The Exponentially Weighted Moving Average

Chapter · June 2010

DOI: 10.1002/9780470400531.eorms0314

1 author:

Marcus B. Perry
University of Alabama

THE EXPONENTIALLY WEIGHTED MOVING AVERAGE

MARCUS B. PERRY
Department of Information Systems, Statistics and Management Science, The University of Alabama, Tuscaloosa, Alabama

DEFINITION

The exponentially weighted moving average (EWMA) is often applied to a time-ordered sequence of random variables. It computes a weighted average of the sequence by applying weights that decrease geometrically with the age of the observations. The EWMA is parameterized by the following. Consider the $n \times 1$ random vector $\mathbf{x}$ given by

$$\mathbf{x} = [\, x_1, x_2, \ldots, x_n \,]', \tag{1}$$

with mean vector $\boldsymbol{\mu}$ and finite $n \times n$ autocovariance matrix Σ. The EWMA is defined by the linear transformation

$$\mathbf{z} = \mathbf{C}\mathbf{x} + z_0\mathbf{b}, \tag{2}$$

where $\mathbf{z}$ is $n \times 1$ and

$$\mathbf{C} = \begin{pmatrix} \lambda & 0 & 0 & \cdots & 0 \\ \lambda(1-\lambda) & \lambda & 0 & \cdots & 0 \\ \lambda(1-\lambda)^2 & \lambda(1-\lambda) & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda(1-\lambda)^{n-1} & \lambda(1-\lambda)^{n-2} & \lambda(1-\lambda)^{n-3} & \cdots & \lambda \end{pmatrix} \tag{3}$$

is a known $n \times n$ matrix and $\mathbf{b}$ is a known $n \times 1$ vector having the form

$$\mathbf{b} = \big((1-\lambda) \quad (1-\lambda)^2 \quad \cdots \quad (1-\lambda)^n\big)', \tag{4}$$

where z_0 is an initial (scalar) value and denotes the starting value for the EWMA. The parameter λ $(0 < \lambda \leq 1)$ is called the *smoothing* coefficient and its value in practice is often selected based upon how fast the process mean changes. For more rapid mean changes one should choose λ to be "large" and for slower (or less frequent) mean changes one should choose λ to be "small." In applications such as forecasting, λ is typically chosen between 0.05 and 0.30 [1] and [2]. It should be noted that when $\lambda = 1$ we have $\mathbf{z} = \mathbf{x}$. This is evident from Equation (2), when $\lambda = 1$ we have $\mathbf{C} = \mathbf{I}$ (where $\mathbf{I}$ is the identity matrix) and $\mathbf{b} = \mathbf{0}$ so that,

$$\mathbf{z} = \mathbf{C}\mathbf{x} + z_0\mathbf{b} = \mathbf{I}\mathbf{x} + z_0\mathbf{0} = \mathbf{x}. \tag{5}$$

To illustrate the EWMA, Fig. 1 shows a plot of $n = 100$ observations randomly generated from a normal process with a time-varying mean. The original process $\{x_i\}$ is shown as the dotted line and the EWMA process $\{z_i\}$ $(\lambda = 0.20)$ is shown as the solid line. For this realization, the initial value z_0 was set to zero. The value of $\lambda = 0.20$ was chosen by minimizing the sum of the squared one-step-ahead prediction errors. This will be discussed in more detail later.

Properties

By applying the expectation operator to Equation (2), it is easily shown that

$$\mathrm{E}(\mathbf{z}) = \mathbf{C}\boldsymbol{\mu} + z_0\mathbf{b}, \tag{6}$$

and by letting $\boldsymbol{\mu} = \mu\mathbf{1}_{n\times1}$ (i.e., stationary mean) and $z_0 = \mu$ we have

$$\begin{aligned}\mathrm{E}(\mathbf{z}) &= \mu\mathbf{C}\mathbf{1}_{n\times1} + \mu\mathbf{b} \\ &= \mu(\mathbf{C}\mathbf{1}_{\mathbf{n}\times\mathbf{1}} + \mathbf{b}) = \mu\mathbf{1}_{n\times1},\end{aligned} \tag{7}$$

