Chapter 2: Statistics Background for Forecasting

Rong Fan

South Dakota State University

rong.fan@sdstate.edu

2.1 Introduction

A time series is a sequence of a collection of observations taken over time. We denoted the value of the time series at time t by y_t, where $t = 1, 2, ...T$.

A fitted value is obtained from estimating the parameters in a model to historial data, which is denoted by $\hat{y}_t$. The difference between y_t and $\hat{y}_t$ is called a residual,

$$e_t = y_t - \hat{y}_t.$$

The forecast of y_t is made at some previous time period, say $t - \tau$, is denoted by $\hat{y}_t(t - \tau)$. The difference

$$e_t(\tau) = y_t - \hat{y}_t(t - \tau)$$

is called the lead -τ forecast error. In particular, the lead -1 forecast error is

$$e_t(1) = y_t - \hat{y}_t(t - 1)$$

A time series plot is a graph of y_t versus the time t, for $t = 1, 2, ...T$.

The following is a time series plot for viscosity readings and a histogram of the data.

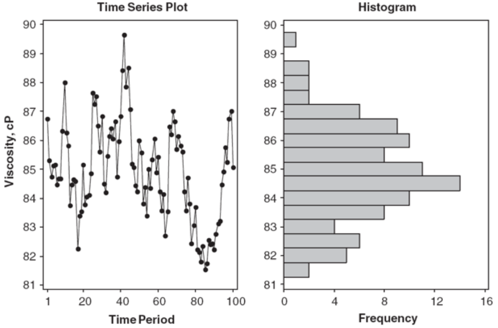


FIGURE 2.1 Time series plot and histogram of chemical process viscosity readings.

The following is a time series plot for beverage production shipments and a histogram of the data.

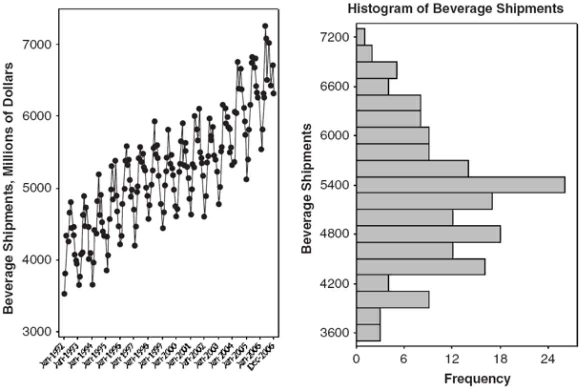


FIGURE 2.2 Time series plot and histogram of beverage production shipments.

Note that although the histograms are similar, the two time series plots display very different characteristics.

A variation of time series plot

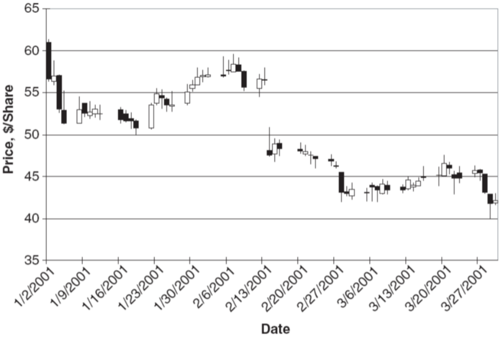


FIGURE 2.4 Open-high/close-low chart of whole foods market stock price. (*Source*: `finance.yahoo.com` .)

2.2 Graphical Displays IV

We now looked at some smoothed versions of data.
A simple Moving Average (MA) of span N assigns weights $1/N$ to the most recent N observations and weight zero to all other observations. In other words, let M_T be the moving average, then the N-span moving average is

$$M_T = \frac{y_T + y_{T-1} + \cdots + y_{T-N+1}}{N} = \frac{1}{N} \sum_{t=T-N+1}^{T} y_t$$

M_T is also a time series and has less variability than the original observations.
If we assume $Var(y_t) = \sigma^2$ and the observations are uncorrelated, then

$$Var(M_T) = Var\left(\frac{1}{N} \sum_{t=T-N+1}^{T} y_t\right) = \frac{1}{N^2} \sum_{t=T-N+1}^{T} Var(y_t) = \frac{\sigma^2}{N}$$

which is smaller than σ^2.

