THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 8, October 1982
Printed in U.S.A.

Quality Evaluation Plan Using Adaptive Kalman Filtering

By M. S. PHADKE

(Manuscript received March 5, 1982)

An important function of the Bell Laboratories Quality Assurance Center and the Western Electric Quality Assurance Directorate is to audit the quality of the products manufactured and the services provided by the Western Electric Company to determine if the intended quality standards are met. Until the sixth period of 1980, the t-rate system was used to make inference on the product quality. Starting the seventh period of 1980, the Quality Measurement Plan (QMP) has been implemented. The QMP is based on an empirical Bayes model of the audit-sampling process using the current and the preceding five periods of data. Because it ignores the time order of the data, it is slow in responding to drifts in the process mean. The Quality Evaluation Plan (QEP) has been designed to take into account the time order of the data and to be more sensitive to drifts in the process mean. In this paper we present the Quality Evaluation Plan, which uses the entire time series of data on a given product to determine if that product meets the quality standard. The time series is modeled by a stochastic process, which allows for the possibility that the process mean may drift or fluctuate around a fixed value. An adaptive Kalman filtering theory is developed for filtering out the sampling variance and obtaining the best estimate of the true defect index and its confidence interval. Thus, in QEP the best estimate of the true defect index is obtained by a combination of adaptive exponential smoothing and shrinkage to the mean. The QEP computations are recursive, and the total computing efforts of QEP and QMP are roughly equal. The paper contains several examples to illustrate the QEP.

I. INTRODUCTION

An important function of the Bell Laboratories Quality Assurance Center and the Western Electric Quality Assurance Directorate is to

audit the quality of the products manufactured, and the services provided by the Western Electric Company to determine if the intended quality standards are met. This is achieved by dividing the products and services into some 3000 homogeneous classes. A small sample is taken from each class during each period (there are eight rating periods in a year). Based on this data, an inference is made in each period regarding the compliance of each class to the quality standard.

Until the sixth period of 1980, the *t*-rate system, evolved from the work of Dodge and others,[1] was used to rate the product quality. Starting with the seventh period of 1980, the Quality Measurement Plan (QMP) was implemented. The QMP, developed by A. B. Hoadley,[2] is based on an empirical Bayes model of the audit-sampling process. It uses the current and the preceding five periods of data. It represents a considerable improvement in the statistical power for detecting substandard quality as compared with the old rules based on the *t*-rate. However, QMP ignores the time order of the observations, so it is less sensitive to drifts in the process mean. The Quality Evaluation Plan (QEP) has been designed to take into account the time order of the data and to be more sensitive to drifts in the process mean.

The object of this paper is to present the Quality Evaluation Plan, which uses the entire time series of data on a given class to determine if that class meets the quality standard. The time series is modeled by a stochastic process, which allows for the possibility of the process mean to (*i*) drift or (*ii*) fluctuate around a fixed value. An adaptive Kalman filtering theory is developed for filtering out the sampling variance and obtaining the best estimate of the true defect index and its confidence interval. Some of the salient features of QEP are: (*i*) the best estimate of the defect index is obtained by an adaptive exponential smoothing process, making QEP more responsive to shifts and drifts in the process mean; (*ii*) the QMP model is a special case of the general model proposed here; and (*iii*) the computational method is recursive.

This paper is divided into nine sections. We describe the model in Section II. Section III gives the Kalman filter solution of the model. Adaptive estimates of the model parameters are developed in Section IV, and in Section V we modify the Kalman filter solution of Section III to reflect the fact that the model parameters are estimated. The solution thus obtained is the adaptive Kalman filter. The construction of the box chart for displaying the results, and the rules for the exception report are spelled out in Section VI. The selection of the starting values for the estimation of the model parameters and the adaptive Kalman filter solution is discussed in Section VII. The algorithm has been tried on a number of rating classes and also on simulated data. We present representative examples in Section VIII.

Some numerical comparison between QEP and QMP is also presented in that section. Finally, in Section IX, we discuss the features of the QEP and other potential applications of the adaptive Kalman filtering methodology developed in this paper. A summary of the QEP formulae is given in Appendix C.

Parts of the derivation of QEP are heuristic. The heuristic has a sound theoretical foundation under two assumptions: (*i*) the audit sample size for a rating class does not vary in time by orders of magnitude, and (*ii*) the maximum likelihood estimates of the time series parameters fall within their feasible region. These assumptions are satisfied in about 95 percent of the audit examples. QEP appears to work for the other 5 percent as well, but this has not been fully tested.

II. DESCRIPTION OF THE MODEL

Let θ_t denote the true defect index in period t for the particular rating class under study. Thus,

$$\theta_t = \frac{\left(\begin{array}{c}\textit{Total number of defects present in} \\ \textit{the production of period } t\end{array}\right)}{\left(\begin{array}{c}\textit{Total number of defects allowed in} \\ \textit{that production under the quality standard}\end{array}\right)}.$$

In deriving the present QMP, Bruce Hoadley[2] assumed that over a time window of six periods the successive values of θ_t are independently and identically distributed around a fixed mean, called the long-term process mean. Consequently, the time order of the past observations is ignored in estimating the current defect index. Hence, QMP responds to a drift in the defect index only through having the moving window, which means a slow response. In our model we will overcome this deficiency by explicitly allowing for drift and serial correlation.

The mathematical analysis of serially correlated data is greatly simplified when the random variables involved are normally distributed. The audit problem can be put in this framework by the square root transformation described in the following paragraph.

For the chosen sample, let e_t be the expected number of defects under standard quality, x_t be the observed number of defects, and $I_t = x_t/e_t$ be the observed defect index; then x_t has a Poisson distribution with mean $e_t\theta_t$. It is well known that the distribution of $\sqrt{x_t}$ can be approximated by the Gaussian density with mean $\sqrt{e_t\theta_t}$ and variance ¼. Let $Y_t = \sqrt{I_t}$. The distribution of Y_t is approximately normal with mean $\sqrt{\theta_t}$ and variance $0.25/e_t$. We shall denote $\sqrt{\theta_t}$ by ζ_t and refer to it as the transformed true defect index, or simply as the true defect index.

When the observations are defined in terms of demerits or defectives we will take x_t and e_t equal to the observed and the expected equivalent defects, respectively, as defined by Hoadley.[3] In this case the distribution of x_t is approximately Poisson with mean $\theta_t e_t$, so we can still use the square root transform defined in the previous paragraph.

Autoregressive-integrated-moving average models with appropriate trend terms may be used to characterize a wide range of serial correlations and trends. However, since the available data on each product is limited, it is essential that we keep the structure simple, involving only a few parameters. Thus, we propose the following model for the variation of ζ_t:

$$\zeta_t = m_t + \nu_{1t}, \tag{1}$$

where m_t is the trend term (including the mean) and ν_{1t} is the deviation from the trend. The successive values of ν_{1t} will be assumed to be independently distributed with zero mean and variance σ_{1t}^2.

Since the exact nature of the drift is not known to begin with, we shall assume that m_t is a random walk. We found in control engineering literature[4] that the random walk model serves well in tracking a variety of trends in unknown parameters, and therefore we chose to use it in the present problem. Thus,

$$m_t = m_{t-1} + \nu_{2t}, \tag{2}$$

where ν_{2t} is a sequence of independently distributed random variables with mean zero and variance σ_{2t}^2. Further, the sequence ν_{2t} will be assumed to be independent of the sequence ν_{1t}.

Equations (1) and (2) thus characterize the variation of the defect index—the component m_t describes the low-frequency (smooth) changes, while ν_{1t} describes the high-frequency changes in ζ_t. If we take $\sigma_{2t}^2 = 0$ and $\sigma_{1t}^2 = \sigma_1^2 =$ constant, then these equations imply that the ζ_t's and hence θ_t's are independently and identically distributed. Thus, the QMP model is a special case of the general model of this paper.

The transformed observed defect index, Y_t, is the transformed true defect index plus the sampling error, η_t. Thus,

$$Y_t = \zeta_t + \eta_t. \tag{3}$$

As discussed earlier, the expected value of Y_t is ζ_t and the variance of Y_t is $0.25/e_t$, so η_t has zero mean and its variance is equal to $0.25/e_t$. We assume that the successive random variables η_t are independent. Also, since the origins of η_t, ν_{1t} and ν_{2t} are unrelated, we assume that these three series are mutually uncrosscorrelated. Further, the distributions of ν_{1t}, ν_{2t}, and η_t are assumed to be normal. The justification for this assumption comes from the fact that Y_t's are approximately normally distributed.

The problem at hand is to make an inference on θ_n given data up to and including the nth time period. In particular, we wish to determine the posterior probability of the event that θ_n exceeds one.

III. KALMAN FILTER SOLUTION

In the Kalman filter terminology, ζ_n and m_n are the unobserved state variables about which we wish to make inference using the observations $Y_1, \cdots, Y_n$. Let us, for now, assume that the model parameters σ_{1t}^2, σ_{2t}^2, and $\sigma_{\eta t}^2$ are known for $t = 1, \cdots n$ and that the means and the variances of ζ_0 and m_0 are known. Then the Kalman filter provides recursive formulae for estimating the posterior means and variances of ζ_n and m_n. The derivation of the general Kalman filter may be found in a number of books (e.g., see Jazwinsky[5] or Gelb[4]). A simple derivation for the special case of the audit model is given in Appendix A. The desired recursive formulae are given below.

