distributed with zero mean and standard deviation ζ_i , equation (7) may be re-expressed:

$$y_{it} \sim N(\gamma_i \Delta W(t|\alpha_i, \delta_i), \zeta_i^2) \quad (8)$$

One of the objectives of the model is to forecast sales of new parts, relying only on demand histories for parts that preceded them. In order to do this, we assume that the part-specific standard deviation ζ_i may be obtained by scaling a common standard deviation σ_y by γ_i , the same factor used to scale the mean of the distribution in (8). Thus:

$$y_{it} \sim N(\gamma_i \Delta W(t|\alpha_i, \delta_i), [\gamma_i \sigma_y]^2) \quad (9)$$

Outliers

Inspection of the demand histories displayed in Figure 1 suggests that demand is subject to the occasional “spike”, and we have found it politic to accommodate these spikes as outliers in the model. This can be achieved expediently using a *binary selection model*, as described by Congdon (2003, sec. 3.6.1). Using this construction, each observation of demand y_{it} is associated with a latent binary variable $\zeta_{it} \in \{0, 1\}$, which identifies y_{it} as an outlier iff $\zeta_{it} = 1$. Observations identified as outliers are considered to be drawn from a distribution with the same mean as that in (9) but a standard deviation four times as large.⁹ This results in:

$$y_{it} \sim N(\gamma_i \Delta W(t|\alpha_i, \delta_i), [\gamma_i(1 + 3\zeta_{it})\sigma_y]^2) \quad (10)$$

Since the occurrence of an outlier is *ipso facto* a rare event, the prior for the indicator ζ_{it} is a Bernoulli distribution such that the probability that $\zeta_{it} = 1$ is 5%.

⁹ A scale inflation factor of 3 or 4 is suggested for general use in characterizing discrepant observations by West and Harrison (1997, p. 400).

2.3 Autoregression

Though the model for demand in (10) captures the dynamics of part demand fairly well, short-term forecasts are improved by the addition of a *latent autoregressive process* that West and Harrison (1997, p. 300) suggest as a “catch-all noise model used as a purely empirical representation of residual variation”. The value of this process for part i in period t is denoted x_{it} ; though specific to one particular part, x_{it} is also scaled by γ_i , allowing priors for the process to be drawn from common distributions. Thus (10) is modified:

$$y_{it} \sim \mathbf{N}(\gamma_i[\Delta\mathbf{W}(t|\alpha_i, \delta_i) + x_{it}], [\gamma_i(1 + 3\zeta_{it})\sigma_y]^2) \quad (11)$$

where:

$$x_{it} \sim \mathbf{N}(\lambda_{i1}x_{i,t-1} + \lambda_{i2}x_{i,t-2}, \sigma_x^2) \quad (12)$$

The autoregression coefficients are drawn from a common multivariate normal distribution that is itself specified using the noninformative *multivariate Jeffrey’s prior* (Sun and Berger 2006), so that:

$$(\lambda_{i1}, \lambda_{i2})^\top \sim \mathbf{N}(\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda), \quad p(\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda) \propto |\boldsymbol{\Sigma}_\lambda|^{-2}. \quad (13)$$

To prevent overfitting, we found it necessary to provide informative priors for starting values x_{i0} and $x_{i,-1}$ of the autoregression; it seemed reasonable *a priori* to assert that there was little “residual variation” (to use West and Harrison’s terminology) at the outset of a part’s life, and so the starting priors were concentrated around zero:

$$(x_{i0}, x_{i,-1})^\top \sim \mathbf{N}(\boldsymbol{\mu}_{x_0}, \boldsymbol{\Sigma}_{x_0}), \quad \boldsymbol{\mu}_{x_0} = (0, 0)^\top, \quad \boldsymbol{\Sigma}_{x_0} = \text{diag}(2, 2) \quad (14)$$

As a further bulwark against overfitting in the model, the variance σ_x^2 in (12) is fixed at a constant multiple of the conditional variance, σ_y^2 , of y_{it} . This multiple—which by analogy with the “signal to noise ratio” discussed by e.g. West and Harrison (1997, p. 40), apportions the unexplained variation in the model between the transient and autoregressive components—constitutes a tuning parameter of the model; we found that setting $\sigma_x^2 = (0.8 \sigma_y)^2$ yielded good results.

3 Estimation

This section discusses the procedure used to perform Bayesian updating of the model, and summarizes the results of updating with the full sample described in Section 1.

3.1 Gibbs Sampler

Updating is carried out using a Gibbs sampling routine, schematic descriptions of which now abound in the literature—see (Gilks, Richardson, and Spiegelhalter 1996, chp. 1), for example. The individual steps of this particular sampler are described below. Many of the steps rely on standard results concerning conjugate updating in Bayesian analysis, which may be found in reference texts such as (Gelman, Carlin, Stern, and Rubin 2003) or (Bernardo and Smith 1994). Where such closed-form updates are not available, we resort to Metropolis-Hastings sampling (also discussed by Gilks et al.); proposals are generated using Geweke and Tanizaki’s (2003) *Taylored chain* procedure, details of which are provided in Section C.

In the following, each step is introduced by the conditional distribution from which a sample is to be drawn. Variables of which the sampled quantity is conditionally independent are omitted from the conditioning set. We use the abbreviations: $\zeta_i = (\zeta_{i1}, \dots, \zeta_{iT_i})$, $\mathbf{x}_i = (x_{i1}, \dots, x_{iT_i})$ and $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i})$. Also, for each product j , let $\text{parts}(j) = \{i \mid \text{prod}(i) = j, \text{ for } i = 1, \dots, N\}$ be the set of associated parts. In the interests of brevity, we specify draws for a_j , σ_α , μ_a and σ_a only; samples for d_j , σ_δ , μ_d and σ_d , and g_j , σ_γ , μ_g and σ_g are generated in a parallel fashion.

$$\gamma_i \mid \mathbf{y}_i, \sigma_y, \alpha_i, \delta_i, \mathbf{x}_i, \zeta_i, g_{\text{prod}(i)}, \sigma_\gamma, s_i$$

The kernel of the full conditional distribution is given by the expression:

$$\left[\prod_{t=1}^{T_i} \text{N}(y_{it} \mid \gamma_i (\Delta W(t \mid \alpha_i, \delta_i) + x_{it}), [\gamma_i (1 + 3\zeta_{it}) \sigma_y]^2) \right] \times \text{N}_{[0, \infty)}(\gamma_i \mid g_{\text{prod}(i)}, \sigma_\gamma^2) \times \varphi_i,$$

where:

$$\varphi_i = \begin{cases} \text{N}(s_i \mid \gamma_i, [0.2\gamma_i]^2) & \text{if the entire lifecycle of part } i \text{ has been observed} \\ 1 & \text{otherwise.} \end{cases}$$

In either case, sampling is carried out using the Geweke-Tanizaki Taylored chain to generate a proposal in a Metropolis-Hastings step.

$$\alpha_i \mid \mathbf{y}_i, \gamma_i, \sigma_y, \delta_i, \mathbf{x}_i, \zeta_i, a_{\text{prod}(i)}, \sigma_\alpha$$

The full conditional is proportional to the expression:

$$\left[\prod_{t=1}^{T_i} \text{N}(y_{it} \mid \gamma_i (\Delta W(t \mid \alpha_i, \delta_i) + x_{it}), [\gamma_i (1 + 3\zeta_{it}) \sigma_y]^2) \right] \times \text{N}_{[0, \infty)}(\alpha_i \mid a_{\text{prod}(i)}, \sigma_\alpha^2).$$

This is sampled using the Taylored chain proposal in a Metropolis-Hastings step.

$$\delta_i \mid \mathbf{y}_i, \gamma_i, \sigma_y, \alpha_i, \mathbf{x}_i, \zeta_i, d_{\text{prod}(i)}, \sigma_\delta$$