Wiley Encyclopedia of Operations Research and Management Science, edited by James J. Cochran

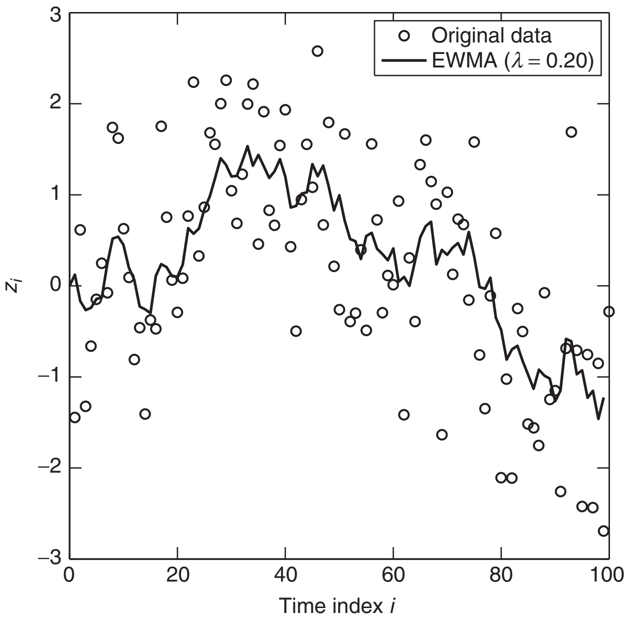

Figure 1. EWMA of a single simulated realization.

where $\mathbf{1}_{n\times 1}$ is a $n \times 1$ vector of ones. In addition, the $n \times n$ autocovariance matrix of $\mathbf{z}$ is given by

$$\mathrm{Var}(\mathbf{z}) = \mathrm{Var}(\mathbf{Cx} + z_0\mathbf{b}) = \mathbf{C}\Sigma\mathbf{C}' = \sigma_x^2\mathbf{CRC}', \tag{8}$$

where $\mathbf{R}$ denotes the $n \times n$ autocorrelation matrix of $\mathbf{x}$. The ith diagonal element of Equation (8) is the variance of the EWMA process at time i, and the (i,j)th off-diagonal element is the autocovariance of the EWMA process at lag$_{(i-j)}$=lag$_{(j-i)}$. It can be shown that the variance of z_i at any time $i > 0$ can be written explicitly as

$$\mathrm{Var}(z_i) = \sigma_x^2\lambda^2 \left(\sum_{j=0}^{i-1}(1-\lambda)^{2j} + 2\sum_{k=0}^{i-2}\sum_{\ell=k+1}^{i-1}(1-\lambda)^{k+\ell}\rho_{\ell-k}\right), \tag{9}$$

where $\rho_m = \rho_{-m}$ denotes the lag m autocorrelation of the original process $\{x_i\}$. It should be noted that $\mathrm{Var}(z_i) \to \sigma_z^2$ as $i \to \infty$, where σ_z^2 denotes the steady-state variance of the EWMA process and can be written as

$$\sigma_z^2 = \sigma_x^2 \left(\frac{\lambda}{2-\lambda} + 2\lambda^2\sum_{k=0}^{\infty} \times \sum_{\ell=k+1}^{\infty}(1-\lambda)^{k+\ell}\rho_{\ell-k}\right), \tag{10}$$

where the last term converges, as $|(1-\lambda)^{k+\ell}\rho_{\ell-k}| < 1$ for all k and ℓ. Note that for uncorrelated processes Equation (10) reduces to

$$\sigma_z^2 = \sigma_x^2\left(\frac{\lambda}{2-\lambda}\right). \tag{11}$$

In the next section, some discussion on EWMA forecasts is provided. In subsequent sections, application of the EWMA to sales forecasting and quality control applications are presented. Finally, this tutorial will close with a discussion.