Conditional on the data up to time t, the distribution of m_t is normal with mean $\hat{m}_t$ and variance q_t,

$$\text{i.e.,} \quad m_t | t \sim N(\hat{m}_t, q_t),$$

$$\text{where} \quad \hat{m}_t = \omega_{2t}\hat{m}_{t-1} + (1 - \omega_{2t})Y_t \tag{4}$$

$$q_t = (1 - \omega_{2t})(\sigma_{1t}^2 + \sigma_{\eta t}^2) \tag{5}$$

$$\omega_{2t} = \frac{\sigma_{1t}^2 + \sigma_{\eta t}^2}{\sigma_{1t}^2 + \sigma_{\eta t}^2 + \sigma_{2t}^2 + q_{t-1}}. \tag{6}$$

Likewise, conditional on the data up to time t, the distribution of ζ_t is normal with mean $\hat{\zeta}_t$ and variance p_t,

$$\text{i.e.,} \quad \zeta_t \sim N(\hat{\zeta}_t, p_t),$$

$$\text{where} \quad \hat{\zeta}_t = \omega_{2t}\omega_{1t}\hat{m}_{t-1} + (1 - \omega_{2t}\omega_{1t})Y_t \tag{7}$$

$$p_t = (1 - \omega_{2t}\omega_{1t})\sigma_{\eta t}^2 \tag{8}$$

$$\omega_{2t}\omega_{1t} = \frac{\sigma_{\eta t}^2}{\sigma_{\eta t}^2 + \sigma_{1t}^2 + \sigma_{2t}^2 + q_{t-1}}. \tag{9}$$

To use these recursive equations the starting values m_0 and q_0 must be specified. The choice of these values is discussed in Section VII. For now, we note that as $t \to \infty$, the effect of the starting values reduces to zero.

Notice that eq. (4) is an adaptive, exponential smoothing equation. The smoothing constant, ω_{2t}, is a function of time and is determined by the relative values of the different variances as given by eq. (6). Observe that $V(Y_t | m_t) = \sigma_{1t}^2 + \sigma_{\eta t}^2$ so that $\sigma_{1t}^2 + \sigma_{\eta t}^2$ measures the uncertainty in using Y_t for estimating m_t; also, $\sigma_{2t}^2 = V(m_t | m_{t-1})$ and $q_{t-1} = V(m_{t-1} | t - 1)$. Thus, $\sigma_{2t}^2 + q_{t-1}$ is a measure of uncertainty in

using $\hat{m}_{t-1}$ for estimating m_t. It is clear from eqs. (4) and (6) that the weights given to Y_t and $\hat{m}_{t-1}$ are inversely proportional to their respective uncertainties in estimating m_t.

Equations (7) and (9) are the analogous equations for computing the posterior mean of ζ_t. Note that $V(Y_t|\zeta_t) = \sigma_{\eta t}^2$ is the uncertainty in using Y_t to estimate ζ_t. Further, $V(\zeta_t|m_{t-1}) = \sigma_{1t}^2 + \sigma_{2t}^2$ and $V(m_{t-1}|t-1) = q_{t-1}$ so that $\sigma_{1t}^2 + \sigma_{2t}^2 + q_{t-1}$ is the uncertainty in using $\hat{m}_{t-1}$ to estimate ζ_t. The weights on Y_t and $\hat{m}_{t-1}$ are thus seen to be inversely proportional to the respective uncertainties.

From eq. (8) we note that the posterior variance of ζ_t conditional on data up to time t is smaller than $\sigma_{\eta t}^2$ by the factor $(1 - \omega_{2t}\omega_{1t})$. Thus, the factor $(1 - \omega_{2t}\omega_{1t})$ represents the advantage of filtering in estimating ζ_t. Similarly, from eq. (5) we see that the factor $(1 - \omega_{2t})$ is the benefit of filtering in estimating m_t.

To compare the QMP model given in Ref. 2 with the QEP model we shall rewrite eqs. (7) and (9) as follows:

$$\hat{\zeta}_t = \omega_{1t}\hat{m}_t + (1 - \omega_{1t})Y_t$$

$$\omega_{1t} = \frac{\sigma_{\eta t}^2}{\sigma_{\eta t}^2 + \sigma_{1t}^2}.$$

Analogous to the QMP, eq. (7) expresses $\hat{\zeta}_t$ as a weighted sum of $\hat{m}_t$, the estimated current mean level, and Y_t, the current observation. The weight ω_{1t} is analogous to the shrinkage constant given in Ref. 2, and we will also call it a shrinkage constant.

The discussion of this section was based on the assumption that σ_{1t}^2, σ_{2t}^2, and $\sigma_{\eta t}^2$ are known quantities. However, in the audit problem σ_{1t}^2 and σ_{2t}^2 are not known and must be estimated from the observed data. In the following section we will derive the estimates of σ_{1t}^2 and σ_{2t}^2, and in Section V we will modify eqs. (4) through (9) to accommodate the fact that σ_{1t}^2 and σ_{2t}^2 are estimated.

IV. ESTIMATION OF THE MODEL PARAMETERS

Consider the case where $\sigma_{1t}^2 = \sigma_1^2 =$ constant, $\sigma_{2t}^2 = \sigma_2^2 =$ constant, and $\sigma_{\eta t}^2 = \sigma_\eta^2 =$ constant.

Let us define $Z_t = Y_t - Y_{t-1}$. Under the assumed model $E(Z_t) = 0$ and the autocovariances of Z_t are given by

$$E(Z_t^2) = 2\sigma_1^2 + \sigma_2^2 + 2\sigma_\eta^2, \tag{10}$$

$$E(Z_t Z_{t-1}) = -\sigma_1^2 - \sigma_\eta^2, \tag{11}$$

and

$$E(Z_t Z_{t-l}) = 0, \qquad l \geq 2. \tag{12}$$

Thus, Z_t is a first-order moving average [MA(1)] process of zero mean, i.e., Z_t can be represented as

$$Z_t = a_t + \beta a_{t-1}, \tag{13}$$

where a_t is a white noise series of variance σ^2. The parameters β and σ^2 are related to σ_1^2, σ_2^2, and σ_η^2 through the autocovariance function; i.e.,

$$E(Z_t^2) = (1 + \beta^2)\sigma^2 = 2\sigma_1^2 + \sigma_2^2 + 2\sigma_\eta^2, \tag{14}$$

and

$$E(Z_t Z_{t-1}) = \beta\sigma^2 = -\sigma_1^2 - \sigma_\eta^2. \tag{15}$$

Solving eqs. (14) and (15) we have

$$\sigma_1^2 = -\beta\sigma^2 - \sigma_\eta^2 \tag{16}$$

and

$$\sigma_2^2 = (1 + \beta)^2\sigma^2. \tag{17}$$

The nonnegativity of σ_1^2 and the invertibility of the model given by eq. (13) impose the following restriction on β: $-1 < \beta \leq -\sigma_\eta^2/\sigma^2$. Thus, the feasible region for β and σ^2 is the one enclosed by the lines: $\beta = -1$, $\beta\sigma^2 = -\sigma_\eta^2$, and $\sigma^2 = \infty$. The region is shown in Fig. 1.

The parameters β and σ^2 can be estimated using a suitable time-series method. Once β and σ^2 are known, eqs. (16) and (17) may be used to estimate σ_1^2 and σ_2^2.

Contrary to the assumption made at the beginning of this section, σ_{1t}^2, σ_{2t}^2, and $\sigma_{\eta t}^2$ may in fact vary from period to period. Through eqs. (14) and (15) this implies that β and σ^2 may vary with time. In other words, Z_t is like a $MA(1)$ process with changing parameters. Therefore,

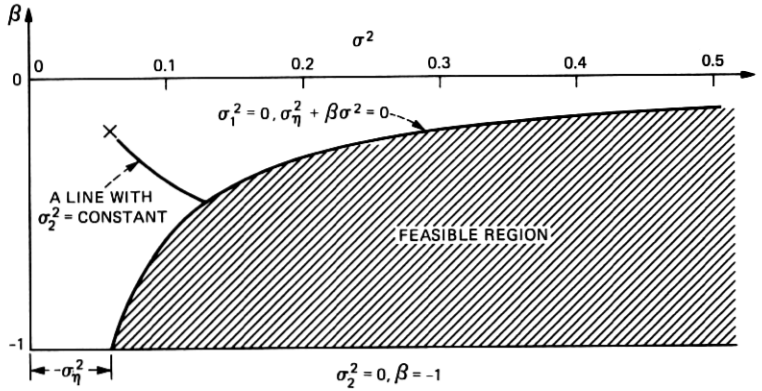

Fig. 1—Feasible region for β and σ^2.

we need an adaptive, recursive estimation method for estimating β and σ^2, rather than the usual time series estimation methods. The recursive method of Phadke[6] will therefore be used here. The method discounts the past data exponentially and thus can respond to changes in the model parameters. The necessary recursion formulae are given below. Appendix B gives the derivation of these formulae.

$$\hat{\beta}_t = \beta_0 - R_t^{-1}\nu_t \tag{18}$$

$$\hat{\sigma}_t^2 = S_t(\hat{\beta}_t)/A_t, \tag{19}$$

where

$$\nu_t = \lambda\nu_{t-1} + 2a_t\frac{da_t}{d\beta} \tag{20}$$

$$R_t = \lambda R_{t-1} + 2\left(\frac{da_t}{d\beta}\right)^2 \tag{21}$$

$$a_t = Z_t - \beta_0 a_{t-1} \tag{22}$$

$$\frac{da_t}{d\beta} = -a_{t-1} - \beta_0\frac{da_{t-1}}{d\beta} \tag{23}$$

$$S_t(\beta_0) = \lambda S_{t-1}(\beta_0) + a_t^2 \tag{24}$$

$$S_t(\hat{\beta}_t) = S_t(\beta_0) + (\hat{\beta}_t - \beta_0)\nu_t + \tfrac{1}{2}(\hat{\beta}_t - \beta_0)^2 R_t \tag{25}$$

$$A_t = \lambda A_{t-1} + 1. \tag{26}$$

The choice of the starting values for these recursions will be studied in Section VII. The parameter λ, $0 < \lambda \leq 1$, determines how fast the old data is discounted in estimating the model parameters. $\lambda = 1$ implies that the entire past data is used. The smaller the λ the faster the past data is discounted.