2.2 Graphical Displays V

A "centered" version of the moving average is

$$M_T = \frac{1}{2S+1} \sum_{t=-S}^{S} y_{t-i}$$

notice here $N = 2S + 1$.

A time series plot and its five-period moving average:

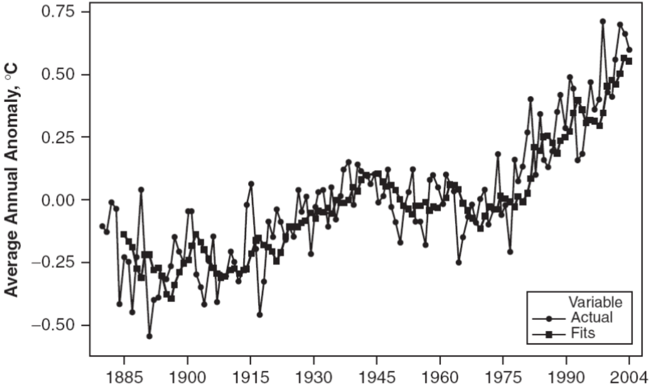


FIGURE 2.5 Time series plot of global mean surface air temperature anomaly, with five-period moving average. (*Source*: NASA–GISS.)

Moving averages are commonly used in stock trend analysis. A time series plot of Whole Foods market stock price with 50-day MA is shown as

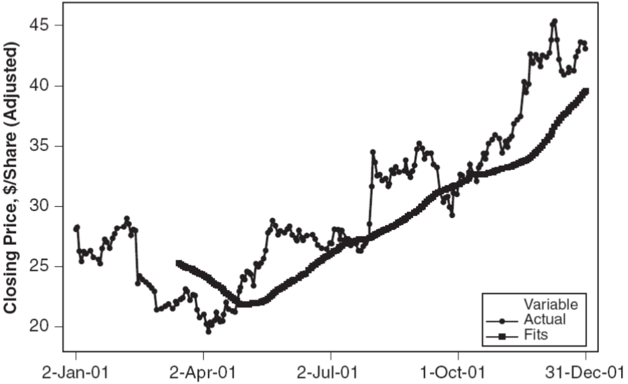


FIGURE 2.6 Time series plot of whole foods market stock price, with 50-day moving average. (*Source*: finance.yahoo.com .)

The candlestick plot for SP500 with its 50 MA.

A linear data smoother, or a linear filter replaces each observation y_t with a linear combination of the other data points. Notice the simple moving average is a linear filter with equal weights. There are cases where unequal weights are used.

For example, the Hanning filter is defined as follows:

$$M_t^H = \frac{1}{4}y_{t+1} + \frac{1}{2}y_t + \frac{1}{4}y_{t-1}$$

2.2 Graphical Displays I

Recall the median is more resistant to outliers than mean, so moving medians are an alternative to moving averages when the time series may contain unusual values or outliers.

The moving median of span N is

$$m_t^{[N]} = med(y_{t-S}, ...y_t, ..., y_{t+S})$$

where $N = 2S + 1$.

In particular, the case $N = 3$ is very popular:

$$m_t^{[3]} = med(y_{t-1}, y_t, y_{t+1})$$

This smoother would process the data three values at a time, and replace the three original observations by their median.