As above, but this time sampling from:

$$\left[\prod_{t=1}^{T_i} \mathbf{N}(y_{it} \mid \gamma_i(\Delta \mathbf{W}(t \mid \alpha_i, \delta_i) + x_{it}), [\gamma_i(1 + 3\zeta_{it})\sigma_y]^2) \right] \times \mathbf{N}_{[0, \infty)}(\delta_i \mid d_{\text{prod}(i)}, \sigma_\delta^2).$$

$$\zeta_{it} \mid y_{it}, \gamma_i, \sigma_y, \alpha_i, \delta_i, x_{it}$$

Let:

$$\begin{aligned} \hat{y}_{it} &= \gamma_i(\Delta \mathbf{W}(t \mid \alpha_i, \delta_i) + x_{it}), \\ p_0 &= (1 - 0.05) \times \mathbf{N}(y_{it} \mid \hat{y}_{it}, [\gamma_i \sigma_y]^2), \\ p_1 &= 0.05 \times \mathbf{N}(y_{it} \mid \hat{y}_{it}, [4\gamma_i \sigma_y]^2). \end{aligned}$$

Sample ζ_{it} from the Bernoulli distribution with success probability $p_1 / (p_0 + p_1)$.

$$\mathbf{x}_i \mid \mathbf{y}_i, \gamma_i, \sigma_y, \alpha_i, \delta_i, \zeta_i, \boldsymbol{\mu}_{x_0}, \boldsymbol{\Sigma}_{x_0}$$

Following West and Harrison (1997, example 9.6), \mathbf{x}_i is given by the state vector of the *dynamic linear model*, or *DLM*, specified by $(\mathbf{F}_t, \mathbf{G}_t, V_t, \mathbf{W}_t)$, for $t \in 1, \dots, T_i$, where:

$$\mathbf{F}_t = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{G}_t = \begin{pmatrix} \lambda_{i1} & \lambda_{i2} \\ 1 & 0 \end{pmatrix}, \quad V_t = [(1 + 3\zeta_{it})\sigma_y]^2, \quad \mathbf{W}_t = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & 0 \end{pmatrix}.$$

First and second moments of the multivariate normal prior for the initial state configuration (\mathbf{m}_0 and \mathbf{C}_0 in West and Harrison's formulation) are $\boldsymbol{\mu}_{x_0}$ and $\boldsymbol{\Sigma}_{x_0}$, respectively.

Procedures for sampling the state of such a DLM (generally under the moniker *forward filtering/backwards sampling* algorithms) are described by Frühwirth-Schnatter (1994), West and Harrison (1997) and Durbin and Koopman (2001), amongst others.

$$\lambda_{i1}, \lambda_{i2} \mid \mathbf{x}_i, \sigma_x, \boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda$$

Draw from the posterior distribution of the coefficients in the linear regression $x_{it} \sim \mathbf{N}(\lambda_{i1}x_{t-1t} + \lambda_{i2}x_{t-2t}, \sigma_x^2)$, given a prior $(\lambda_{i1}, \lambda_{i2})^\top \sim \mathbf{N}(\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda)$ —see e.g. (Gelman et al. 2003, chp. 8).

$$\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda \mid \lambda_1, \dots, \lambda_N$$

Conjugate updating for parameters of a multivariate normal distribution.

$$a_j \mid \{\alpha_k \mid k \in \text{parts}(j)\}, \sigma_\alpha$$

Sampling is carried out using a device due to Griffiths (2004): Specifically, for $k \in$

parts(j), let:

$$\tilde{\alpha}_k = a_j + \sigma_\alpha \Phi^{-1} \left[\frac{\Phi\left(\frac{\alpha_i - a_j}{\sigma_\alpha}\right) - \Phi\left(\frac{-a_j}{\sigma_\alpha}\right)}{1 - \Phi\left(\frac{-a_j}{\sigma_\alpha}\right)} \right], \quad (15)$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

Then as Griffiths demonstrates, supposing that $\tilde{\alpha}_k \sim N(a_j, \sigma_\alpha^2)$ and drawing from the conditional distribution $a_j \mid \{\tilde{\alpha}_k \mid k \in \text{parts}(j)\}, \sigma_\alpha$ (a straightforward application of semi-conjugate updating) is equivalent to drawing from $a_j \mid \{\alpha_k \mid k \in \text{parts}(j)\}, \sigma_\alpha$ given that $\alpha_k \sim N_{[0, \infty)}(a_j, \sigma_\alpha^2)$.

$$\sigma_\alpha \mid \alpha_1, \dots, \alpha_N, a_1, \dots, a_J$$

Again, using Griffiths's device, draw from $\sigma_\alpha \mid \tilde{\alpha}_1, \dots, \tilde{\alpha}_N$, given that $\tilde{\alpha}_i - a_{\text{prod}(i)} \sim N(0, \sigma_\alpha^2)$, where $\tilde{\alpha}_i$ is defined in equation (15).

$$\mu_a, \sigma_a \mid a_1, \dots, a_J$$

Two step semi-conjugate updating for parameters of a normal distribution, drawing μ_a first from $\mu_a \mid a_1, \dots, a_J, \sigma_a$, and then σ_a from $\sigma_a \mid a_1, \dots, a_J, \mu_a$.

$$\sigma_y \mid \mathbf{y}_1, \alpha_1, \delta_1, \gamma_1, \mathbf{x}_1, \zeta_1, \dots, \mathbf{y}_N, \alpha_N, \delta_N, \gamma_N, \mathbf{x}_N, \zeta_N$$

Let $r_{it} = (1 + 3\zeta_{it})^{-1}(y_{it}/\gamma_i - \Delta W(t|\alpha_i, \delta_i) - x_{it})$, for $i = 1, \dots, N, t = 1, \dots, T_i$. Then $r_{it} \sim N(0, \sigma_y^2)$, and conjugate updating applies.

3.2 Posterior Estimates and Diagnostics

Figure 6 (inspired by similar figures on e.g. p. 351 of Gelman and Hill 2006) illustrates the results of updating the model with the full sample of demand histories described in Section 1.1. Here the Gibbs sampler was run in a single chain for 4,000 iterations, with samples from the first 1,200 discarded; no thinning of the remaining samples was performed.

On the left of the Figure are displayed posterior distributions for the quantities associated with a random selection of parts ("pt."- 8, 10, 29 and 35) and randomly selected products ("pd."-, 3, 4, 5 and 8), as well as population-level parameters. Parameters α_i , δ_i , γ_i and λ_{i1} are summarized for the selected parts, and the fitted value of y_{it} in the fifth period of each part's life cycle is displayed as " $y_{.5}$ ". Also shown are the product-level location parameters a_j , as well as scale parameters σ_α , σ_a and σ_y . Each posterior distribution is summarized graphically by a condensed box-and-whisker plot,¹⁰ with the distribution's mean and standard deviation given numerically.

¹⁰ "Boxes" delimit the interquartile range of the distributions, and "whiskers" extend 1.5 times the interquartile range from the ends of the boxes—see Tukey (1977) for further details.