The estimates $\hat{\sigma}_t^2$ and $\hat{\beta}_t$ are uncorrelated and have the following approximate variances:

$$V(\hat{\beta}_t) \simeq 2\hat{\sigma}_t^2/R_t \tag{27}$$

$$V(\hat{\sigma}_t^2) \simeq 2\hat{\sigma}_t^4/A_t. \tag{28}$$

The estimated values of σ^2 and β may be substituted in eqs. (16) and (17) to compute σ_{1t}^2 and σ_{2t}^2. When $\sigma_{\eta t}^2$ varies with time, we should use the exponentially smoothed value, $\overline{\sigma_{\eta,t}^2}$, as defined below:

$$\overline{\sigma_{\eta,t}^2} = \lambda\overline{\sigma_{\eta,t-1}^2} + (1-\lambda)\sigma_{\eta t}^2, \tag{29}$$

in place of σ_η^2 in eq. (16). Thus,

$$\hat{\sigma}_{1t}^2 = -\hat{\beta}_t\hat{\sigma}_t^2 - \overline{\sigma_{\eta,t}^2},$$

and

$$\hat{\sigma}_{2t}^2 = (1 + \hat{\beta}_t)^2 \hat{\sigma}_t^2.$$

In the rare case of extreme variations (order of magnitude variations) in $\sigma_{\eta t}^2$, eq. (29) has not been fully tested, so we urge caution for such cases.

It is possible that eqs. (18) and (19) would yield infeasible values of $\hat{\beta}_t$ and $\hat{\sigma}_t^2$. In that case we propose the following truncation rules, which take us close to the maximum point on the feasible boundary.

Step 1: Truncate $\hat{\beta}_t$ to the region $[-1, 0]$. Denote the truncated value by β_t^*, where

$$\beta_t^* = \begin{cases} \beta_0 - \nu_t/R_t & \text{if } -1 \le \beta_0 - \nu_t/R_t \le 0 \\ 0 & \text{if } \beta_0 - \nu_t/R_t > 0 \\ -1 & \text{if } \beta_0 - \nu_t/R_t < -1 \end{cases}.$$

Step 2: Compute σ_t^{*2} and $\hat{\sigma}_{2t}^2$:

$$\sigma_t^{*2} = S_t(\beta_t^*)/A_t = \{S_t(\beta_0) + (\beta_t^* - \beta_0)\nu_t + \tfrac{1}{2}(\beta_t^* - \beta_0)^2 R_t\}/A_t$$

$$\hat{\sigma}_{2t}^2 = (1 + \beta_t^*)^2 \sigma_t^{*2}.$$

Step 3: If β_t^* and σ_t^{*2} belong to the feasible region, i.e., if $(-\beta_t^* \sigma_t^{*2} - \overline{\sigma_{\eta t}^2}) \ge 0$ then $\hat{\beta}_t = \beta_t^*$, $\hat{\sigma}_t^2 = \sigma_t^{*2}$, and $\hat{\sigma}_{1t}^2 = -\hat{\beta}_t \hat{\sigma}_t^2 - \overline{\sigma_{\eta,t}^2}$.

Step 4: If β_t^* and σ_t^{*2} do not belong to the feasible region, i.e., if $(-\beta_t^* \sigma_t^{*2} - \overline{\sigma_{\eta,t}^2}) < 0$, then set $\hat{\sigma}_{1t}^2 = 0$. Compute $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ by solving the following two equations:

$$\hat{\sigma}_{1t}^2 = -\hat{\beta}_t \hat{\sigma}_t^2 - \overline{\sigma_{\eta t}^2} = 0,$$

and

$$\hat{\sigma}_{2t}^2 = (1 + \hat{\beta}_t)^2 \hat{\sigma}_t^2.$$

The resulting $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ are given by

$$\hat{\beta}_t = \frac{-(2 + \hat{\sigma}_{2t}^2/\overline{\sigma_{\eta,t}^2}) + \sqrt{(2 + \hat{\sigma}_{2t}^2/\overline{\sigma_{\eta,t}^2})^2 - 4}}{2}$$

and

$$\hat{\sigma}_t^2 = -\overline{\sigma_{\eta,t}^2}/\hat{\beta}_t.$$

Note that when these truncations are applied there will be a larger degree of approximation involved in using eqs. (27) and (28) for computing the variances of $\hat{\beta}_t$ and $\hat{\sigma}_t^2$. However, these variances enter only into the secondary terms of the adaptive Kalman filter to be derived in the next section. Consequently, we may ignore the effect of truncation.

V. ADAPTIVE KALMAN FILTER

In deriving the Kalman filter solution of Section III, it was assumed that σ_{1t}^2 and σ_{2t}^2 are known quantities. Since in the audit problem these parameters are estimated using the observed data, we need to make due modifications to the Kalman filter solution. We will refer to the resulting formulae as the adaptive Kalman filter.

Consider the distribution of m_t conditional on data up to time t:

$$\hat{m}_t = E(m_t|t) = EE(m_t|t, \sigma_{1t}^2, \sigma_{2t}^2) = E(\omega_{2t}\hat{m}_{t-1} + (1 - \omega_{2t})Y_t);$$

hence,

$$\hat{m}_t \simeq \hat{\omega}_{2t}\hat{m}_{t-1} + (1 - \hat{\omega}_{2t})Y_t \tag{30}$$

where

$$\hat{\omega}_{2t} = \frac{\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2}{\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2 + \hat{\sigma}_{2t}^2 + q_{t-1}}. \tag{31}$$

The distribution of ω_{2t} conditional on data up to time t is very complicated. So the expected value of ω_{2t} cannot be simply computed. Therefore, in eq. (30) we have approximated $E(\omega_{2t}|t)$ by $\hat{\omega}_{2t}$, the maximum posterior density point. This approximation would be very good when $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ lie inside the feasible region. However, if $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ lie on the boundary of the feasible region, $\hat{\omega}_{2t}$ will be a less accurate approximation of $E(\omega_{2t}|t)$. The extent of the inaccuracy will depend on the values of R_t and A_t. For large R_t and A_t the likelihood of β_t and σ_t^2 will drop very rapidly as one goes away from the feasible boundary nearest to the maximum likelihood point. Thus, the inaccuracy would be smaller for larger values of R_t and A_t.

Now consider the variance of m_t given data up to time t:

$$\begin{aligned} q_t = V(m_t|t) &= EV(m_t|t, \sigma_{1t}^2, \sigma_{2t}^2) + VE(m_t|t, \sigma_{1t}^2, \sigma_{2t}^2) \\ &= E[(1 - \omega_{2t})(\sigma_{1t}^2 + \sigma_{\eta t}^2)] + V[\omega_{2t}\hat{m}_{t-1} + (1 - \omega_{2t})Y_t]; \end{aligned}$$

hence,

$$q_t \simeq (1 - \hat{\omega}_{2t})(\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2) + (Y_t - \hat{m}_{t-1})^2 V(\omega_{2t}). \tag{32}$$

The effect of $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ lying on the feasible boundary will be to introduce an inaccuracy in the term $(1 - \hat{\omega}_{2t})(\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2)$ of eq. (32), as discussed above. Knowing the variance of β_t and σ_t^2, the variance of ω_{2t} can be derived via the Taylor series approximation as follows. We have

$$\begin{aligned} \omega_{2t} &= \frac{\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2}{\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2 + \hat{\sigma}_{2t}^2 + q_{t-1}} \\ &= \frac{-\beta_t\sigma_t^2 - \overline{\sigma_{\eta t}^2} + \sigma_{\eta t}^2}{(1 + \beta_t + \beta_t^2)\sigma_t^2 + q_{t-1} + \sigma_{\eta t}^2 - \overline{\sigma_{\eta t}^2}} \end{aligned}$$

$$\simeq \hat{\omega}_{2t} + \left(\frac{\partial \omega_{2t}}{\partial \beta_t}\right)(\beta_t - \hat{\beta}_t) + \left(\frac{\partial \omega_{2t}}{\partial \sigma_t^2}\right)(\sigma_t^2 - \hat{\sigma}_t^2).$$

Hence, to the first-order approximation the variance of ω_{2t} is

$$\begin{aligned} V(\omega_{2t}) &\simeq \left(\frac{\partial \omega_{2t}}{\partial \beta}\right)^2 V(\hat{\beta}_t) + \left(\frac{\partial \omega_{2t}}{\partial \sigma^2}\right)^2 V(\hat{\sigma}_t^2) \\ &= \{2\hat{\sigma}_t^6[1 + \hat{\omega}_{2t}(1 + 2\hat{\beta}_t)]^2/R_t \\ &\quad + 2\hat{\sigma}_t^4[\hat{\beta}_t + (1 + \hat{\beta}_t + \hat{\beta}_t^2)\hat{\omega}_{2t}]^2/A_t\} \\ &\quad \div (\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2 + \hat{\sigma}_{2t}^2 + q_{t-1})^2. \end{aligned} \tag{33}$$

Note that the smoothing constant ω_{2t} is restricted to the interval [0, 1]. The most noninformative distribution on this interval is the uniform distribution whose variance is 1/12. The $\hat{\omega}_{2t}$ computed by eq. (31) clearly adheres to the interval [0, 1]. However, because of the approximations involved, the computed $V(\omega_{2t})$ may come out larger than 1/12. In that case we propose to truncate it to 1/12.