COMPUTING EWMA FORECASTS

The EWMA forecast of a future observation at time $n + \ell$ is calculated by

$$\hat{x}_{n+\ell} = z_n = \lambda \sum_{j=0}^{n-1} (1-\lambda)^j x_{n-j} + (1-\lambda)^n z_0, \tag{12}$$

where for large n the last term is negligible and the forecast at the time $n + \ell$ can be written as

$$\hat{x}_{n+\ell} = z_n = \lambda \sum_{j=0}^{n-1} (1-\lambda)^j x_{n-j}. \tag{13}$$

Note that the forecasts generated by Equation (12) are the same for all lead times $\ell > 0$. To update forecasts with each new observation, one can use

$$z_n = \lambda x_n + (1-\lambda) z_{n-1}, \tag{14}$$

which can also be written as

$$z_n = z_{n-1} + \lambda(x_n - z_{n-1}), \tag{15}$$

where Equation (15) is written as the sum of the previous period's forecast and a fraction of the previous period's forecast error.

Suppose one uses z_n to forecast x_{n+1}, then for long running processes the variance of the one-step-ahead prediction error $(z_n - x_{n+1})$ is given by

$$\sigma^2_{\text{pred}} = \sigma^2_z + \sigma^2_x \left(1 - 2\lambda \sum_{j=0}^{n-1} (1-\lambda)^j \rho_{j+1} \right), \tag{16}$$

where σ^2_z is the steady-state EWMA variance, σ^2_x is the steady-state variance of the original process $\{x_i\}$, and ρ_m denotes the lag m autocorrelation of $\{x_i\}$. Notice that for uncorrelated processes $\rho_m = 0$ $(m \neq 0)$ and Equation (16) reduces to

$$\sigma^2_{\text{pred}} = \sigma^2_z + \sigma^2_x, \tag{17}$$

which is the sum of the steady-state variances of the EWMA process and the original process. More generally, the variance of the prediction error is defined for all lead times $\ell > 0$ by

$$\sigma^2_{\text{pred}} = \sigma^2_z + \sigma^2_x \left(1 - 2\lambda \sum_{j=0}^{n-1} (1-\lambda)^j \rho_{j+\ell} \right). \tag{18}$$

Initial Value for z_0

As noted previously, the influence of z_0 on z_n is negligible for large n. Thus, if n is large, one can choose any value for z_0 and its effect on z_n should be null. Some guidance for selecting z_0 in practice is given in Abraham and Ledolter [1], where these authors suggest using the arithmetic average of available historical data as the starting value.

Choice of Smoothing Coefficient

The smoothing coefficient λ determines how much the current observation affects the forecast. As mentioned previously, a small λ applies more weight to the previous observations and less weight to the current observation (and vice versa for large λ). Selection of values for λ in practice is typically accomplished via simulation. Specifically, forecasts are computed for a range of values of λ and are compared to the actual observations $x_1, x_2, \ldots, x_n$. For each value of λ, the one-step-ahead forecast errors are computed by

$$e_{i-1} = x_i - z_{i-1}, \tag{19}$$

for $i = 1, 2, \ldots, n$, and the sum of squared one-step-ahead forecast errors are computed by

$$\text{SSE}(\lambda) = \sum_{i=1}^{n} e^2_{i-1}, \tag{20}$$

where the smoothing coefficient λ that minimizes Equation (20) is then used to compute future forecasts.

To illustrate, consider the time series shown in Fig. 2. The data were simulated from an independent normal process with a time-varying mean. Suppose one considers values of $\lambda \in [0.05, 0.30]$ in increments of