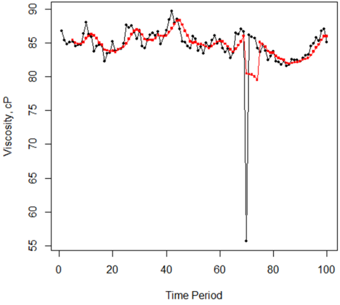
2.2 Graphical Displays I

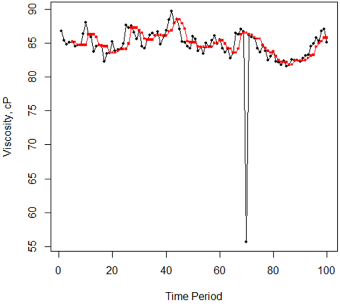
Example: Consider the sequence of observations:
15 18 13 12 16 14 16 17 18 15 18 200 19 14 21 24 19 25

If we apply $m_t^{[3]}$ to the data above, we obtain
— 15 13 13 14 16 17 17 18 18 19 19 19 21 21 24 —

Moving medians can be applied more than once if desired to obtain an even smoother series of observations. For example, apply $m_t^{[3]}$ to the smoothed data above results in
— — 13 13 14 16 17 17 18 18 19 19 19 21 21 — —

Recall the viscosity example with sensor malfunction. Here is the plot with 5 MA. One can see a comparison between the moving average which is contaminated by 5 periods and the moving median of span 5. The moving median minimizes the impact of the invalid data point.





2.3 Numerical Description of Time Series Data I

A time series is said to be strictly stationary if its properties are not affected by a change of time. That is, the joint probability distribution of y_t, y_{t+1}... y_{t+n} is exactly the same as the joint probability distribution of y_{t+k}, y_{t+k+1},..., y_{t+k+n}

Recall the joint probability distribution of y_0, y_1 ... y_n is

$$F_{y_0,y_1...y_n}(x_0, x_1, ...x_n) = P(y_0 < x_0, y_1 < x_1, ..., y_n < x_n), \qquad (1)$$

where y_0, y_1, ..., y_n are random variables and x_0, x_1 ... x_n are numbers. The pharmaceutical product sales data is shown below:

2.3 Numerical Description of Time Series Data II

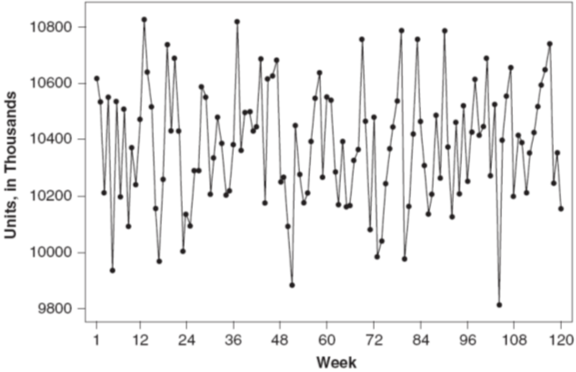


FIGURE 2.8 Pharmaceutical product sales.

Note that the time series seems to vary around a fixed level, which is one of the features of stationary time series.

The Whole Foods Market stock price data is shown below:

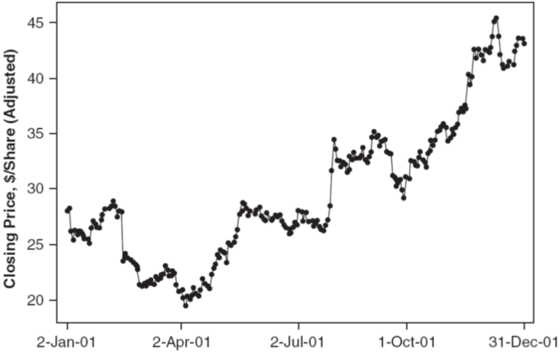


FIGURE 1.7 Whole Foods Market stock price, daily closing adjusted for splits.

Note that the time series tends to wander around or drift, with no obvious fixed level. This indicates it is a nonstationary time series.