When $\hat{\beta}_t$ and $\hat{\sigma}_{1t}^2$ lie on the boundary of their feasible region, the use of the Taylor series approximation would yield inaccurate estimates of $V(\omega_{2t})$. Since the contribution of this variance is secondary in computing $V(m_t|t)$, we may ignore the effect of truncation.

We can proceed in an analogous way to compute $E(\zeta_t|t) = \hat{\zeta}_t$ and $V(\zeta_t|t) = p_t$ to yield:

$$\hat{\zeta}_t = \hat{\omega}_{2t}\hat{\omega}_{1t}\hat{m}_{t-1} + (1 - \hat{\omega}_{2t}\hat{\omega}_{1t})Y_t, \tag{34}$$

where

$$\hat{\omega}_{1t} = \frac{\sigma_{\eta t}^2}{\sigma_{\eta t}^2 + \hat{\sigma}_{1t}^2} \tag{35}$$

and

$$p_t = (1 - \hat{\omega}_{2t}\hat{\omega}_{1t})\sigma_{\eta t}^2 + (Y_t - \hat{m}_{t-1})^2 V(\omega_{2t}\omega_{1t}), \tag{36}$$

where

$$\begin{aligned} V(\omega_{2t}\omega_{1t}) &\simeq [2\hat{\sigma}_t^6(1 + 2\hat{\beta}_t)^2\hat{\omega}_{1t}^2\hat{\omega}_{2t}^2/R_t \\ &\quad + 2\hat{\sigma}_t^4(1 + \hat{\beta}_t + \hat{\beta}_t^2)\hat{\omega}_{1t}^2\hat{\omega}_{2t}^2/A_t] \\ &\quad \div (\hat{\sigma}_{1t}^2 + \sigma_{\eta t}^2 + \hat{\sigma}_{2t}^2 + q_{t-1})^2. \end{aligned} \tag{37}$$

For the reasons discussed in the case of $V(\omega_{2t})$, if eq. (37) yields a value of $V(\omega_{2t}\omega_{1t}) > 1/12$, then we will truncate it to 1/12.

The effects of $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ lying on the feasible boundary are similar to those explained in connection with $\hat{m}_t$ and q_t.

VI. BOX CHART AND THE EXCEPTION REPORT

In Ref. 2 Bruce Hoadley has proposed a format for displaying the

conditional distribution of θ_t given data up to time t. He has also proposed exception rules in terms of this distribution. We shall use the same reporting format and the exception rules.

The conditional distribution of θ_t will be summarized by a box chart that shows the 99, 95, 5, and 1 percentiles of the distribution, the best estimate of θ_t, denoted by $\hat{\theta}_t$, the mean level, $\hat{M}_t$, and the current defect index I_t. By applying the inverse square root transformation, we have

$$\hat{M}_t = \hat{m}_t^2 \quad \text{and} \quad \hat{\theta}_t = \hat{\xi}_t^2.$$

The quantiles of θ_t are once again obtained by squaring the quantiles of ξ_t. Since ξ_t is restricted to be positive, and we have approximated its density by the normal distribution, we may have to truncate some of the extreme quantiles to zero. If we take this fact into account, the desired quantiles of θ_t are:

$$Q_1 = 99\% \text{ quantile} = [\max(\hat{\xi}_t - 2{\cdot}326\sqrt{p_t}, 0)]^2$$

$$Q_2 = 95\% \text{ quantile} = [\max(\hat{\xi}_t - 1{\cdot}645\sqrt{p_t}, 0)]^2$$

$$Q_3 = 5\% \text{ quantile} = (\hat{\xi}_t + 1{\cdot}645\sqrt{p_t})^2$$

$$Q_4 = 1\% \text{ quantile} = (\hat{\xi}_t + 2{\cdot}326\sqrt{p_t})^2.$$

A sample box chart is shown in Fig. 2.

The exception rules are:

(*i*) Below Normal: A rating class will be declared below normal if the posterior probability of θ_t being larger than one exceeds 0.99.

(*ii*) Alert: An alert will be declared for a rating class if the posterior probability of θ_t being greater than one exceeds 0.95 but it is less than or equal to 0.99.

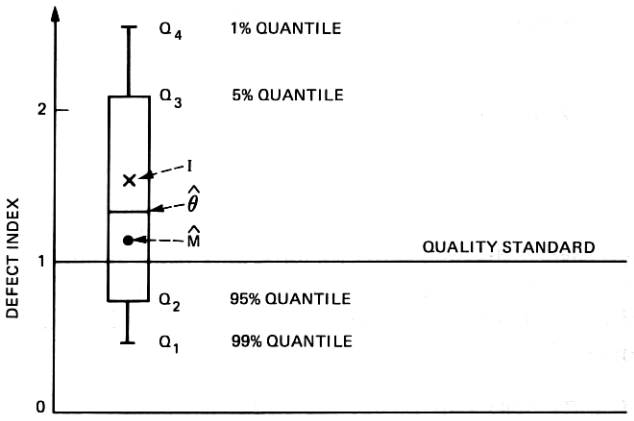

Fig. 2—Sample box chart of the conditional distribution of θ_t.

So in terms of the quantiles derived above, we will declare below normal if $Q_1 > 1$ and alert if $Q_1 \leq 1$ but $Q_2 > 1$.

VII. CHOICE OF STARTING VALUES

The Kalman filter solution described in Section III and the estimation of model parameters described in Section IV are both recursive procedures that must be appropriately initialized. Note that since each of these procedures discounts the past data, the effect of initialization diminishes to zero as more data is accumulated on any rating class. So any biases introduced by the initialization process are transient and temporary. The best way to choose the initial values is by analyzing the historical data on all rating classes. Pending such an analysis, we shall tentatively choose the initial parameter values, as follows.

We will take $\hat{m}_0 = 1.0$ and $\hat{q}_0 = 0.134$. This amounts to choosing a very diffused prior distribution on the mean level. On the square-root defect-index scale the lower and the upper one percentiles of this distribution are 0.149 and 1.851, respectively; while on the defect-index scale the lower and the upper one percentiles are 0.022 and 3.428, respectively. The mean and the median of this distribution are equal to one on either scale. Consistent with this we will choose $Y_0 = 1.0$.

The parameter $\overline{\sigma^2_{\eta,0}}$ should be taken equal to the variance of the transformed defect index associated with the planned equivalent expectancy for a period's sample for the particular rating class. In the present analysis we will take $e_0 = e_1$ and $\overline{\sigma^2_{\eta,0}} = \sigma^2_{\eta,0} = 0.25/e_0$.

The parameter λ determines how many periods of data are effectively used in estimating the time series parameters β_t and σ^2_t and, hence, the parameters σ^2_{1t} and σ^2_{2t}. We will take $\lambda = 0.95$, which implies that effectively $1/(1 - \lambda) = 20$ periods of data are used in estimating the model parameters.

We also need to specify the values of β_0, a_0, $da_0/d\beta$, $S_0(\beta_0)$, e_0, R_0, and A_0. All these variables enter into the recursive maximum likelihood estimation of β and σ^2. We shall take $\beta_0 = -0.6$, which is an approximate midpoint of the feasible range of β. The quantities a_0 and $da_0/d\beta$ will be taken equal to their respective expected values, namely, zero in each case; and A_0 will be set equal to its steady-state value, namely, $1/(1 - \lambda)$. We will take $\nu_0 = 0$, $S_0(\beta_0) = 0.625/[e_0(1 - \lambda)]$, and $R_0 = 20.0/e_0$.

The above starting values imply that at $t = 0$ the mean and the variance of β are respectively -0.6 and 0.063. The variance of the uniform distribution on the $(-1, 0)$ interval is $1/12 = 0.083$. Since the feasible interval for β is smaller than $(-1, 0)$, the variance of 0.063 represents a fairly diffused initialization.

Also, the above starting values imply that the mean and the variance of σ^2 at $t = 0$ are 2.5 $\sigma^2_{\eta 0}$ and $0.625(\sigma^2_{\eta 0})^2$. Therefore, by the gamma

density assumption, the 95-percent confidence interval on σ^2 is ($1.2\ \sigma_{\eta 0}^2$, $4.28\ \sigma_{\eta 0}^2$), which is a very wide interval.