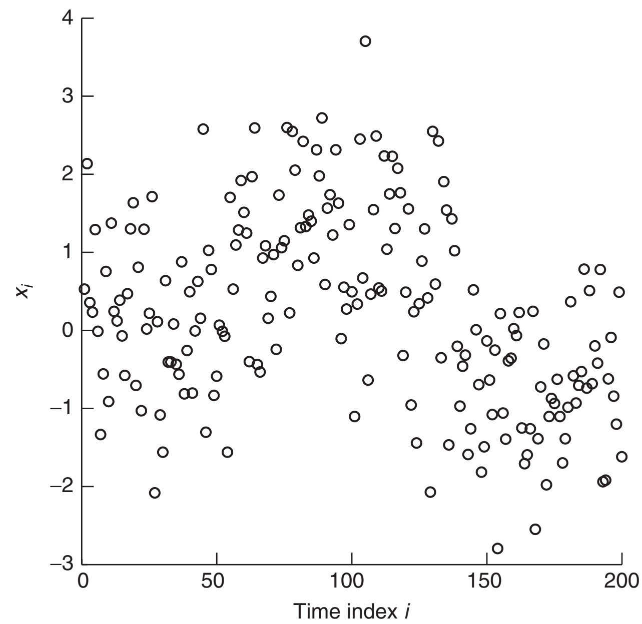

Figure 2. Single realization of simulated time series.

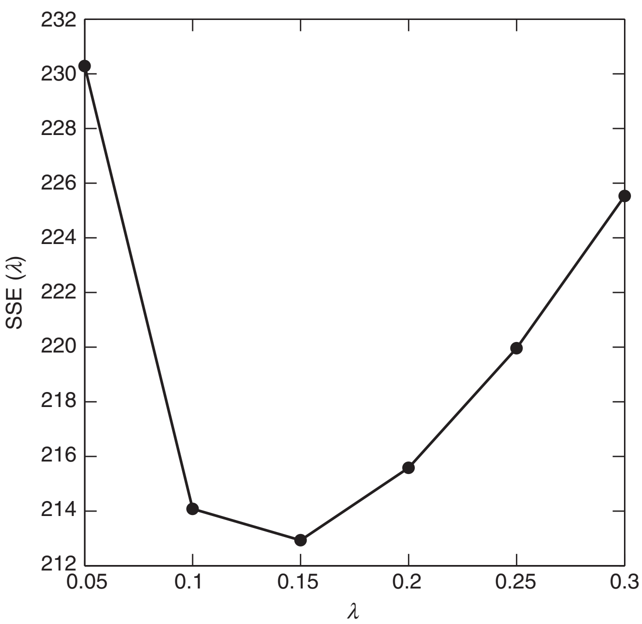

Figure 3. Plot of λ versus SSE(λ) for simulated series.

0.05, and computes SSE(λ) for each of these cases (i.e., $\lambda = 0.05, 0.10, 0.15, \ldots, 0.30$). Figure 3 shows a plot of λ versus SSE(λ) for the series shown in Fig. 2. Since $\lambda = 0.15$ yields the minimum sum of squared (one-step-ahead) prediction errors, this value will be used to compute future forecasts. The EWMA process $\{z_i\}$ with $\lambda = 0.15$ is shown in Fig. 4, superimposed on the original process $\{x_i\}$. The $(n + \ell)$th forecast $(\ell > 0)$ for this process is then computed by

$$\hat{x}_{n+\ell} = z_{n-1} + 0.15(x_n - z_{n-1}). \tag{21}$$

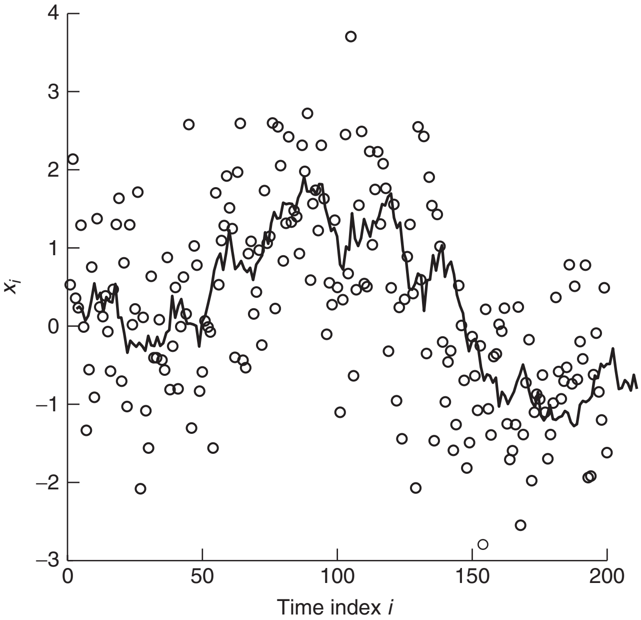

Figure 4. Original series $\{x_i\}$ and EWMA series $\{z_i\}$ ($\lambda = 0.15$) superimposed on same plot.