2.3 Numerical Description of Time Series Data IV

A stationary time series has a constant mean (assuming the density function is $f(y)$)

$$\mu_y = E(y) = \int_{-\infty}^{\infty} y f(y)\ dy \tag{2}$$

and constant variance

$$\sigma_y^2 = Var(y) = \int_{-\infty}^{\infty} (y - \mu_y)^2 f(y)\ dy \tag{3}$$

They can be estimated by sample mean and sample variance:

$$\bar{y} = \hat{\mu}_y = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{4}$$

$$s^2 = \hat{\sigma}_y^2 = \frac{1}{T}\sum_{t=1}^{T} (y_t - \bar{y})^2 \tag{5}$$

2.3 Numerical Description of Time Series Data V

If a time series is stationary, then for a fixed k (which is called the lag), the joint distribution of y_t and y_{t+k} is the same for any t.

To study the joint distribution and the nature of the time series, one can plot a scatter diagram of all of the data pairs that are separated by the same lag k, i.e. y_t vs y_{t+k} for all t.

2.3 Numerical Description of Time Series Data VI

Figure 2.10 is a scatter diagram for the pharmaceutical product sales for lag $k = 1$.

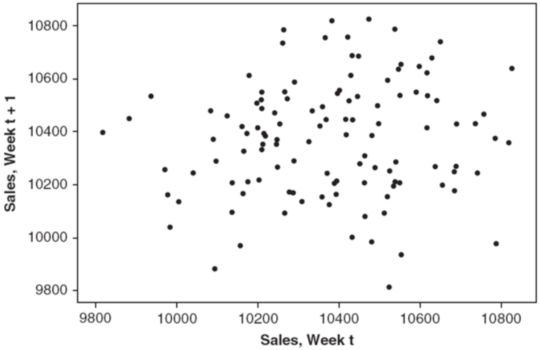


FIGURE 2.10 Scatter diagram of pharmaceutical product sales at lag $k = 1$.

It exhibits little structure. y_t and y_{t+k} seem to be uncorrelated.

Figure 2.11 is a scatter diagram for the chemical viscosity readings for lag $k = 1$.

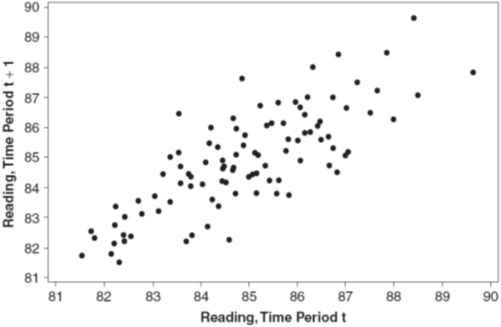


FIGURE 2.11 Scatter diagram of chemical viscosity readings at lag $k = 1$.

The pairs of adjacent observations y_t, y_{t+1} are positively correlated.

2.3 Numerical Description of Time Series Data VIII

Note that the behavior inferred from the scatter diagrams can be observed from the time series plots.

The covariance between y_k and its value at another time, say y_{t+k} is called the autocovariance at lag k, defined by

$$\gamma_k = Cov(y_t, y_{t+k}) = E[(y_t - \mu)(y_{t+k} - \mu)] \tag{6}$$

The collection of γ_k is called the autocovariance function. Note that $\gamma_0 = \sigma_y^2$.

The autocorrelation coefficient at lag k for a stationary time series is

$$\rho_k = \frac{Cov(y_t, y_{t+k})}{Var(y_t)} = \frac{\gamma_k}{\gamma_0} \tag{7}$$

The collection of ρ_k is called the autocorrelation function(ACF). Note that $\rho_0 = 1$.

2.3 Numerical Description of Time Series Data IX

If a time series has a finite mean and autocovariance function, it is said to be **weakly stationary** of order 2. An equivalent way of stating this is:

1. $E(y_t) = \mu$ does not depend on t.
2. $Var(y_t) = \gamma(0)$ does not depend on t.
3. $Cov(y_t, y_{t+k}) = \gamma(k)$.

We say y_t is **Gaussian**, if the joint probability distributions of all sets of $(y_{t1}, y_{t2}, ...y_{tT})$ are multivariate normal distributions.

It can be shown that if a time series is Gaussian and weakly stationary, then it is **strictly stationary**.