The values of σ_1^2 and σ_2^2 implied by the above starting values are $0.5\ \sigma_{\eta 0}^2$ and $0.4\ \sigma_{\eta 0}^2$, respectively.

VIII. ILLUSTRATIVE EXAMPLES

To illustrate the properties of the quality evaluation plan we shall now present six examples. The first three examples are the simulated examples, while the latter three use real audit data.

Example 1: Figure 3a shows the response of QEP to a sudden shift in the quality level. For the first ten periods the observed defect index fluctuates randomly around the fixed level 3.0. From the eleventh to the twentieth period the observed defect index is fixed at 1.0. In each period the expectancy at standard is 5.0. Notice that starting with the eleventh period the estimated mean level rapidly approaches the new mean level. Also, starting with the eleventh observation the product gets off the exception report. Figure 3b shows the corresponding results for QMP. It is clear that in terms of both $\hat{M}_t$ and $\hat{\theta}_t$ the response of QEP is quicker than the response of QMP.

Example 2: Figure 4a displays the response of QEP to a linear trend in the quality level. As in the case of Example 1, for the first ten periods the observed defect index randomly fluctuates around the fixed level 3.0. From the eleventh to the twentieth period the observed defect index has a linear downtrend, as plotted. In all twenty periods the expectancy at standard is 5.0. Notice that both $\hat{M}_t$ and $\hat{\theta}_t$ follow the trend with a small lag. Also note that the QEP algorithm recognizes that the process has a drift rather than random fluctuation. Consequently, $\hat{M}_t$ and $\hat{\theta}_t$ are very close while following the drift. Figure 4b shows the results of QMP for the same data. Here again, $\hat{M}_t$ and $\hat{\theta}_t$ follow the trend, but the lag is much larger. This is manifested in the fact that QEP gets the product off below normal in the seventeenth period while with QMP that happens a period later.

Example 3: This example illustrates that QEP and QMP have similar behaviors when the defect index fluctuates about a fixed value for a long period of time. Figure 5a shows the results of QEP when the defect index fluctuates around the fixed level of 2.0, while Fig. 5b shows the results of QMP with the same data. Note that both the methods declare below normals and alerts in the same periods.

Example 4: Figure 6a gives the data for repaired remreed grids, rating class OC038TT, for periods 7801 through 7904. The periods are numbered 1 through 12 in the figure. The QEP results are also shown in the figure. Similar results with QMP are plotted in Fig. 6b. In response to the drift in the quality we see that QEP attaches a heavier weight to the current data. Consequently, with QEP the mean level, $\hat{M}_t$

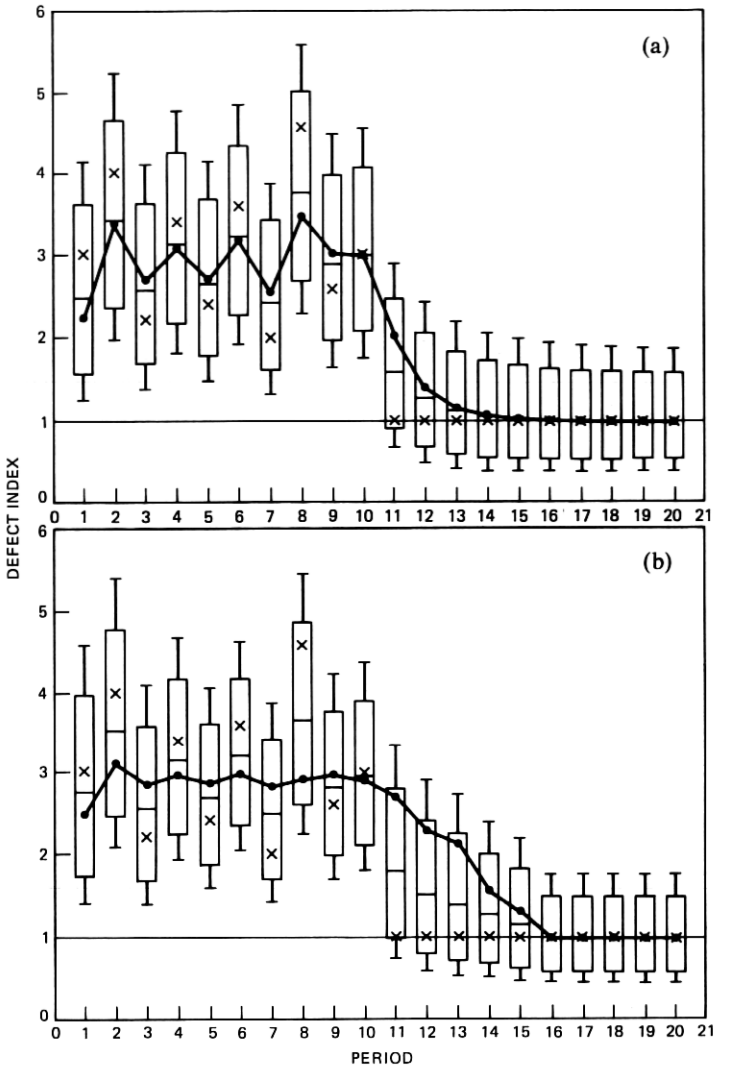

Fig. 3—Response to a sudden shift in the quality level for: (a) QEP, and (b) QMP.

follows the drift more closely than with QMP. In the period 5, recognizing the drift, QEP takes the product off the exception report while QMP still calls it an alert. Also period 7 is an alert according to QMP while, according to QEP, it is off the exception report. These differences between QEP and QMP are clearly seen to be the result of the fact that QEP exponentially discounts the past data, while QMP considers every observation in the six-period window to be equally important.

Example 5: Figures 7a and 7b give the results of QEP and QMP,

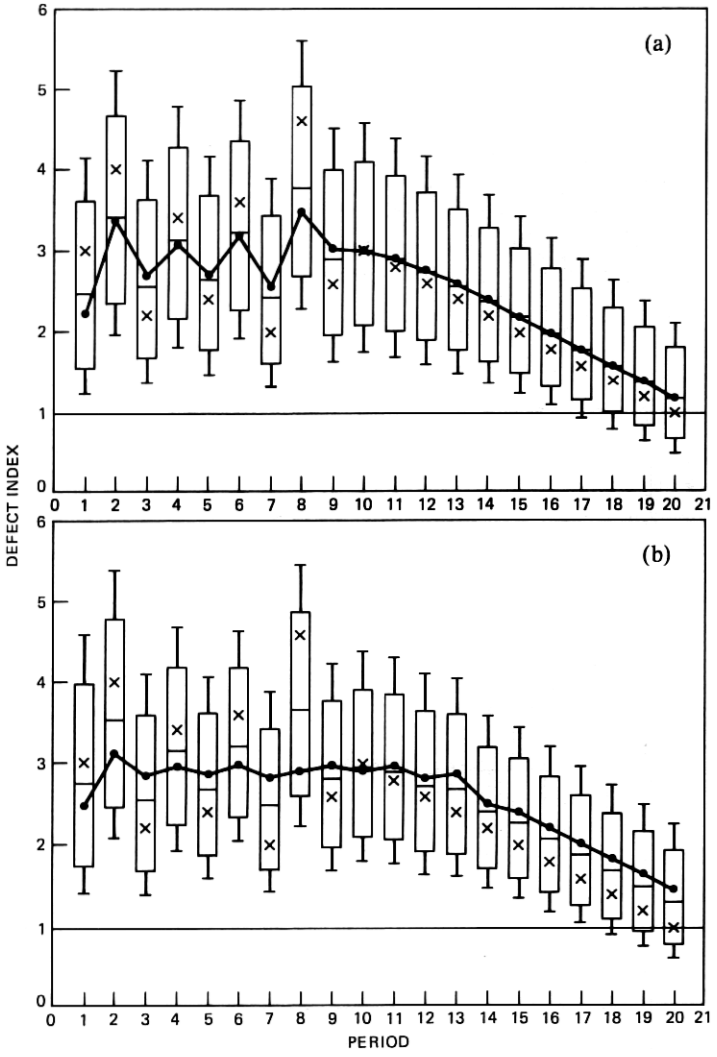

Fig. 4—Response to a linear trend in the quality level for: (a) QEP, and (b) QMP.

respectively, for the rating class OH060CM, consisting of modular telephone chords. The periods covered by the chart are 7701-7808. As we saw in Example 4, QEP follows the drift more closely. In terms of the exception report, there are several differences. In periods 8, 15, and 16 QMP declares below normal, while QEP calls it only an alert. In period 10 QMP calls it an alert while QEP does not declare any exception. These differences are once again a result of the fact that QEP recognizes the drift and hence heavily discounts the past data.