BAYESIAN FORECASTING OF PARTS DEMAND

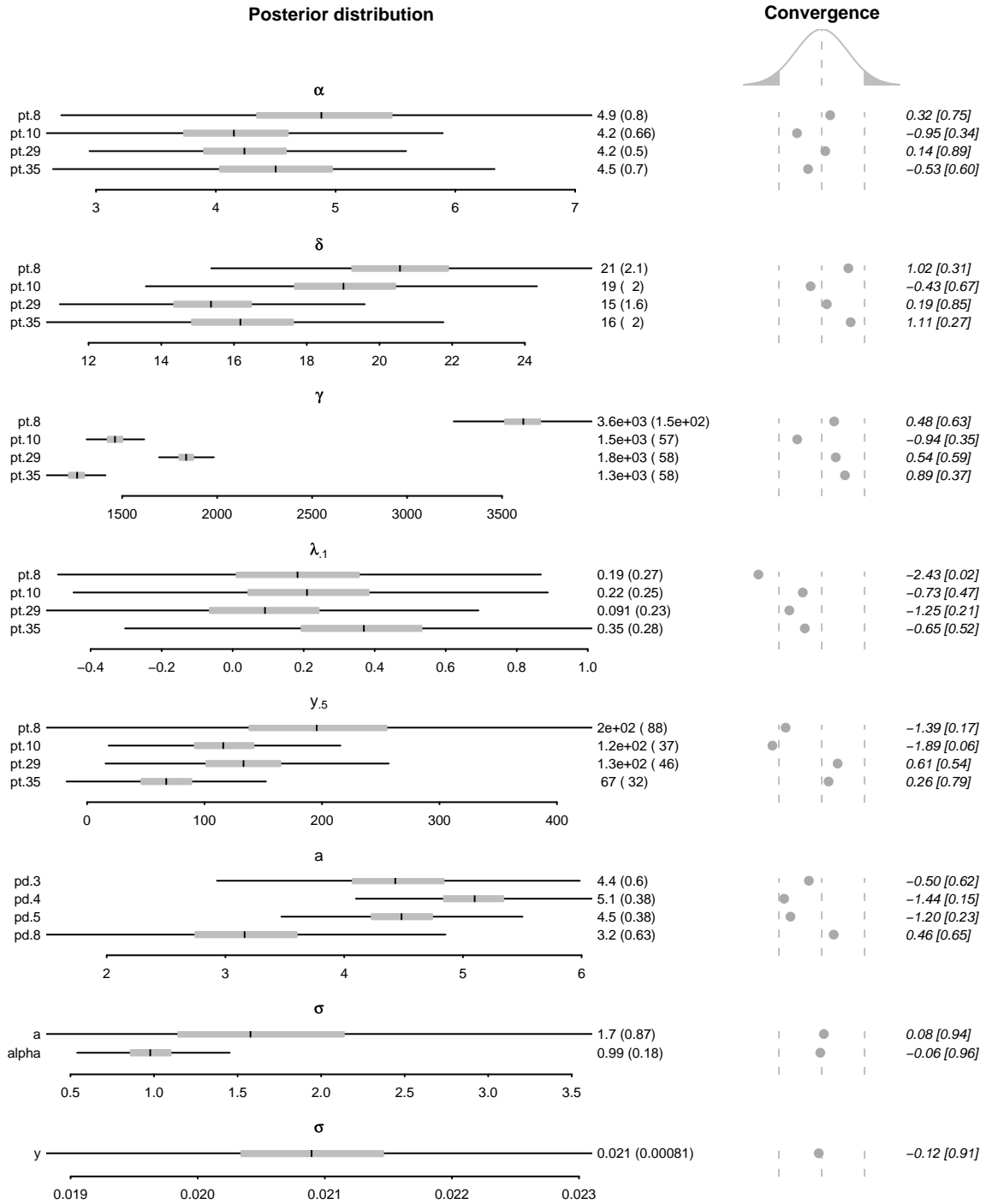


Figure 6. Posterior distributions and convergence diagnostics

The right-hand side of the Figure plots Geweke’s (1992) convergence diagnostic for each of the quantities in question. Derived from a comparison of the first and last segments of the Markov chain associated with a given quantity, Geweke’s statistic z has an asymptotically standard normal distribution if the chain is stationary (i.e. convergence has occurred). On the diagram, values of z are plotted on the 5th and 95th percentiles of the standard normal distribution; note that with 27 quantities displayed, we would expect around 3 values of z to fall outside of the percentile bounds even if convergence has been achieved.

4 Testing Forecast Performance

To test the forecasting effectiveness of the model, we use it to produce forecasts for the demand series sample described in Section 1.1, and compare its performance with that of the three other forecasting methods.

4.1 Setup

Testing begins with the collection of 45 parts in Section 1.1, which are used in 9 products in all. Each part i is associated with a period S_i , the period in which the part was first used (and in which demand for the part was first observed), and a life cycle length L_i , the number of periods during which the part was in use. Of the parts in the collection, all the parts from 3 randomly-chosen products—16 parts in total—are held in reserve to allow for calibration of the forecasting methods. There remain 29 holdout parts, belonging to 6 products, with life cycles together spanning 54 planning periods. The forecast horizon is a single planning period, since this encompasses the lead time of the bulk of the manufacturing parts at issue. Forecasts are made using the candidate methods in each of the spanned planning periods. In forecast period t , the following information is made available to the forecast methods:

Reserved parts Full demand histories are provided for all of the reserved parts, as sequences of the form $(y_{i1}, \dots, y_{iL_i})$, for each reserved part i , where $y_{it'}$, $t' = 1, \dots, L_i$, is the demand observed for the part in the t' th period of its life cycle (recall that L_i is the length of part i ’s lifecycle).

Holdout parts For each holdout part i such that $S_i \leq t + 1$, let $T_i = \min(L_i, t + 1 - S_i)$. Assemble a series which is empty if $T_i = 0$, and otherwise is of the form $(y_{i1}, \dots, y_{iT_i})$, where for $t' = 1, \dots, T_i$, value $y_{it'}$ is the demand for part i observed in period $S_i + t' - 1$.

In summary, the data provided to the forecasting methods in each period consists of a collection of demand histories, $\mathbf{y}_1, \dots, \mathbf{y}_N$, of lengths T_1, \dots, T_N , where demand history \mathbf{y}_i comprises the sequence $(y_{i1}, \dots, y_{iT_i})$ for $T_i > 0$, and the empty sequence otherwise. A

forecast \hat{y}_{iT_i+1} is required for each part i for which $T_i < L_i$,¹¹ corresponding to the actual demand for part i in the period $S_i + T_i + 1$.

The four candidate forecasting methods are described below.

Bench This is an all-but-trivial “benchmark” method that serves as a baseline for comparison. For $T_i > 0$, it simply repeats the observation from the previous period. For the first period of a part’s life cycle (i.e. when $T_i = 0$), it uses the mean of the first-period demands for other parts of the same product, if there are any suitable demand histories in the data provided to the method, and the mean of the first-period demands of all parts otherwise:

$$\hat{y}_{iT_i+1} = \begin{cases} y_{iT_i} & \text{if } T_i > 0 \\ \text{mean}\{y_{i'1} \mid i' \in \text{prec}_i\} & \text{otherwise.} \end{cases} \quad (16)$$

Here, the expression “mean S ” denotes the arithmetic mean of the set S , and the set prec_i is defined as follows:

$$\text{prec}_i = \begin{cases} P_i & \text{if } P_i \neq \emptyset, \\ \{1, \dots, N\} & \text{otherwise} \end{cases} \quad (17)$$

where

$$P_i = \{i' \in 1 \dots N \mid \text{prod}(i) = \text{prod}(i') \text{ and } T_{i'} > 0\}.$$

Judg This emulates the forecasting method currently used in Sun’s supply chain. It relies on the provision of *in period* $t - 1$ of a forecast \hat{z}_{jt} of total quarterly demand for the product j in the quarter into which planning period t falls.¹² For example (with minor abuse of notation), $\hat{z}_{j\text{Q2P1}}$ is a forecast of platform demand in quarter Q2 that is published in planning period Q1P3 (the period immediately prior to Q2P1). Then with prec_i defined as in equation (17), define an “attach rate” for part i in period $t \in \{0, \dots, T_i\}$:

$$\rho_{it} = \begin{cases} y_{it} / \hat{z}_{\text{prod}(i)t} & \text{if } t > 0, \\ \text{mean}\{\rho_{i'1} \mid i' \in \text{prec}_i\} & \text{otherwise.} \end{cases} \quad (18)$$

Now (recalling that $\hat{z}_{\text{prod}(i)T_i+1}$ is available in period T_i) let:

$$\hat{y}_{iT_i+1} = \rho_{iT_i} \hat{z}_{\text{prod}(i)T_i+1} \quad (19)$$

ExpS As an exemplar of the exponential smoothing techniques outlined in Section 1.2, we use the forecast package developed for the R statistical programming envi-

¹¹ Note that this precludes all of the reserved parts and those of the holdout parts whose life cycles concluded on or before t .

¹² In practice, this forecast is provided by the Company’s sales and marketing units.

ronment (Venables and Smith 2002) by Hyndman and Khandakar (2008). An embodiment of the forecasting framework developed by Hyndman, Koehler, Snyder, and Grose (2002)—itself an outgrowth of the structural approach to exponential smoothing detailed at length by Hyndman et al. (2008)—the forecast package provides for the automatic selection, estimation and extrapolation of an exponential smoothing model from a taxonomy of such models that incorporate additive or multiplicative seasonal and error components and/or additive, multiplicative or damped trends. Its performance is demonstrably superior in a number of applications, as Hyndman et al. (2002) relate.