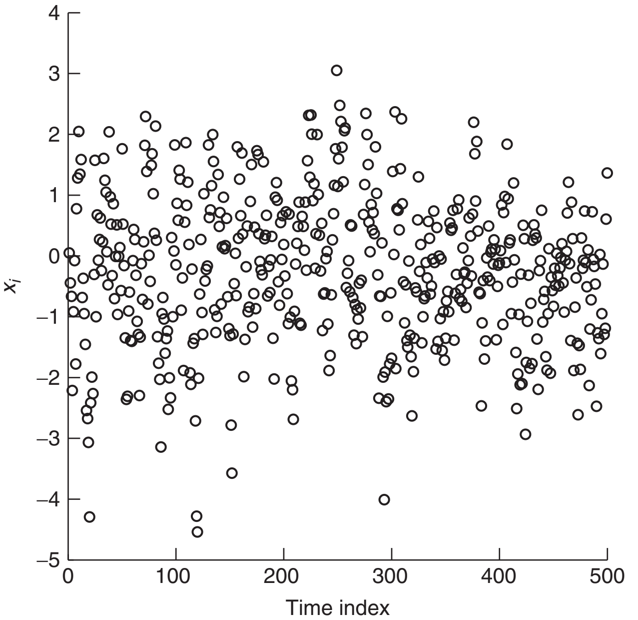

Figure 5. Single realization of simulated AR(1) process.

Consider the second example, where the process under study is well modeled by the first-order autoregressive process, or AR(1), given by

$$x_i = \mu + \phi(x_{i-1} - \mu) + \epsilon_i, \tag{22}$$

for $i = 1, 2, \ldots, n$, where $|\phi| < 1$ denotes the autoregressive parameter and the ϵ_i's are uncorrelated with constant variance σ_ϵ^2. A single realization of the AR(1) process with $\mu = 0$, $\phi = 0.60$, and $\sigma_\epsilon^2 = 1$ is shown in Fig. 5. Considering a range of $\lambda \in [0.15, 0.90]$

in increments of 0.15, Fig. 6 shows the value of λ that yields the minimum SSE(λ) is $\lambda = 0.60$. This is no surprise since the lag 1 autocorrelation for the AR(1) process with $\phi = 0.60$ is $\rho_1 = 0.60$. Figure 7 shows the EWMA process $\{z_i\}$ with $\lambda = 0.60$ superimposed on the original process $\{x_i\}$. For this example, $\lambda = 0.60$ should be used to compute future forecasts, where the forecast at time $n + \ell$ ($\ell > 0$) is given by

$$\hat{x}_{n+\ell} = z_{n-1} + 0.60(x_n - z_{n-1}). \quad (23)$$

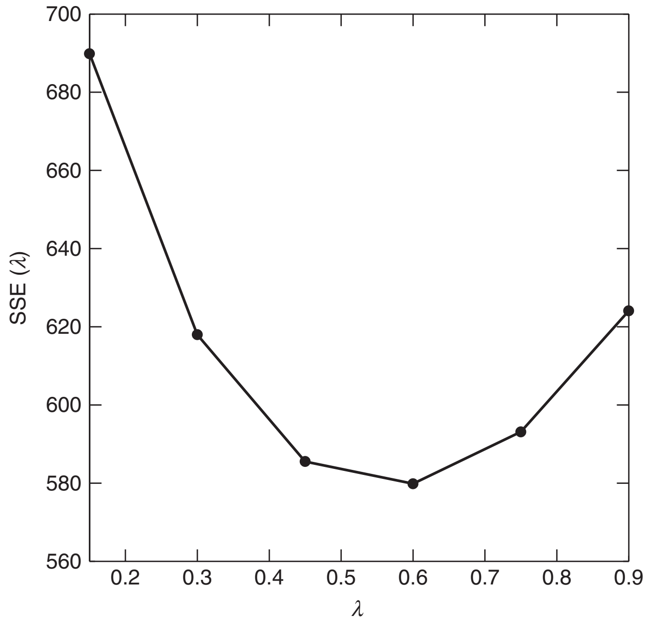

Figure 6. Plot of λ versus SSE(λ) for realization of AR(1) process.