2.3 Numerical Description of Time Series Data X

Example(White Noise): suppose ϵ_t is a sequence of independent random variables with $E(\epsilon_t) = 0$, and $Var(\epsilon_t) = \sigma^2$. Then it is a weakly stationary process and is called a white noise process.

1. Mean $E(\epsilon_t) = 0$ and $Var(\epsilon_t) = \sigma^2$ do not depend on t.
2. The autocovariance function is

$$Cov(\epsilon_t, \epsilon_{t+k}) = \begin{cases} \sigma^2, & k = 0 \\ 0, & k \neq 0 \end{cases}$$

   which depends only on k.
3. The ACF is

$$\rho(k) = Cor(\epsilon_t, \epsilon_{t+k}) = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases}$$

2.3 Numerical Description of Time Series Data XI

Example: suppose ϵ_t is a white noise process. Define $y_t = \mu + \epsilon_t + \epsilon_{t-1}$, where μ is a constant. Then y_t is stationary.

1. Mean $E(y_t) = \mu$.
2. The autocovariance function is

$$Cov(y_t, y_{t+k}) = \gamma(k) = \begin{cases} 2\sigma^2, & k = 0 \\ \sigma^2, & k = 1 \\ 0, & k > 1 \end{cases}$$

   which depends only on k.
3. The ACF is

$$\rho(k) = Cor(y_t, y_{t+k}) = \begin{cases} 1, & k = 0 \\ 0.5, & k = 1 \\ 0, & k > 1 \end{cases}$$

2.3 Numerical Description of Time Series Data XII

General case: suppose ϵ_t is a white noise process. Then

$$y_t = \mu + \epsilon_t + \psi_1 \epsilon_{t-1} + \psi_2 \epsilon_{t-2} + \ldots + \psi_q \epsilon_{t-q}$$

where μ, ψ_i are constants. Then y_t is stationary.

To estimate the autocovariance and ACFs from a time series of finite length, say y_1, y_2,...,y_T. The usual estimate of the autocovariance function is

$$c_k = \hat{\gamma}_k = \frac{1}{T}\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y}), \qquad k = 0, 1, 2, ..., K$$

and the ACF is estimated by sample ACF

$$r_k = \hat{\rho}_k = \frac{c_k}{c_0}, \qquad k = 0, 1, 2, ..., K$$

We want $T \geq 50$ to get a reliable estimate of ACF. The sample autocorrelations should be calculated up to lag K, where K is about $T/4$.

2.3 Numerical Description of Time Series Data XIV

It is important to determine if the autocorrelation coefficient at a particular lag is zero. Assume that the true value of the autocorrelation coefficient at lag k is zero, i.e. $\rho_k = 0$, then the standard error is

$$se(r_k) \tilde{=} \frac{1}{\sqrt{T}}$$

which can be used to find the significance limits for sample ACF.

For the data of the chemical process viscosity readings example (See text page 40):

Table: Chemical Process Viscosity Readings

Time	Reading	Time	Reading	Time	Reading	Time	Reading
1	86.7418	26	87.2397	51	85.5722	76	84.7052
2	85.3195	27	87.5219	52	83.7935	77	83.8168
3	84.7355	28	86.4992	53	84.3706	78	82.4171
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
24	84.8495	49	84.2112	74	84.2339	99	87.0048
25	87.6416	50	85.9952	75	83.5737	100	85.0572

The sample ACF at lag $k = 1$ is calculated as

$$
\begin{aligned}
c_0 &= \frac{1}{100} \sum_{t=1}^{100-0} (y_t - \bar{y})(y_{t+0} - \bar{y}) \\
&= \frac{1}{100} \sum_{t=1}^{100} (y_t - \bar{y})^2 \\
&= \frac{1}{100} [(86.7418 - 84.9153)^2 + (85.3195 - 84.9153)^2 + \cdots \\
&\quad + (85.0572 - 84.9153)^2] \\
&= 280.9332
\end{aligned}
$$