A number of considerations recommended the forecast package as candidate for comparison: a) The package provides access to state-of-the-art exponential forecasting models, together with sophisticated techniques for selecting, fitting and forecasting with them; b) by operating entirely automatically, the package represents a plausible representative of the best “off the shelf” technology that might be available to operations managers; and c) again, by dint of its automatic operation, the package avoids the possibility that incompetence or invidiousness on our part might affect its performance in the comparison.

Forecasting the value $\hat{y}_{i T_i+1}$ given the (possibly empty) demand history \mathbf{y}_i is very straightforward:

- (1) Load the package forecast into R.
- (2) Attempt to fit an exponential smoothing model to \mathbf{y}_i using the package’s `ets(.)` function.
- (3) If the attempt is unsuccessful (that is, if `ets(.)` returns an error—almost invariably because \mathbf{y}_i is too short), use the forecast produced by applying the `Bench` method to the same series.
- (4) Otherwise, apply the function `forecast(.)` to the model produced in step 3. The point forecast used for $\hat{y}_{i T_i+1}$ is the mean component of the list returned by `forecast(.)`.¹³

Mod This forecasting method relies on *posterior predictive distributions* calculated by Bayesian updating of the model described in the previous sections, given the demand histories provided in each period of the test. With demand histories $\mathbf{y}_1, \dots, \mathbf{y}_N$, the posterior distribution for the quantity $y_{i T_i+1}$ whose value is to be forecast is given by $p(y_{i T_i+1} | \mathbf{y}_1, \dots, \mathbf{y}_N)$, which is itself calculated by marginaliz-

¹³ R code for steps 2 – 4 is (roughly) as follows:

```
m <- ets(y)
if (class(m) == "try-error") return(benchmark.forecast(y))
forecast(m, 1)$mean$
```

ing over the posterior density for the model parameters (here represented by θ):

$$p(y_{i T_i+1} | \mathbf{y}_1, \dots, \mathbf{y}_N) = \int p(y_{i T_i+1} | \theta) p(\theta | \mathbf{y}_1, \dots, \mathbf{y}_N) d\theta \quad (20)$$

$$\propto \int p(y_{i T_i+1} | \theta) p(\mathbf{y}_1, \dots, \mathbf{y}_N | \theta) p(\theta) d\theta \quad (21)$$

The integral expression in equation (21) is readily approximated using the Gibbs sampler described in the previous section, as described by Albert (2008), for example. For the point forecast $\hat{y}_{i T_i+1}$, we use the mean of the posterior predictive distribution of $y_{i T_i+1}$.

4.2 Results

The exercise described in the previous section resulted in rolling one-step-ahead forecasts for the 29 parts in the holdout set. A preliminary illustration of the relative performance of the forecast methods in the comparison is provided by the Tukey (1977)-style box-and-whisker plots in Figure 7. These plots display the mean absolute errors, root mean percentage errors and mean absolute percentage errors produced for all the holdout parts by the methods ExpS, Judg and Mod, expressed as ratios to the corresponding errors from the benchmark method, Bench (formal definitions of the error metrics used may be found in Section D). Note that in connection with these relative error measures, a value less than one indicates that the method's performance exceeds that of the benchmark, while a metric greater than one suggests that the (much cheaper) benchmark has produced a smaller error.

By this standard (note that a horizontal line at the value one is drawn across each panel of the Figure), it appears that the forecast performance of the model (Mod) generally exceeds that of the benchmark—for every metric, fully 75% of the series (delimited by the upper “hinge” of the appropriate box) have errors under Mod smaller than those of the benchmark. The performance of the exponential smoothing and judgment-based methods (ExpS and Judg, resp.) is not nearly so compelling, with 50% or more of all series results (the line across each box indicates the median of the sample) worse than the benchmark.

These conclusions are confirmed by the summary statistics in Table 1: Overall, the relative metrics of ExpS and Judg exceed one, while those of Mod are less than one. ExpS and Judg are evenly matched in regard of the relative MAE metric, while ExpS fares slightly better according to the relative MSE, and Judg has a lower mean relative MAPE (of course both methods are outperformed by the benchmark method on all counts).

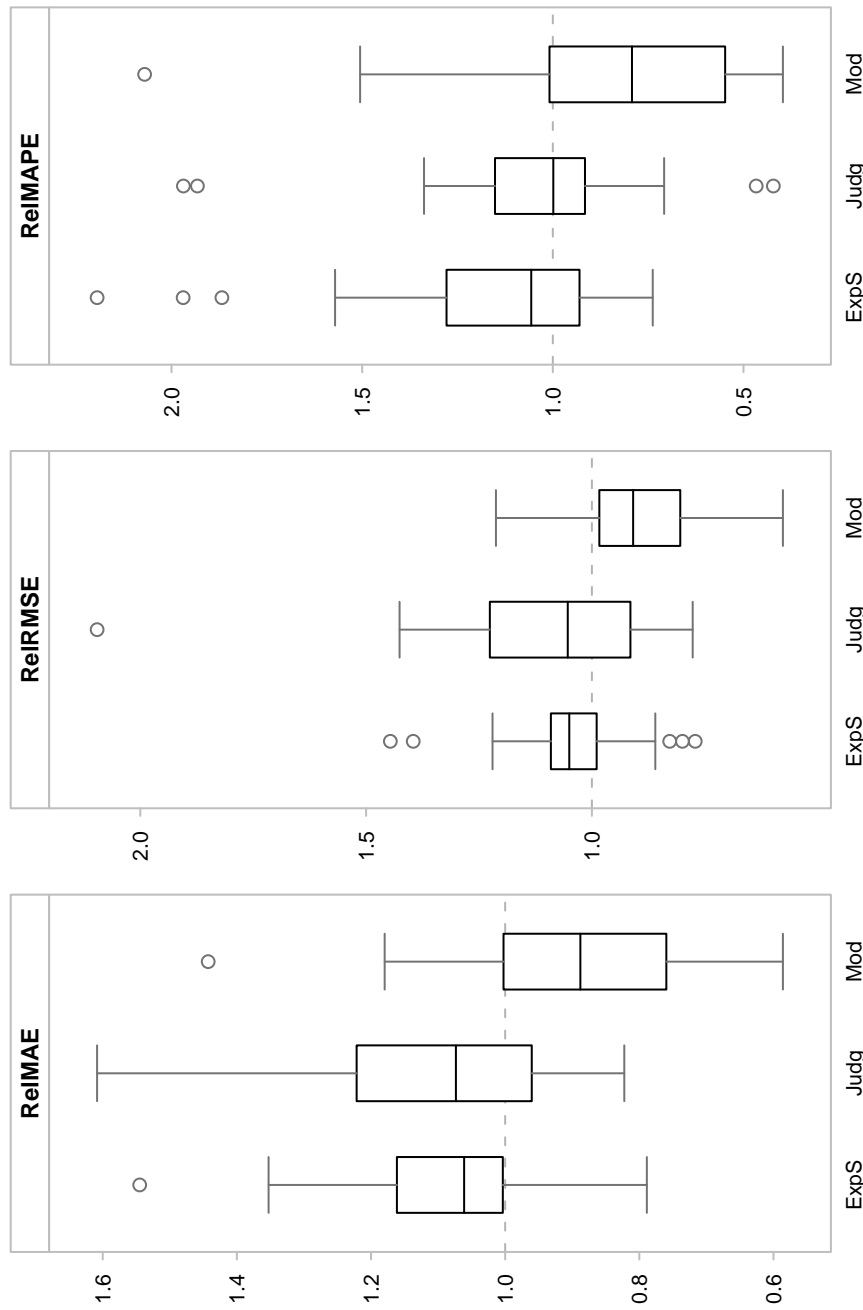


Figure 7. Relative performance statistics by series

	ExpS	Judg	Mod
RelMAE	1.08	1.08	0.88
RelRMSE	1.04	1.08	0.88
RelMAPE	1.13	1.01	0.78

Table 1. Summary relative performance statistics

4.3 Statistical Significance of the Results

Though the sample distributions illustrated in Figure 7 do appear to indicate superior predictive performance on the part of the model, as compared with the other methods in the comparison, one might reasonably wonder about the statistical basis for making such a statement of superior performance. Investigations into such matters date back at least to (Granger and Newbold 1977), though the seminal recent work is that of Diebold and Mariano (1995)—later modified in Harvey, Leybourne, and Newbold (1997). Extensions to Diebold and Mariano’s work—dealing with rolling forecasts, for example, where the forecasting model is updated period-by-period—have been produced by White (2000); Giacomini and White (2006) and Hansen (2005).