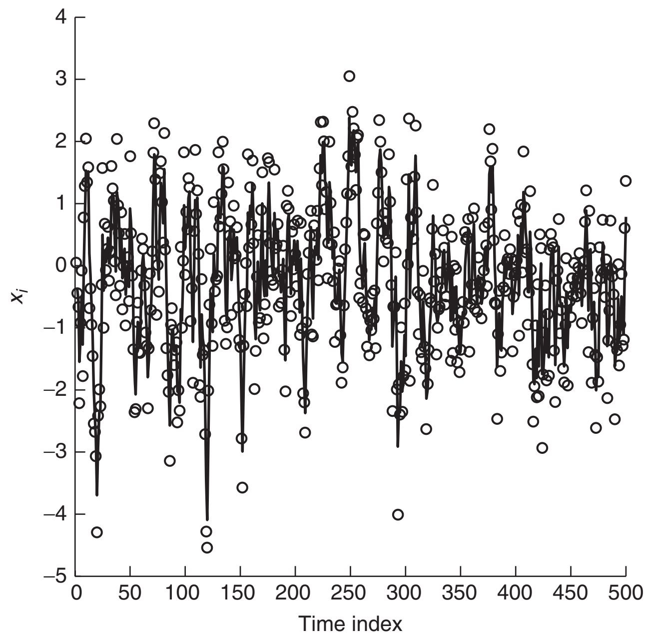

Figure 7. Original series $\{x_i\}$ and EWMA series $\{z_i\}$ ($\lambda = 0.60$) superimposed on same plot.

In the next section, some of the most common applications of the EWMA are discussed to include forecasting and application of the EWMA to quality engineering problems.

APPLICATIONS OF THE EWMA

One of the most common applications of the EWMA is generating one-step-ahead forecasts of time series. For example, consider Fig. 8, where the monthly Australian sales of rose wine is plotted over 173 months, from approximately January 1980 to April 1995. The data and its original source can be obtained from the Time Series Data Library.[1] Suppose we consider values of $\lambda \in [0.05, 0.30]$ in increments of 0.05, then the plot in Fig. 9a shows the value of λ, which minimizes SSE(λ) is $\lambda = 0.15$. The plot in Fig. 9b shows the EWMA with $\lambda = 0.15$ superimposed on the same plot as $\{x_i\}$. The forecast for x_{174} is then given by $\hat{x}_{174} = z_{173} = 47.23$. The actual sales in the 174th month is $x_{174} = 45.00$, and the forecast error is given by $e_{173} = -2.23$.

Another common application of the EWMA is in quality control applications. Roberts [3] was the first to suggest the use of EWMA as the basis for a process monitoring and control scheme (i.e., control chart). The idea is to chart the z_i at each time i and compare to an upper and lower control limit. If z_i exceeds one of these control limits, then the process is deemed "out of statistical control." Specifically, at each time $i > 0$, the EWMA z_i is compared to the control limits

$$\mu_0 \pm L\sqrt{\mathrm{Var}(z_i)}, \tag{24}$$

where μ_0 is the expected value of the process, $\mathrm{Var}(z_i)$ is defined in Equation (9) and L is a constant expressed in standard deviation units.[2] If z_i exceeds one of the control limits in Equation (24), then one would conclude with strong evidence that the process mean has shifted. It is common in practice to use the steady-state control limits, which is given by

$$\mu_0 \pm L\sqrt{\sigma_z^2}, \tag{25}$$

where σ_z^2 is defined in Equation (10).