$$
\begin{aligned}
c_1 &= \frac{1}{100}\sum_{t=1}^{100-1}(y_t - \bar{y})(y_{t+1} - \bar{y}) \\
&= \frac{1}{100}[(86.7418 - 84.9153)(85.3195 - 84.9153) + \cdots \\
&\quad + (87.0048 - 84.9153)(85.0572 - 84.9153)] \\
&= 220.3137
\end{aligned}
$$

$$
r_1 = \frac{c_1}{c_0} = \frac{220.3137}{280.9332} = 0.7842
$$

$$
\begin{aligned}
c_2 &= \frac{1}{100} \sum_{t=1}^{100-2} (y_t - \bar{y})(y_{t+2} - \bar{y}) \\
&= \frac{1}{100} [(86.7418 - 84.9153)(84.7355 - 84.9153) \\
&\quad + (85.3195 - 84.9153)(85.1113 - 84.9153) + \cdots \\
&\quad + (86.7312 - 84.9153)(85.0572 - 84.9153)] \\
&= 176.4401
\end{aligned}
$$

$$
r_2 = \frac{c_2}{c_0} = \frac{176.4401}{280.9332} = 0.62805
$$

Similar, we are able to obtain $r_3, r_4, \cdots, r_{25}, \cdots$

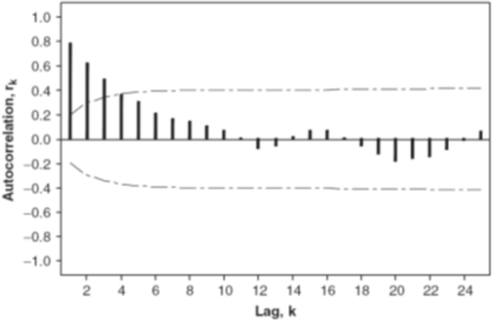


FIGURE 2.12 Sample autocorrelation function for chemical viscosity readings, with 5% significance limits.

Notice the ACF values decrease 0.78 to 0, followed by a sinusoidal pattern about 0. This pattern is typical of stationary time series.

Another stationary example would be the pharmaceutical product sales with the sample ACF values plotted as follows:

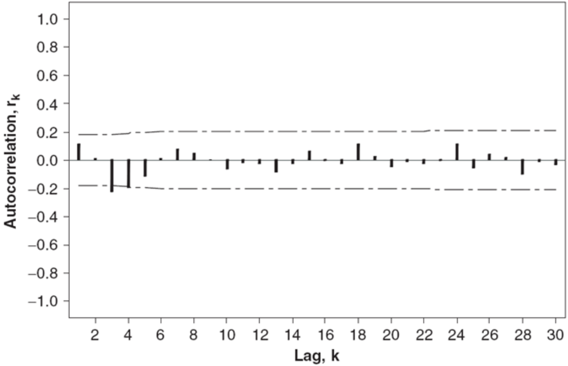


FIGURE 2.14 Autocorrelation function for pharmaceutical product sales, with 5% significance limits.

What does the sample ACF of a nonstaionary time series look like? Let us look at the Whole Foods Market stock example:

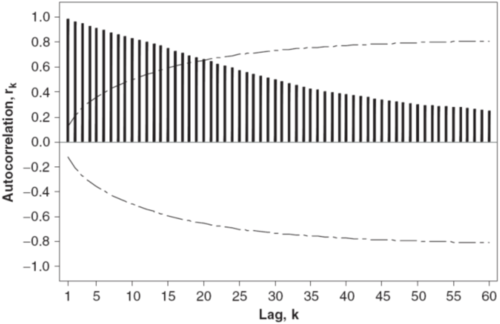


FIGURE 2.15 Autocorrelation function for Whole Foods market stock price, with 5% significance limits.

Notice that the sample ACF decays very slowly and the sample autocorrelations are still large at long lags. This is characteriestic of a nonstationary time series. Generally, if the sample ACF does not dampen out within about 15 to 20 lags, the time series is nonstationary.