The framework adopted in (Diebold and Mariano 1995) and its successors relies on the examination of differences between sequences of forecast errors produced by applying different models to a single series; multiple series with differing scales—the focus of the comparison in this section—are largely outside of its ambit. Unfortunately, as Ashley (2003) demonstrates, in order to draw definitive conclusions from a single series, a substantial number (often exceeding 30) of holdout observations are required. In view of the relatively short series in our sample, therefore, direct application of the Diebold-Mariano framework seems inappropriate.

An approach more directly applicable in this context is described by Koning, Franses, Hibon, and Stekler (2005), who conduct a statistical analysis of comparative forecast performance across more than 3,000 time series.¹⁴ (Many of the same techniques are employed by Eugster, Hothorn, and Leisch (2008) to compare the performance of computer algorithms.) This approach begins by ranking forecast methods in terms of their forecast accuracy on each series in the comparison. Proceeding formally, given a metric of forecast

¹⁴ We should note that the approach of Koning et al. (2005) (and indeed the entire enterprise of significance testing for predictive superiority) is somewhat controversial—see (Thompson 1992), (Armstrong 2007) and (Stekler 2007).

	MAE	RMSE	MAPE
S statistic	20.8	20.5	20
p-value	1.19e−04	1.36e−04	1.73e−04

Table 2. Friedman S statistics and p-values

accuracy $\psi \in \{\text{MAE}, \text{RMSE}, \text{MAPE}\}$,¹⁵ each forecast method $m \in 1, \dots, M$ is assigned a rank $R_{mi} \in 1, \dots, M$ for each part i in the holdout set, such that $R_{mi} < R_{m'i}$ iff $\psi_{mi} < \psi_{m'i}$.¹⁶

Next, the non-parametric test of Friedman (1940)—extending (Kendall and Babington Smith 1939)—is used to test for the null hypothesis that all the methods are actually equal in statistical accuracy (the alternative hypothesis being that at least two methods differ in accuracy). We compute the Friedman statistic S as follows:

$$S = \frac{12N}{M(M+1)} \sum_{m=1}^M \left(\bar{R}_m - \frac{m+1}{2} \right)^2, \quad \text{where } \bar{R}_m = \frac{1}{N} \sum_{i=1}^N R_{mi}. \quad (22)$$

Under the null hypothesis, the S statistic has an asymptotic χ^2 distribution with $M - 1$ degrees of freedom. Table 2 sets out the values of S calculated for the methods Bench, ExpS, Judg and Mod, ranked according to their performance as measured by each of the forecast metrics, along with the associated p-values. Clearly, there are substantial ground for rejecting the hypothesis that all methods are of equal accuracy.

For more detailed comparisons of predictive performance, Koning et al. (2005) use another non-parametric test due to McDonald and Thompson (1967), based on earlier work of Nemenyi (1963) and Wilcoxon (1945). This is a multiple testing procedure involving component null hypotheses that methods m and m' have equal predictive performance, for all pairs of methods (m, m') . According to this test, such a component null hypothesis is rejected iff:

$$|\bar{R}_m - \bar{R}_{m'}| \geq r_\alpha, \quad (23)$$

where \bar{R}_m is defined in equation (22) and r_α is a constant chosen so that the *experiment-wise error rate* (that is the probability of making any incorrect judgments when all methods actually have equal predictive performance) equals α . In large samples, as shown by Hollander and Wolfe (1999, p. 296):

$$r_\alpha \approx q_\alpha \left[\frac{NM(M+1)}{12} \right]^{\frac{1}{2}}, \quad (24)$$

¹⁵ See Section D for definitions of these metrics.

¹⁶ Equal accuracies are accorded the average of all ranks in the tie—see (Hollander and Wolfe 1999, p. 273). Full details of all the tests, used in this section may be found in (*Ibid.*, chp. 7).

where q_α is the upper α^{th} percentile point of the distribution of the range of M independent standard normal variables.

In the interests of perspicuity, Koning et al. (2005) recommend a graphical representation of the results of the Wilcoxon, Nemenyi, McDondald-Thompson test. Figure 8 is such an illustration of the application of the test (with α set to 10%) to the forecast comparison in this section. Here, for each forecast metric, each method is represented by a vertical line centered at \bar{R}_m with length r_α ; if the lines of any pair of methods do not overlap, we can reject the hypothesis that their performance is equal. In addition, a horizontal line is drawn at the top of the line corresponding to the best performing method, so that all methods with confidence intervals above this line perform significantly worse than the best method. From the Figure, it appears that the model's performance *is* significantly superior (at the 10% level) to that of the other methods, at least as measured by the RMSE and MAPE metrics, though the comparisons of MAE measurements are somewhat less definitive.¹⁷

5 Conclusions

Based upon the evaluation in the previous section, it does appear that the Bayesian forecasting model presented here is well-suited to forecasting parts demand at Sun. The statistical calculations of Section 4.3 notwithstanding, however, we would be wary of recommending the model as a panacea for inventory forecasting problems in general. After all, the model was developed specifically to encompass specific characteristics of Sun's parts demands; despite its relatively poor showing here, the exponential smoothing system that underlies the ExpS forecasting method of the previous section performs well in a wide variety of applications, as Hyndman et al. (2002) demonstrate. And as Thompson (1992) points out—commenting on work of Fildes (1992), dealing with demands experienced by a telecommunications company—statistical demonstrations of forecasting efficacy in narrowly-defined domains may not extend beyond those domains. This is not to say, of course, that the constituents of the model presented here (a parameteric representation of life cycles, hierarchical priors for pooling demand information and so on) would not be of use in other contexts. More broadly, a number of general observations might also be made:

First, it is notable that a statistical model as complex as that set out in Section 2 is required to exceed that forecast performance of the trivial method Bench. By contrast, both an intuitively appealing process involving the consensus judgments of trained professionals (Judg) and a sophisticated general-purpose forecasting package (ExpS) perform no better

¹⁷ In regard of the failure of the test to reject the null hypothesis (albeit barely) in the MAE case, note that controlling experimentwise error rates is considered by some researchers an unduly conservative approach to statistical inference—see Cox (1965), for example.