To illustrate the application of the EWMA control chart, consider the statistical monitoring of an industrial process to assess its quality over time. Suppose that the expected (or in-control) mean value of the original process $\{x_i\}$ is $\mu_0 = 0$, that samples taken over time are uncorrelated, and the variance of $\{x_i\}$ is given by $\sigma_x^2 = 1$. Suppose further that following the 50th sample collected, the process mean shifts from $\mu_i = 0$ to $\mu_i = 1$, so that the 51st sample collected is the first obtained from the changed process. Then the plot in Fig. 10a shows a single realization of this process. The plot in Fig. 10b shows the EWMA control chart; which signaled at $n = 58$, which implies eight observations were sampled from the changed process before the control chart issued a signal. Notice that the control chart limits reach steady-state after roughly 15 observations.

The above illustration is an example of how the EWMA can be used as the basis for a process change detection algorithm. Although the process in the latter example is uncorrelated, the EWMA can be used to monitor autocorrelated processes as well [4–6]. For more details on the development, design, and application of the EWMA control chart see Roberts [3], Hunter [7], Montgomery [2], and Lucas and Saccucci [8].

The above examples are not the only applications of the EWMA; however, they are common applications of the EWMA in the management science and quality engineering disciplines. In general, application of the EWMA is widespread throughout several disciplines, including health, finance, criminal justice, sports, industry, demography, and many more.

DISCUSSION

In this tutorial, the EWMA was discussed. As previously noted, EWMA is frequently

---

[1] The Time Series Data Library is accessible at http://www.robjhyndman.com/TSDL/.

[2] The constant L is typically chosen based upon an acceptable false alarm rate.

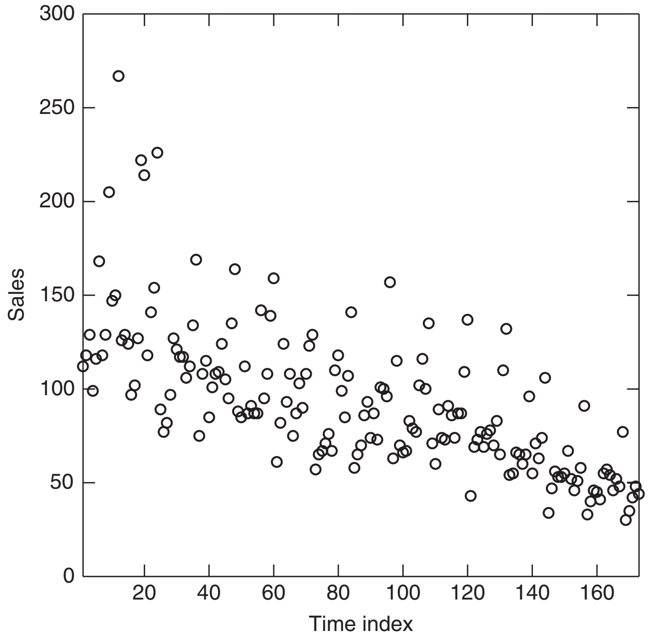

Figure 8. Plot of Australian rose wine over 173 months.