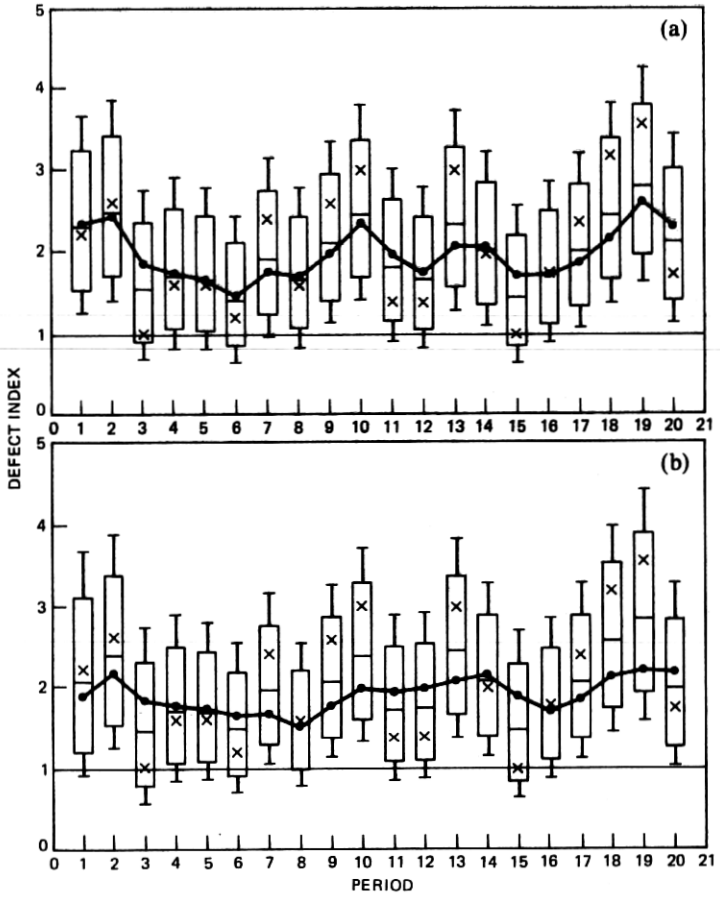

Fig. 5—Response to a random fluctuation in the quality level for: (a) QEP, and (b) QMP.

Example 6: The last example to be considered is the rating class MV104MJ. The results of both the methods are shown in Fig. 8. In this example the quality fluctuates more or less randomly about a fixed mean and, as expected, the two methods give comparable results.

The average values of the weights ω_{1t} and ω_{2t}, and the equivalent expectancies e_t for the three audit examples are tabulated in Table I. Notice that average value of ω_{2t} for OC038TT and OH060CM is 0.48 compared with 0.55 for MV104MJ. This is a direct consequence of the fact that MV104MJ does not exhibit a drift while the others do. The high-frequency fluctuation about the mean function M_t is depicted by ω_{1t}. Relative to the sampling variance $(0.25/e_t)$ OC038TT exhibits a smaller fluctuation than OH060CM. This concurs with the average values of ω_{1t} for these rating classes.

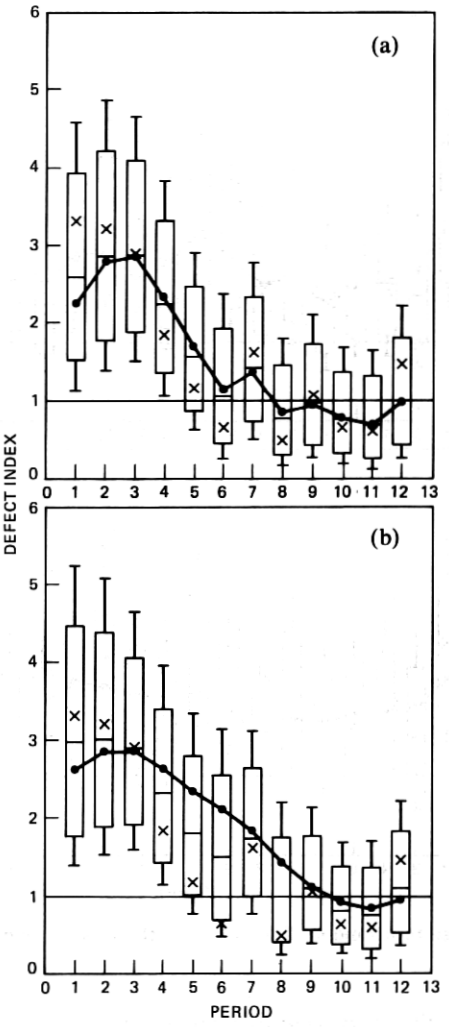

Fig. 6—Results for the rating class OC038TT, periods 7801 through 7904 for (a) QEP, and (b) QMP.

Through the preceding examples it is quite apparent that QEP and QMP could give somewhat different results. Now the key question is: Which method yields a more precise estimate of the unobserved "true defect index"? The only way to answer this question decisively is to take a 100-percent sample of a number of rating classes to find out the true defect indices and compare them with the QEP and QMP results. This is obviously an impossible task. A feasible way to answer the question is by using the models to predict one step ahead and compare the mean-squared prediction errors. Note that $\hat{M}_{t-1}$ is a predictor of I_t.

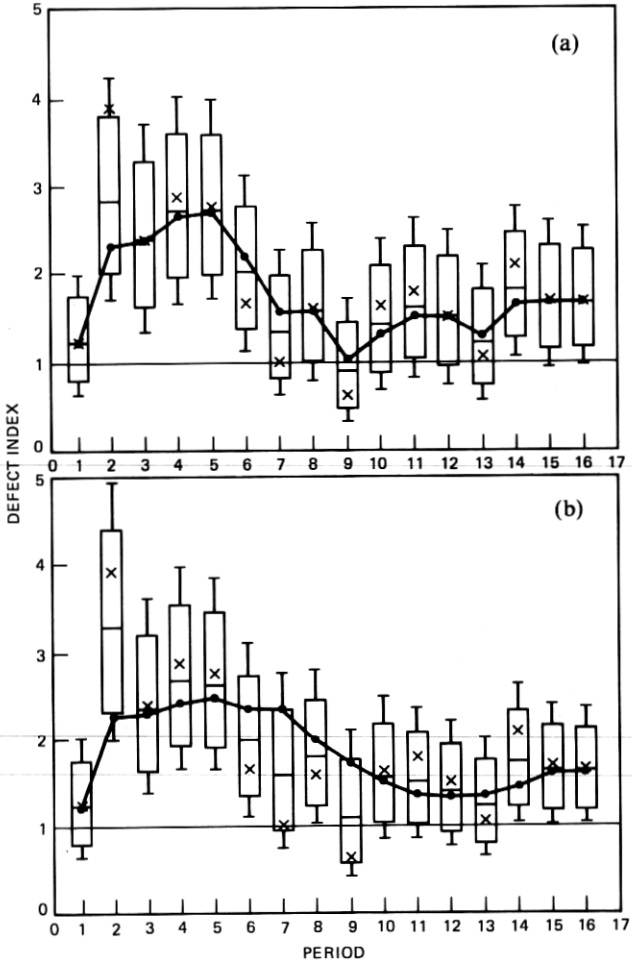

Fig. 7—Results for the rating class OH060CM, periods 7701 through 7808 for: (a) QEP, and (b) QMP.

The mean-squared errors for the three audit data examples, viz. Examples 4, 5, and 6, are given in Table II. For the rating class OC038TT we notice that the mean-squared prediction error (m.s.p.e.) of QMP is 33 percent larger than that of QEP; for OH060CM the m.s.p.e. of QMP is 11 percent larger, and for MV104MJ the m.s.p.e. of QMP is only 3 percent larger. Thus, whenever there is a drift in the quality we may expect QEP to perform better than QMP, whereas if the quality fluctuates randomly around a fixed mean, both QMP and QEP would give similar results.

Effect of truncation: In addition to the numerical examples cited

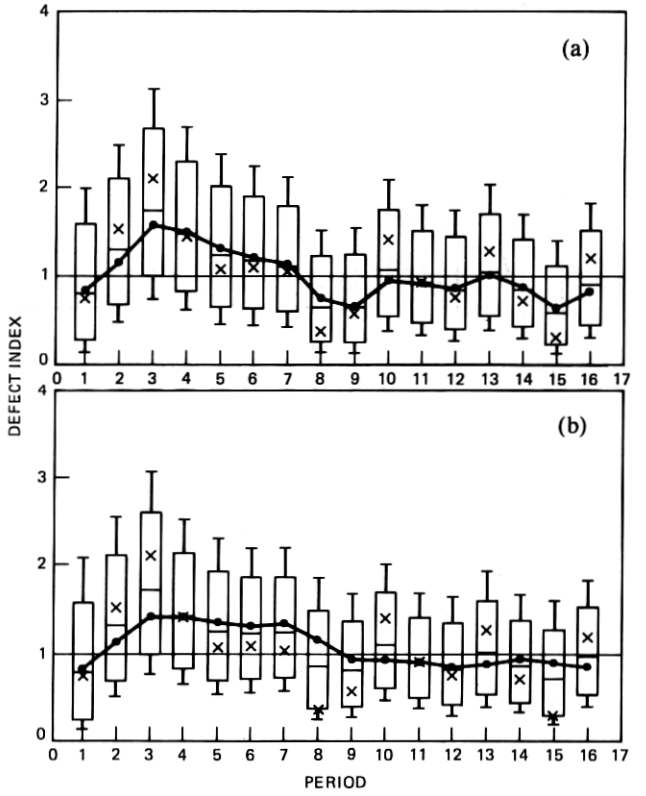

Fig. 8—Results for the rating class MV104MJ, periods 7701 through 7808 for: (a) QEP, and (b) QMP.

above, a limited numerical study was made with thirteen representative rating classes. Each rating class had about fourteen periods of data. This represents a total of 182 test periods. Among these examples, truncation occurred in only 7 percent of the periods. Except for one case, all the truncations caused $\hat{\sigma}_{1t}^2 = 0$. These cases of truncation could be recognized broadly as situations where the variance of the observed defect indices was much smaller than that for the Poisson distribution. In each case of truncation the confidence intervals computed by QEP looked reasonable and comparable to those obtained by QMP, so we can tentatively conclude that the effect of truncation is negligible. Of course, an extensive trial of QEP may suggest some modifications to the truncation rules.