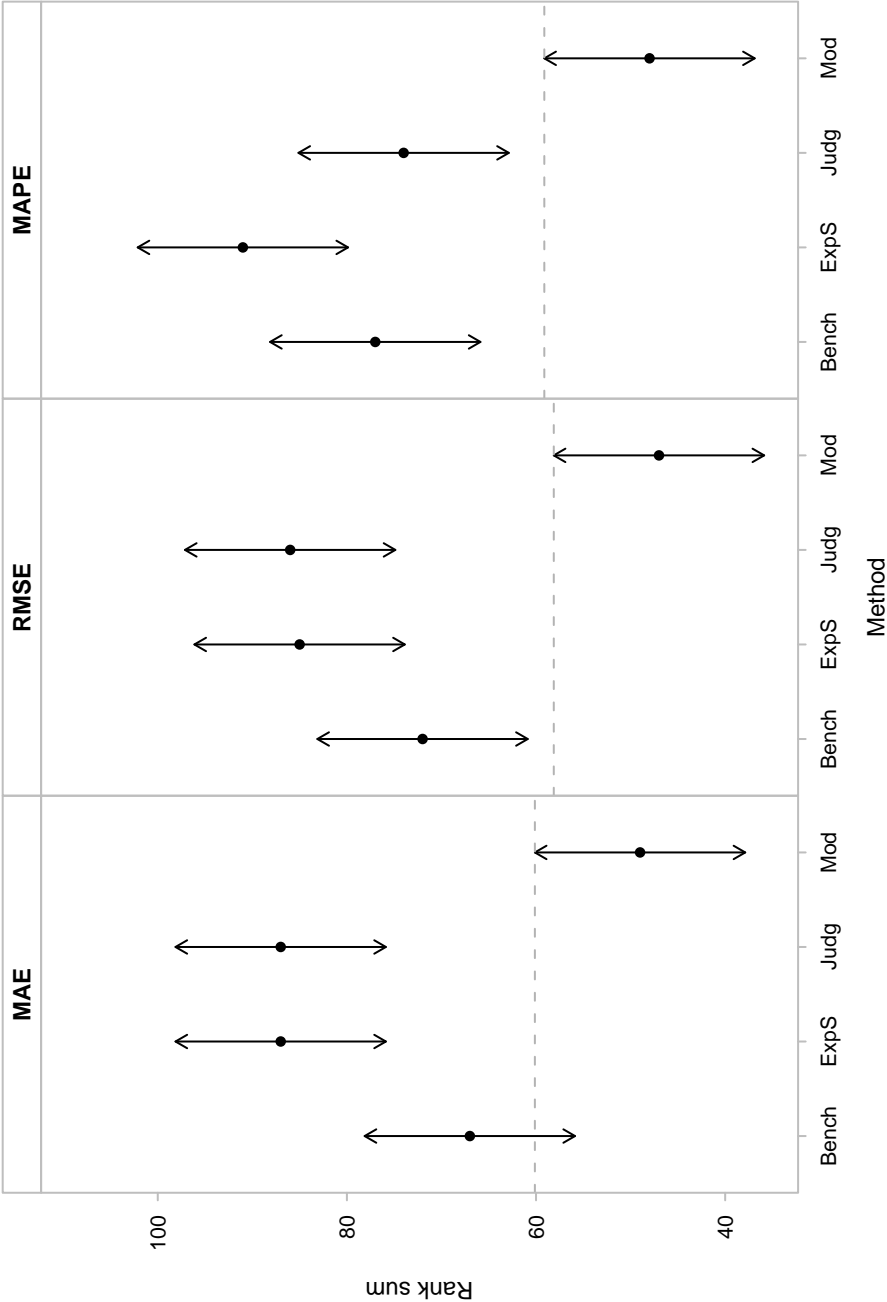


Figure 8. Results of Wilcoxon-Nemenyi-McDonald-Thompson test

than the benchmark (and possibly even a little worse). Thus while familiar and relatively straightforward classes of statistical model such as the ARIMA processes can provide interesting theoretical insights into supply chain demands (as Gilbert (2005) demonstrates, for example), researchers and practitioners should be wary of assuming that they effectively characterize real-world situations.

Given somewhat forbidding complexity of a model such as that described here, it might be felt preferable to explore policies that produce more predictable demands, or to adopt manufacturing strategies that are tolerant of sizable forecast errors. Certainly as an upstream participant in a fairly long demand chain (which involves resellers, OEM's, systems integrators and the like), there's little doubt that Sun experiences the "demand amplification" first illustrated by Forrester (1961) and now familiarly known as the *bullwhip effect* (Lee, Padmanabhan, and Whang 2004). Furthermore, practices that may be used to mitigate the bullwhip effect—in the form of information sharing and "lean" manufacturing procedures—are also well documented (Mason-Jones and Towill 2000; Womack, Jones, and Roos 2007). Such practices are usually neither cheap nor easy to realize, however; efforts at information sharing may involve costly IT investments, and may run afoul of commercial and legal impediments, and the continued competitive success of Toyota (the inventor of the lean approach) in the automobile industry attests to the difficulties associated with the implementation of lean manufacturing. Set against such considerations, a model like the one detailed in this paper—which was developed by a very small team, and which operates almost automatically once deployed—may be an attractive stop-gap. Furthermore, information sharing and lean manufacturing are hardly incompatible with improved forecasting tools, and even with effectively damped demand, short part life cycles will doubtless continue to be the norm in high technology industries, so that a model which properly captures the associated dynamics should continue to be relevant.

Finally, the computational expense of the Markov chain Monte Carlo algorithms required to support a complex Bayesian model operationally might prompt misgivings. The sheer power and affordability of modern computers should help dispel such reservations. For example, the forecast comparison delineated in the previous section entailed the calculation of some 582 individual forecasts, each of which required 4,000 iterations in the Gibbs sampler: Split across 16 CPU's (an investment of less than \$10,000 in 2009), the entire exercise was completed in less than 6 hours.¹⁸ While well short of interactive responsiveness, and certainly ill-matched to forecasting in a retail setting—where tens of thousands of stock keeping units may require forecasts—such a turn-around time is quite adequate for

¹⁸ For a discussion of the programming techniques involved in such an undertaking, see (Rossini, Tierney, and Li 2007).

applications discrete manufacturing like Sun's, where good demand forecasts are required for a few hundred expensive, highly-integrated parts.¹⁹

At the time of writing, limited deployment of the model has begun, but its adoption throughout an organization as extensive and far-flung as Sun's supply chain rests, of course, on the effective negotiation of factors beyond mere technical efficacy, as Rogers (2003) details. Our experience suggests that supply chain managers are initially reluctant to entrust mission-critical forecasting to something as untried and abstruse as a Bayesian statistical model. However, the same experience suggests that a track record of incontestably good forecast performance on the part of the model (usually garnered by running in parallel with normal operations, as it is currently) can help considerably to allay such reluctance, and—not incidentally—help improve the model, too.

References

- J. Albert. *Bayesian Computation with R*. Springer, 2008.
- J. Armstrong and F. Collopy. Error measures for generalizing about forecasting methods - empirical comparisons. *International Journal of Forecasting*, 8(1):69–80, 1992.
- J. S. Armstrong. Significance tests harm progress in forecasting. *International Journal of Forecasting*, 23:321–327, 2007.
- J. S. Armstrong and K. C. Green. Demand forecasting: Evidence-based methods. Working paper 24/05, Dept. of Econometrics and Business Statistics, Monash University, 2005.
- K. Arrow, S. Karlin, and H. Scarf, editors. *Studies in the Mathematical Theory of Inventory Production*. Stanford University Press, 1958.
- R. Ashley. Statistically significant forecasting improvements: How much out-of-sample data is likely necessary. *International Journal of Forecasting*, 19:229–239, 2003.
- Y. Aviv. A time-series framework for supply chain inventory management. *Operations Research*, 51(2):210–227, 2003.
- J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, 1994.
- J. Boylan. Intermittent and lumpy demand: A forecasting challenge. *Foresight: The International Journal of Applied Forecasting*, 1:36–42, June 2005.
- R. G. Brown. *Statistical Forecasting for Inventory Control*. McGraw-Hill Book Company, 1959.
- P. Congdon. *Applied Bayesian Modelling*. Wiley, 2003.
- D. R. Cox. A remark on multiple comparison methods. *Technometrics*, 7(2):223–224, 1965.
- F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13:253–263, 1995.

¹⁹ In fact, the well-established practice of *ABC-based* inventory management Heizer and Render (1999, p. 441) relies on the fact that a small proportion of the parts in most inventories account for much of the dollar value therein, and so merit special attention.

- G. T. Duncan, W. L. Gorr, and J. Szczypula. Forecasting analogous time series. In J. S. Armstrong, editor, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pages 195–213. Kluwer, 2001.
- J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford, 2001.
- W. Enders. *Applied Econometric Time Series*. Wiley, 1995.
- M. J. A. Eugster, T. Hothorn, and F. Leisch. Exploratory and inferential analysis of benchmark experiments. Technical Report 030, Department of Statistics, University of Munich, 2008. URL <http://www.stat.uni-meunchen.de>.
- R. Fildes. The evaluation of extrapolative forecasting methods. *International Journal of Forecasting*, 8:81–98, 1992.
- J. W. Forrester. *Industrial Dynamics*. Pegasus, 1961.
- M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.
- S. Frühwirth-Schnatter. Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15:183–102, 1994.
- S. Frühwirth-Schnatter and S. Kaufmann. Model-based clustering of multiple time series. *Journal of Business and Economic Statistics*, 25(1):78–89, January 2008.
- E. S. Gardner. Exponential smoothing: The state of the art. *Journal of Forecasting*, 4(4):1–28, 1985.
- E. S. Gardner. Evaluating forecast performance in an inventory control system. *Management Science*, 36:490–499, 1990.
- Everette Jr. Gardner. Exponential smoothing: The state of the art—part II. *International Journal of Forecasting*, 22(4):637–666, 2006.
- A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.
- A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.
- A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Press, Boca Raton, 2nd edition, 2003.
- J. Geweke. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics*, volume 4, pages 169–193. Clarendon Press, 1992.
- J. Geweke and H. Tanizaki. Note on the sampling distribution for the Metropolis-Hastings algorithm. *Communications in Statistics – Theory and Methods*, 32(4):775–789, 2003.
- R. Giacomini and H. White. Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578, November 2006.
- K. Gilbert. An ARIMA supply chain model. *Management Science*, 51(2):305–310, February 2005.
- W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 1996.
- C. W. J. Granger and P. Newbold. *Forecasting Economic Time Series*. Academic Press, 1977.