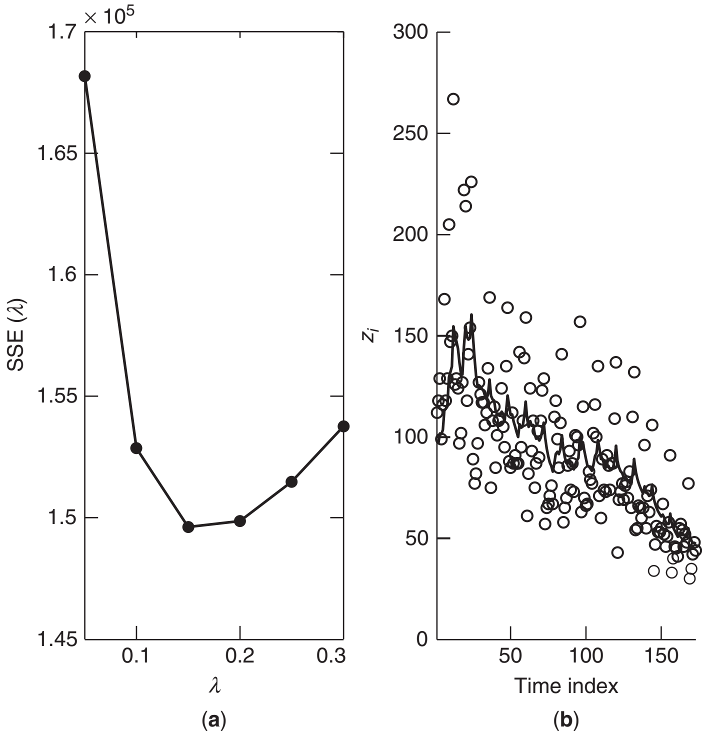

Figure 9. (a) Plot of λ versus SSE(λ). (b) EWMA of rose wine data with $\lambda = 0.15$.

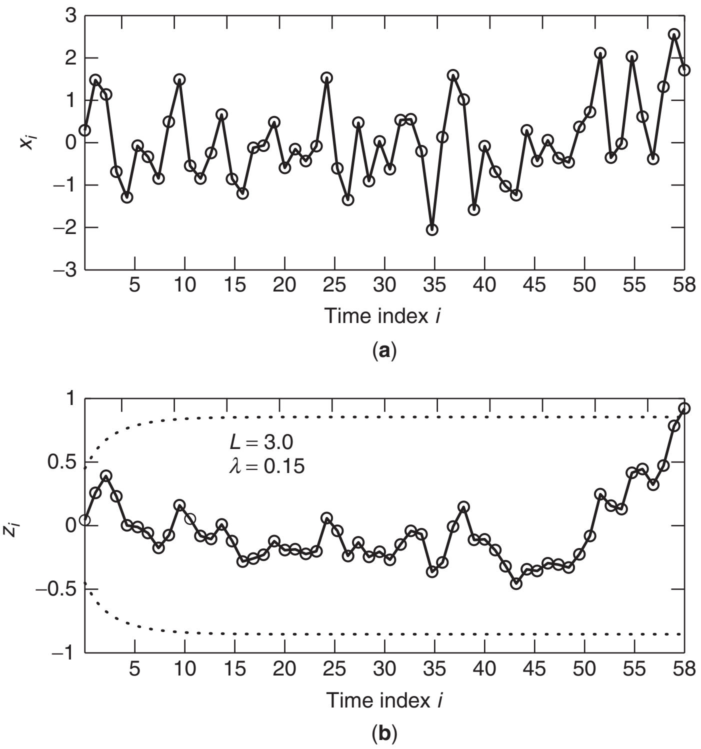

Figure 10. (a) Plot of original series $\{x_i\}$. (b) EWMA control chart with $L = 3$ and $\lambda = 0.15$. Control chart signaled at $n = 58$.

used to compute short-term forecasts of time series (e.g., sales, stocks, and so on). It is also used as a basis for developing process monitoring and control schemes. Further, some properties of the EWMA, to include its expected value and variance, as well as the ℓ-step ahead forecast error variance were provided. Some discussion on computing EWMA forecasts was also included, as well as a commonly used method for selecting a value for λ in practice. Finally, some common applications of the EWMA in the management science and quality engineering disciplines were discussed.

REFERENCES

1. Abraham B, Ledolter J. Statistical methods for forecasting. New York: John Wiley & Sons, Inc.; 1983.
2. Montgomery DC. Introduction to statistical quality control. New York: John Wiley & Sons, Inc.; 2005.
3. Roberts SW. Control chart tests based on geometric moving averages. Technometrics 1959;1:239–250.
4. Zhang NF. A statistical control chart for stationary process data. Technometrics 1998;40:24–38.
5. Montgomery DC, Mastrangelo CM. Some statistical process control methods for autocorrelated processes. J Qual Tech 1991;23(3):179–204.
6. Lu CW, Reynolds MR Jr. EWMA control charts for monitoring the mean of autocorrelated processes. J Qual Tech 1999;31(2):189–206.
7. Hunter JS. A one-point plot equivalent to the shewhart chart with western electric runs rules. Qual Eng 1990;2(1):13–19.
8. Lucas JM, Saccucci MS. Exponentially weighted moving average control schemes: properties and enhancements. Technometrics 1990;32:1–29.