One such modification may be to view the likelihood function of β_t and σ_t^2 as the posterior-probability density function. Then the Bayes estimates of β_t and σ_t^2 may be used in place of the maximum likelihood estimates used in this paper. Because of the complexity of the feasible

Table I—Computed QEP weights for the examples

Rating Class	Average Value of ω_{1t}	ω_{2t}	e_t
OC038TT	0.83	0.48	3.7
OH060CM	0.67	0.48	7.9
MV104MJ	0.75	0.55	4.6

Table II—Comparison of the mean-squared prediction error

Rating Class	Mean Squared Prediction Error QEP	QMP
OC038TT	0.69	0.92
OH060CM	0.94	1.04
MV104MJ	0.29	0.30

region, computing Bayes estimates would involve extensive numerical effort, which may be unnecessary.

IX. DISCUSSIONS

In summary, the QEP model consists of two parts—the system model and the observation model. The system model states that the transformed true-defect index is equal to the process mean that follows the random walk model plus process fluctuation, which is statistically independent from period to period. The random walk model takes care of the process drift. The observation model states that the transformed observed defect index is equal to the transformed true-defect index plus sampling error with a known variance. The different parameters of the QEP model are estimated from the observed data by the recursive, exponentially discounted, maximum likelihood method. The successive transformed true defect indices and the process mean levels are then estimated by the adaptive Kalman filtering algorithm.

From the derivation of the plan and the illustrative examples the following advantages of QEP are apparent:

(*i*) The QEP model takes into account the time order of the observations, while in QMP the time order of the observations is ignored.

(*ii*) The best estimate of the process mean level is obtained by an adaptive exponential smoothing procedure. This makes QEP more responsive to the shifts and drifts in the process level. This is evidenced by the lower mean-squared prediction error for the examples discussed in Section VIII.

(*iii*) The QMP model is a special case of the QEP model. However, the two algorithms are quite different.

(*iv*) The computations are recursive. The entire past data are summarized by ten numbers.

(*v*) The computational efforts of QEP and QMP algorithms are comparable.

In the light of the advantages listed above it is proposed that QEP be considered as a serious alternative to QMP for official rating. In preparation for using QEP it is suggested that it be tried on all rating classes for a number of rating periods, and the resulting exception reports carefully compared with those from the QMP and the *t*-rate system. Such a study would aid us in fine tuning the starting conditions, quantifying the effect of truncation, and perhaps in making some other minor modifications for improving the performance of the QEP.

For small expectancies, the square root transform of the Poisson distribution has a significantly different variance than 0.25, assumed in Section II. Since the audit samples can at times be very small it would be necessary to use the correct variances. A study of this aspect will be done in a later memorandum.

The adaptive Kalman filtering methodology derived in this paper, with appropriate extensions and modifications, can be put to many other applications. In the field of quality control, Phadke[7] had developed a sequential empirical Bayes acceptance sampling plan. The adaptive Kaman filtering method developed in this paper would be particularly suited for updating the empirical prior distribution. Another potential application is in combining the traditional $\bar{X}$ and R control charts into a single box chart. Here the adaptive Kalman filter would permit one to take into account serial correlation in the data as well as process drifts and shifts, and changes in the process variance. Yet another application is in adaptive time series forecasting.

X. ACKNOWLEDGMENTS

I thank A. B. Hoadley for several discussions and suggestions at various stages of this project. I also thank Lakshman Sihna for useful discussions on the Kalman filtering theory. The box plots were generated using S. G. Crawford's software. Dolat Salsman and R. A. Cayford helped in developing the QEP computer program.

REFERENCES

1. H. F. Dodge, "A Method of Rating Manufactured Products," B.S.T.J. (April 1928), pp. 350–68.
2. A. B. Hoadley, "The Quality Measurement Plan," B.S.T.J., *60*, No. 2 (February 1981), pp. 215–273.
3. A. B. Hoadley, unpublished work.
4. Arthur Gelb, ed., *Applied Optimal Estimation*, Cambridge, MA: M.I.T. Press, 1974.

5. A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, New York: Academic Press, 1970.
6. M. S. Phadke, unpublished work.
7. M. S. Phadke, unpublished work.

APPENDIX A

Derivation of the Kalman filter solution

Let the conditional distribution of m_{t-1} given data up to time $t-1$ be normal with mean $\hat{m}_{t-1}$ and variance q_{t-1}, i.e.,

$$m_{t-1}|t-1 \sim N(\hat{m}_{t-1}, q_{t-1}). \tag{38}$$

Eq. (2) expresses m_t as a sum of two independent normal random variables m_{t-1} and ν_{2t}. Since the mean and variance of ν_{2t} are respectively 0 and σ_{2t}^2, it follows that

$$m_t|t-1 \sim N(\hat{m}_{t-1}, \sigma_{2t}^2 + q_{t-1}). \tag{39}$$

Substituting eq. (1) in eq. (3) we have

$$Y_t = m_t + \nu_{1t} + \eta_t, \tag{40}$$

which implies that

$$Y_t|m_t \sim N(m_t, \sigma_{1t}^2 + \sigma_{\eta t}^2). \tag{41}$$

In the Bayesian framework we may view eq. (39) as a prior distribution on m_t, and Y_t as an observation of m_t with the distribution specified by eq. (41). Applying the Bayes theorem the distribution of m_t conditional on data up to time t is seen to be

$$\begin{aligned} f(m_t|t) &\propto \exp\left\{-\frac{(m_t-\hat{m}_{t-1})^2}{2(\sigma_{2t}^2+q_{t-1})} - \frac{(Y_t-m_t)^2}{2(\sigma_{1t}^2+\sigma_{\eta t}^2)}\right\} \\ &\propto \exp\left\{-\frac{(m_t-\hat{m}_t)^2}{2q_t}\right\}, \end{aligned} \tag{42}$$

where

$$\hat{m}_t = \omega_{2t}\hat{m}_{t-1} + (1-\omega_{2t})Y_t, \tag{4}$$

$$\omega_{2t} = (\sigma_{1t}^2 + \sigma_{\eta t}^2)/(\sigma_{1t}^2 + \sigma_{\eta t}^2 + \sigma_{2t}^2 + q_{t-1}), \tag{5}$$

and

$$q_t = (1-\omega_{2t})(\sigma_{1t}^2 + \sigma_{\eta t}^2). \tag{6}$$

From eq. (42) it can be inferred that the distribution of m_t conditional on data up to time t is normal with mean $\hat{m}_t$ and variance q_t.

Equations (7) through (9), used for computing the conditional distribution of ζ_t, can be derived analogously as follows. First by substituting eq. (2) in (1) we have

$$\zeta_t = m_{t-1} + \nu_{1t} + \nu_{2t}; \tag{43}$$

hence,

$$\zeta_t | t-1 \sim N(\hat{m}_{t-1}, \sigma_{1t}^2 + \sigma_{2t}^2 + q_{t-1}). \tag{44}$$

From eq. (3) we have

$$Y_t | \zeta_t \sim N(\zeta_t, \sigma_{\eta t}^2). \tag{45}$$

Treating eq. (44) as a prior distribution for ζ_t and applying the Bayes theorem, we readily obtain the distribution of ζ_t conditional on data up to time t as

$$f(\zeta_t | t) \propto \exp\left\{ -\frac{(\zeta_t - \hat{m}_{t-1})^2}{2(\sigma_{1t}^2 + \sigma_{2t}^2 + q_{t-1})} - \frac{(Y_t - \zeta_t)^2}{2\sigma_{\eta t}^2} \right\}$$

$$\propto \exp\left\{ -\frac{(\zeta_t - \hat{\zeta}_t)^2}{2p_t} \right\}, \tag{46}$$

where

$$\hat{\zeta}_t = \omega_{1t}\omega_{2t}\hat{m}_{t-1} + (1 - \omega_{1t}\omega_{2t}) Y_t, \tag{7}$$

$$\omega_{1t}\omega_{2t} = \sigma_{\eta t}^2/(\sigma_{\eta t}^2 + \sigma_{1t}^2 + \sigma_{2t}^2 + q_{t-1}), \tag{8}$$

and

$$p_t = (1 - \omega_{1t}\omega_{2t})\sigma_{\eta t}^2. \tag{9}$$

Thus, the conditional distribution of ζ_t on data up to time t is normal with mean $\hat{\zeta}_t$ and variance p_t. Equations (7) through (9) form the desired recursive equations for computing $\hat{\zeta}_t$ and p_t.