- W. Griffiths. A Gibbs' sampler for the parameters of a truncated multivariate normal distribution. In R. Becker and S. Hum, editors, *Contemporary Issues in Economics and Econometrics: Theory and Application*, pages 75–91. Edward Elgar, 2004.
- P. R. Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, October 2005.
- A. Harvey and R. D. Snyder. Structural time series models for inventory control. *International Journal of Forecasting*, 6:187–198, 1990.
- D. Harvey, S. Leybourne, and P. Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13:281–291, 1997.
- J. Heizer and B. Render. *Principles of Operations Management*. Prentice-Hall, 3rd edition, 1999.
- M. Hollander and D. A. Wolfe. *Nonparametric Statistical Methods*. Wiley, 1999.
- R. J. Hyndman and Y. Khandakar. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3):1–22, July 2008. URL <http://www.jstatsoft.org/v27/i03>.
- R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.
- R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18(3):439–454, 2002.
- R. J. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer, 2008.
- M. G. Kendall and B. Babington Smith. The problem of m rankings. *The Annals of Mathematical Statistics*, 10(3):275–287, 1939.
- A. J. Koning, P. H. Franses, M. Hibon, and H. O. Stekler. The M3 competition: Statistical tests of the results. *International Journal of Forecasting*, 21:397–409, 2005.
- L. Lapide. Evolution of the forecasting function. *Journal of Business Forecasting*, pages 22–28, 2006. Spring Issue.
- H. L. Lee, V. Padmanabhan, and S. Whang. Information distortion in a supply chain: The bullwhip effect. *Management Science*, 50(12):1875–1886, December 2004.
- V. Mahajan, E. Muller, and Y. Wind, editors. *New-Product Diffusion Models*. Kluwer, 2000.
- R. Mason-Jones and D. R. Towill. Coping with uncertainty: Reducing “bullwhip” behaviour in global supply chains. *Supply Chain Forum*, 1:40–44, 2000.
- B. P. M. McCabe and G. M. Martin. Bayesian predictions of low count time series. *International Journal of Forecasting*, 21(2):315–330, 2005.
- B. J. McDonald and W. A. Thompson, Jr. Rank sum multiple comparisons in one- and two-way classifications. *Biometrika*, 54:487–497, 1967.
- W. W. Moe and P. S. Fader. Using advance purchase orders to forecast new product sales. *Marketing Science*, 21(3):347–364, Summer 2002.
- P. Nemenyi. *Distribution-free Multiple Comparisons*. PhD thesis, Princeton University, 1963.
- J. A. Norton and F. M. Bass. A diffusion theory model of adoption and substitution for successive generations of high-technology products. *Management Science*, 33(9):1069–1086, 1987.

- Y. Qi and T. P. Minka. Hessian-based Markov chain Monte Carlo algorithms. In *First Cape Cod Workshop on Monte Carlo Methods*, 2002.
- E. M. Rogers. *Diffusion of Innovations*. Free Press, 5 edition, 2003.
- P. E. Rossi, G. M. Allenby, and R. McCulloch. *Bayesian Statistics and Marketing*. Wiley, 2005.
- A. J. Rossini, L. Tierney, and N. Li. Simple parallel statistical computing in R. *Journal of Computational and Graphical Statistics*, 16:399–420, 2007.
- N. Sanders and K. B. Manrodt. Forecasting practices in US corporations: Survey results. *Interfaces*, 24:92–100, 1994.
- D. C. Schmittlein and V. Mahajan. Maximum likelihood estimation for an innovation diffusion model of new product acceptance. *Marketing Science*, 1(1):57–78, Winter 1982.
- L. Shenstone and R. J. Hyndman. Stochastic models underlying Croston’s method for intermittent demand forecasting. *Journal of Forecasting*, 24:389–402, 2005.
- R. D. Snyder, A. Koehler, and K. Ord. Forecasting for inventory control with exponential smoothing. *International Journal of Forecasting*, 18(1):5–18, 2002.
- D. J. Spiegelhalter, A. Thomas, N. G. Best, W. R. Gilks, and D. Lunn. BUGS: Bayesian inference using Gibbs sampling. Technical report, MRC Biostatistics Unit, Cambridge, England, 2003. URL www.mrc-bsu.cam.ac.uk/bugs.
- V. Srinivasan and C. H. Mason. Nonlinear least squares estimation of new-product diffusion models. *Marketing Science*, 5(2):169–178, Spring 1986.
- H. O. Stekler. Significant tests harm progress in forecasting: Comment. *International Journal of Forecasting*, 23:329–330, 2007.
- D. Sun and J. O. Berger. Objective priors for the multivariate normal model. In *ISBA 8th World Meeting on Bayesian Statistics*, 2006.
- P. A. Thompson. A statistician in search of a population: A commentary on error measures. *International Journal of Forecasting*, 8(1):103–104, 1992.
- J.W. Tukey. *Exploratory data analysis*. Addison-Wesley, 1977.
- A. F. Veinott. Optimal policy for a multi-product, dynamic nonstationary inventory problem. *Management Science*, 12(3):206–222, November 1965.
- W. N. Venables and D. M. Smith. *An Introduction to R*. Network Theory, 2002.
- M. West and P. J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York, 2nd edition, 1997.
- H. White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, September 2000.
- F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.
- J. P. Womack, D. T. Jones, and D. Roos. *The Machine That Changed the World: The Story of Lean Production*. Free Press, 2007.
- P. Yelland. Bayesian forecasting for low-count time series using state-space models: An empirical evaluation for inventory management. *International Journal of Production Economics*, 2008. doi: 10.1016/j.ijpe.2008.08.027. URL <http://dx.doi.org>. In press; available online 22 August 2008.
- P. M. Yelland. A model of the product lifecycle for sales forecasting. Technical Report 127, Sun Microsystems Laboratories, 2004. URL http://research.sun.com/techrep/2004/smli_tr-2004-127.pdf.

BAYESIAN FORECASTING OF PARTS DEMAND

J. Yurkiewicz. 2006 forecasting software survey. *OR/MS Today*, 33(4):40–49, August 2006.
URL <http://www.lionhrtpub.com/orms/orms-8-06/frsurvey.html>.

A Notation

Notation	Meaning
<i>Indexes</i>	
$i, k \in \{1, \dots, N\}$	Indexes ranging over parts.
$t \in \{1, \dots, T_i\}$	Index ranging over planning periods, for part i .
$j \in \{1, \dots, J\}$	Index ranging over product platforms.
$m, m' \in 1, \dots, M$	Indexes ranging over forecasting methods.
$\text{prod}(i)$	Index of the product to which part i belongs.
$\text{parts}(j)$	Indexes of the parts belonging to product j .
<i>Model exposition</i>	
y_{it}	Demand for part i in period t ; aggregated as $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i})$.
s_i	Observed total demand for part i over its entire lifetime (if available).
$W(t \lambda_i, k_i)$	The value of the Weibull CDF indexed by λ_i and k_i .
$\Delta W(t \lambda_i, k_i)$	The difference $\Delta W(t \lambda_i, k_i) - \Delta W(t-1 \lambda_i, k_i)$.
$\Delta W(t \alpha_i, \delta_i)$	Differenced Weibull CDF indexed by α_i and δ_i (see below).
<i>First-level model parameters</i>	
α_i	20 th percentile of the Weibull distribution representing the part life cycle.
δ_i	Difference between 95 th and 20 th percentiles of the Weibull distribution representing the part life cycle.
γ_i	Scale factor associated with part i ; loosely, putative life cycle demand for part i .
ζ_{it}	Equal to 1 if y_{it} is deemed an outlier, 0 otherwise.
x_{it}	Value of the latent autoregressive process for part i in period t .
$\lambda_i = (\lambda_{i1}, \lambda_{i2})^\top$	Coefficients of autoregression for part i .
<i>Prior hierarchy</i>	
a_j, d_j, g_j	Locations of platform-level priors for part parameters α_i, δ_i and γ_i , resp., for $i \in \text{parts}(j)$.
$\mu_\theta, \sigma_\theta$	Parameters (usually mean and std. dev., resp.) of prior distribution for generic parameter θ .
<i>Test setup</i>	
S_i	The period in which part i was first used.
L_i	The length of part i 's entire life cycle.
\hat{y}_{it}	The forecast value of y_{it} .

continued on next page

Notation	Meaning
	<i>Miscellaneous</i>
\emptyset	The empty set.
mean S	The mean of set S .
$\mathcal{I}(\cdot)$	The indicator function, equal to 1 if its argument is true, and 0 otherwise.
\mathbf{x}^\top	The transpose of \mathbf{x} .

B Standard Probability Distributions

Distribution	Description	Density/mass function
$N(\mu, \sigma^2)$	Normal distribution with mean μ and standard deviation σ .	$N(x \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2\sigma^2}(x - \mu)^2\right]$
Weib(λ, k)	Weibull distribution with shape λ and scale k .	$Weib(x \lambda, k) = \lambda k^{-\lambda} \theta^{-1+\lambda} e^{-\left(\frac{\theta}{k}\right)^\lambda},$ $x \geq 0$
Bern(p)	The Bernoulli distribution with success probability p .	$Bern(x p) = p^x(1 - p)^{(1-x)},$ $x \in \{0, 1\}$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution with mean $\boldsymbol{\mu}$ and +ve definite $d \times d$ covariance matrix $\boldsymbol{\Sigma}$.	$N(\boldsymbol{x} \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} \boldsymbol{\Sigma} ^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right]$
$N_{[0, \infty)}(\mu, \sigma^2)$	The normal distribution $N(\mu, \sigma^2)$, truncated on the left at 0.	$N_{[0, \infty)}(x \mu, \sigma^2) = 2N(x \mu, \sigma^2),$ $x \geq 0$
Exp(λ)	The exponential distribution with parameter λ .	$Exp(x \lambda) = \lambda e^{-\lambda x}, x > 0$
Mult($n; p_1, \dots, p_k$)	The multinomial distribution with n trials and bin probabilities p_1, \dots, p_k .	$Mult(x n; p_1, \dots, p_k) = \binom{n}{x_1, \dots, x_k} p_1^{x_1}, \dots, p_k^{x_k},$ $x_j = 0, 1, 2, \dots, n; \sum x_j = n$
Inv- $\chi^2(\nu)$	The inverse chi-squared distribution with ν degrees of freedom.	$Inv-\chi^2(x \nu) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} x^{-(\nu/2+1)} \exp[-1/(2x)],$ $x > 0$

C The Geweke-Tanizaki (2003)²⁰ “Taylored Chain”

With x the current state of the sampler, to produce a proposal x^+ for a target kernel $p(z)$, let $q(z) = \log p(z)$, with $q'(z)$ and $q''(z)$ the first and second derivatives thereof. Proceed by cases:

²⁰ Sampling techniques similar to the Taylored chain are also discussed by Qi and Minka (2002).

Case 1: $q''(x) < -\varepsilon$, where ε is a suitable small constant, such as 0.1.²¹

Rewrite the Taylor expansion of $q(z)$ around x :

$$\begin{aligned} q(z) &\approx q(x) + q'(x)(z-x) + \frac{1}{2}q''(x)(z-x)^2 \\ &= q(x) - \underbrace{\frac{1}{2}(-q''(x)) \left(z - \left(x - \frac{q'(x)}{q''(x)} \right) \right)^2}_{\dagger} \end{aligned}$$

Since $q''(x) < 0$, the component (\dagger) of the latter expression constitutes the exponential part of a normal distribution, which implies that the target kernel in the vicinity of x may be approximated by a normal distribution with mean $x - q'(x)/q''(x)$ and standard deviation $1/\sqrt{-q''(x)}$; sample x^+ accordingly.

Case 2: $q'(x) \geq -\varepsilon$ and $q'(x) < 0$

Approximate $q(z)$ by a line passing through x and x_1^* , the largest mode of $q(z)$ smaller than x :

$$q(z) \approx q(x_1^*) + \underbrace{\frac{q(x_1^*) - q(x)}{x_1^* - x}}_{\dagger} (z - x_1^*)$$

In this case, the component (\dagger) indicates an exponential distribution, and the proposal is:²²

$$\begin{aligned} x^+ &= \hat{x}_1 + w, \quad \text{where } w \sim \text{Exp}(\lambda_1), \\ \lambda_1 &= \left| \frac{q(x_1^*) - q(x)}{x_1^* - x} \right|, \\ \hat{x}_1 &= x_1^* - 1/\lambda_1 \end{aligned}$$

Case 3: $q'(x) \geq -\varepsilon$ and $q'(x) > 0$

Approximate $q(z)$ by a line passing through x and x_2^* , the *smallest* mode of $q(z)$ *larger* than x . The proposal is developed in a manner parallel to that in Case 2:

$$\begin{aligned} x^+ &= \hat{x}_2 - w, \quad \text{where } w \sim \text{Exp}(\lambda_2), \\ \lambda_2 &= \left| \frac{q(x_2^*) - q(x)}{x_2^* - x} \right|, \\ \hat{x}_2 &= x_2^* - 1/\lambda_2 \end{aligned}$$

Case 4: $q'(x) \geq -\varepsilon$ and $q'(x) = 0$

²¹ By ensuring that $|q''(x)| > 0$, using ε rather than 0 reduces the occurrence of proposed values that depart too markedly from the current state.

²² The origin of the proposal, \hat{x}_1 , is offset from the mode x_1^* in order to guarantee irreducibility of the resulting Markov chain; see (Geweke and Tanizaki 2003) for details.

In this instance, x^+ is sampled from a uniform distribution over a range $[x_1, x_2]$, such that $x_1 < x < x_2$. End points x_1 and x_2 are set to suitable modes of $q(\cdot)$, if they can be found, and to user-supplied values otherwise.

D Forecast Error Metrics

Assume that forecast method m provides forecast \hat{y}_{mit} for part i and $t \in 1, \dots, T_i$, and that y_{it} is the corresponding actual value. Then the *mean absolute error*, *root mean square error* and *mean absolute percentage error* resp. for method m and part i are defined as follows:

$$\begin{aligned} \text{MAE}_{mi} &= \frac{1}{T_i} \sum_{t=1}^{T_i} |\hat{y}_{mit} - y_{it}| \\ \text{RMSE}_{mi} &= \left[\frac{1}{T_i} \sum_{t=1}^{T_i} (\hat{y}_{mit} - y_{it})^2 \right]^{\frac{1}{2}} \\ \text{MAPE}_{mi} &= \frac{1}{T_i} \sum_{t=1}^{T_i} 100 \times \left| \frac{\hat{y}_{mit} - y_{it}}{y_{it}} \right| \end{aligned}$$

Relative versions of the above are defined by ratios to the corresponding error metric for the benchmark forecast method, Bench:²³

$$\begin{aligned} \text{RelMAE}_{mi} &= \text{MAE}_{mi} / \text{MAE}_{\text{Bench } i} \\ \text{RelRMSE}_{mi} &= \text{RMSE}_{mi} / \text{RMSE}_{\text{Bench } i} \\ \text{RelMAPE}_{mi} &= \text{MAPE}_{mi} / \text{MAPE}_{\text{Bench } i} \end{aligned}$$

Finally, average metrics for a particular method across all parts in the sample are derived using the geometric mean; for example:

$$\text{RelMAE}_m = \left[\prod_{i=1}^N \text{RelMAE}_{mi} \right]^{\frac{1}{N}}$$

²³ Metric RelMAE_{mi} is closely related to the *mean absolute scaled error* of Hyndman and Koehler (2006), and RelRMSE_{mi} conforms to one definition of Theil's U_2 statistic (Armstrong and Collopy 1992).