APPENDIX B

Estimation of β and σ^2

Given the observed transformed defect indices $Y_0, Y_1, Y_2, \cdots, Y_n$ one can compute $Z_t = Y_t - Y_{t-1}$ for $t = 1, \cdots, n$. The Z_t series follows $MA(1)$ model given by eq. (13). The exponentially discounted probability density function of $a_1, \cdots, a_t$ is given by

$$p(a_1, \cdots, a_t | \sigma^2) = \prod_{j=1}^{t} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-a_j^2/2\sigma^2) \right]^{\lambda_j}, \tag{47}$$

where $\lambda_j = \lambda^{t-j}$. Thus, the exponentially discounted probability density function of $Z_1, \cdots, Z_t$ conditional on the knowledge of a_0, β, and σ^2 is given by

$$p(Z_1, \cdots, Z_t | a_0, \beta, \sigma^2) = \prod_{j=1}^{t} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-a_j^2/2\sigma^2) \right]^{\lambda_j}, \tag{48}$$

where a_j is related to Z_j and β via the recursion relation in eq. (13), i.e.,

$$a_j = Z_j - \beta a_{j-1}. \tag{49}$$

Thus, the conditional, exponentially discounted, log-likelihood function is

$$L_t(\beta, \sigma^2) = -\tfrac{1}{2}(A_t \ln \sigma^2 + S_t/\sigma^2), \tag{50}$$

where

$$A_t = \sum_{j=1}^{t} \lambda^{t-j}, \tag{51}$$

and

$$S_t = \sum_{j=1}^{t} \lambda^{t-j} a_j^2. \tag{52}$$

By differentiating eq. (50) with respect to β and σ^2 and equating the derivatives to zero, it can be shown that L_t is maximum at $(\hat{\beta}_t, \hat{\sigma}_t^2)$, where $\hat{\beta}_t$ is the minimum point of $S_t(\beta)$ and $\hat{\sigma}_t^2 = S_t(\hat{\beta}_t)/A_t$.

In the neighborhood of a point β_0, we can approximate a_j by the linear function:

$$a_j(\beta) = a_j(\beta_0) + (\beta - \beta_0) \left.\frac{da_j(\beta)}{d\beta}\right|_{\beta_0}. \tag{53}$$

Substituting this approximation in eq. (52) we have

$$S_t(\beta) \simeq S_t(\beta_0) + (\beta - \beta_0)\nu_t + \tfrac{1}{2}(\beta - \beta_0)^2 R_t, \tag{54}$$

where ν_t, R_t and $S_t(\beta_0)$ obey the recursion relations shown in eqs. (20) through (26). It is easy to verify that $\hat{\beta}_t$, given by eq. (18) minimizes the approximate $S_t(\beta)$ of eq. (54), and eq. (19) gives $\hat{\sigma}_t^2$.

The matrix of second partial derivatives of L_t is

$$-\begin{bmatrix} \dfrac{R_t}{2\sigma^2} & 0 \\ 0 & \dfrac{A_t}{2\sigma^4} \end{bmatrix},$$

so by the Fisher-information theory, the estimates $\hat{\beta}_t$ and $\hat{\sigma}_t^2$ are uncorrelated and their approximate variances are as given by eqs. (27) and (28).

The above recursive procedure also has a Bayesian interpretation, as given by Phadke.[6]

APPENDIX C

Summary of the Formulae

C.1 Initial conditions

$$\hat{m}_0 = 1.0$$

$$q_0 = 0.134$$

$$Y_0 = 1.0$$

$$\beta_0 = -0.6$$

$$\lambda = 0.95$$

$$a_0 = \frac{da_0}{d\beta} = 0$$

$$e_0 = e_1$$

$$S_0(\beta_0) = \frac{0.625}{e_0(1-\lambda)}$$

$$\nu_0 = 0$$

$$R_0 = 20.0/e_0$$

$$A_0 = 1/(1-\lambda)$$

$$\overline{\sigma^2_{\eta,0}} = 0.25/e_0$$

C.2 Recursive formulae

$$I_t = x_t/e_t$$

$$Y_t = \sqrt{I_t}$$

$$Z_t = Y_t - Y_{t-1}$$

$$\sigma^2_{\eta,t} = 0.25/e_t$$

$$a_t = Z_t - \beta_0 a_{t-1}$$

$$\frac{da_t}{d\beta} = -a_{t-1} - \beta_0 \frac{da_{t-1}}{d\beta}$$

$$S_t(\beta_0) = \lambda S_{t-1}(\beta_0) + a_t^2$$

$$\nu_t = \lambda \nu_{t-1} + 2a_t \frac{da_t}{d\beta}$$

$$R_t = \lambda R_{t-1} + 2\left(\frac{da_t}{d\beta}\right)^2$$

$$A_t = \lambda A_{t-1} + 1$$

$$\beta_t^* = \beta_0 - R_t^{-1}\nu_t, \qquad -1 \le \beta_t^* \le 0$$

$$S_t(\beta_t^*) = S_t(\beta_0) + (\beta_t^* - \beta_0)\nu_t + \tfrac{1}{2}(\beta_t^* - \beta_0)^2 R_t$$

$$\sigma_t^{*2} = S_t(\beta_t^*)/A_t$$

$$\overline{\sigma^2_{\eta,t}} = \lambda\, \overline{\sigma^2_{\eta,t-1}} + (1-\lambda)\sigma^2_{\eta,t}$$

$$\hat{\sigma}^2_{1t} = -\beta^*_t \sigma^{*2}_t - \overline{\sigma^2_{\eta t}}, \qquad \hat{\sigma}^2_{1t} \geq 0$$

$$\hat{\sigma}^2_{2t} = (1+\beta^*_t)^2 \sigma^{*2}_t$$

If $-\beta^*_t \sigma^{*2}_t - \overline{\sigma^2_{\eta t}} \geq 0$, then

$$\hat{\beta}_t = \beta^*_t \;\&\; \hat{\sigma}^2_t = \sigma^{*2}_t.$$

If $-\beta^*_t \sigma^{*2}_t - \overline{\sigma^2_{\eta n}} < 0$, then

$$\hat{\beta}_t = \frac{-(2 + \hat{\sigma}^2_{2t}/\overline{\sigma^2_{\eta t}}) + \sqrt{(2+\hat{\sigma}^2_{2t}/\overline{\sigma^2_{\eta t}})^2 - 4}}{2} \text{ and}$$

$$\hat{\sigma}^2_t = -\overline{\sigma^2_{\eta t}}/\hat{\beta}_t.$$

$$\hat{\omega}_{1t} = \frac{\sigma^2_{\eta t}}{\hat{\sigma}^2_{1t} + \sigma^2_{\eta t}}$$

$$\hat{\omega}_{2t} = \frac{\hat{\sigma}^2_{1t} + \sigma^2_{\eta t}}{\hat{\sigma}^2_{1t} + \sigma^2_{\eta t} + \hat{\sigma}^2_{2t} + q_{t-1}}$$

$$\hat{m}_t = \hat{\omega}_{2t}\hat{m}_{t-1} + (1-\hat{\omega}_{2t})\, Y_t$$

$$\hat{\xi}_t = \hat{\omega}_{1t}\hat{m}_t + (1-\hat{\omega}_{1t})\, Y_t$$

$$V(\omega_{2t}) = \frac{\{2\hat{\sigma}^6_t[1+\hat{\omega}_{2t}(1+2\hat{\beta}_t)]^2/R_t\} + \{2\hat{\sigma}^4_t[\hat{\beta}_t + (1+\hat{\beta}_t+\hat{\beta}^2_t)\hat{\omega}_{2t}]^2/A_t\}}{(\hat{\sigma}^2_{1t} + \hat{\sigma}^2_{2t} + \sigma^2_{\eta t} + q_{t-1})^2}; \; V(\omega_{2t}) \leq \frac{1}{12}$$

$$V(\omega_{1t}\omega_{2t}) = \frac{[2\hat{\sigma}^6_t(1+2\hat{\beta}_t)^2\hat{\omega}^2_{1t}\hat{\omega}^2_{2t}/R_t] + [2\hat{\sigma}^4_t(1+\hat{\beta}_t+\hat{\beta}^2_t)\hat{\omega}^2_{1t}\hat{\omega}^2_{2t}/A_t]}{(\hat{\sigma}^2_{1t} + \hat{\sigma}^2_{2t} + \sigma^2_{\eta t} + q_{t-1})^2}; \; V(\omega_{1t}\omega_{2t}) \leq \frac{1}{12}$$

$$q_t = (1-\hat{\omega}_{2t})(\hat{\sigma}^2_{1t} + \sigma^2_{\eta t}) + (Y_t - \hat{m}_{t-1})^2 V(\omega_{2t})$$

$$p_t = (1-\hat{\omega}_{1t}\hat{\omega}_{2t})\sigma^2_{\eta t} + (Y_t - \hat{m}_{t-1})^2 V(\omega_{1t}\omega_{2t})$$

C.3 Points for the box chart

Current defect index : I_t

Best estimate of the defect index : $\hat{\theta}_t = \hat{\xi}^2_t$

The mean level : $\hat{M}_t = \hat{m}^2_t$

99% quantile : $Q_1 = [\max(\hat{\xi}_t - 2.326\sqrt{p_t}, 0)]^2$

95% quantile : $Q_2 = [\max(\hat{\xi}_t - 1.645\sqrt{p_t}, 0)]^2$

5% quantile : $Q_3 = (\hat{\xi}_t + 1.645\sqrt{p_t})^2$

1% quantile : $Q_4 = (\hat{\xi}_t + 2.326\sqrt{p_t})^2$