Another useful diagnostic tool is the variogram. The variogram G_k measures variances of the differences between observations that are k lags apart, relative to the variance of the differences that are 1 lag apart.

Definition (Variogram)

$$G_k = \frac{Var(y_{t+k} - y_t)}{Var(y_{t+1} - y_t)}, \qquad k = 1, 2 \ldots \tag{8}$$

The values of G_k are plotted as a function of the lag k.

If the time series is stationary, it turns out that

$$G_k = \frac{1 - \rho_k}{1 - \rho_1} \tag{9}$$

2.3 Numerical Description of Time Series Data XXIV

Since for a stationary time series $\rho_k \to 0$, we have $G_k \to 1/(1-\rho_1)$. However, if the time series is nonstationary, G_k will increase monotonically.

$$
\begin{aligned}
G_k &= \frac{Var(y_{t+k}-y_t)}{Var(y_{t+1}-y_t)} && (10)\\
&= \frac{Var(y_{t+k})+Var(y_t)-2cov(y_t,y_{t+k})}{Var(y_{t+1})+Var(y_t)-2cov(y_t,y_{t+1})} && (11)\\
&= \frac{\sigma^2+\sigma^2-2\gamma_k}{\sigma^2+\sigma^2-2\gamma_1} && (12)\\
&= \frac{2\sigma^2-2\gamma_k}{2\sigma^2-2\gamma_1} && (13)\\
&= \frac{\sigma^2-\gamma_k}{\sigma^2-\gamma_1} && (14)\\
&= \frac{1-\frac{\gamma_k}{\sigma^2}}{1-\frac{\gamma_1}{\sigma^2}} = \frac{1-\rho_k}{1-\rho_1} && (15)
\end{aligned}
$$

Estimating the variogram:

Let $d_t^k = y_{t+k} - y_t$ and $\bar{d}^k = \frac{1}{T-k}\sum d_t^k$.
Then an estimate of $Var(y_{t+k} - y_t)$ is

$$s_k^2 = \frac{\sum_{t=1}^{T-k}(d_t^k - \bar{d}^k)^2}{T-k-1}. \qquad (16)$$

Therefore the **sample variogram** is given by

$$\hat{G}_k = \frac{s_k^2}{s_1^2}, \qquad k = 1, 2... \qquad (17)$$

Recall in the viscosity readings example, both data plot and sample ACF suggest that the time series is stationary. The variogram is shown below.

Lag	Variogram	Plot Variogram
1	1.0000	
2	1.7238	
3	2.3562	
4	2.9527	
5	3.2230	
6	3.6659	
7	3.8729	
8	3.9634	
9	4.1541	
10	4.3259	
11	4.6161	
12	4.9923	
13	4.8752	
14	4.5393	
15	4.2971	
16	4.3065	
17	4.6282	
18	4.9006	
19	5.2050	
20	5.4711	
21	5.3873	
22	5.3109	
23	5.0395	
24	4.6880	
25	4.3416	

FIGURE 2.16 JMP output for the sample variogram of the chemical process viscosity data from Figure 2.19.

2.3 Numerical Description of Time Series Data XXVII

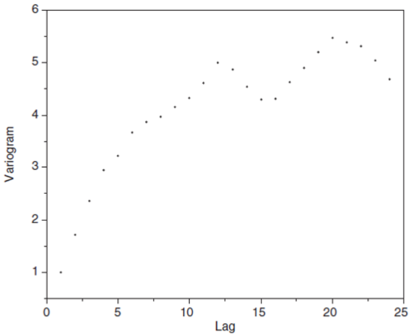


FIGURE 2.17 JMP sample variogram of the chemical process viscosity data from Figure 2.9.

Notice that the sample variogram converges to a level and then fluctuates around it which is consistent with a stationary time series.

Remark: it is not super complicated to compute the sample variogram "by hand".

2.3 Numerical Description of Time Series Data XXVIII

Recall in the Whole Foods example, we have non-stationarity. The variogram is shown below.

Lag	Variogram	Plot Variogram
1	1.0000	
2	2.0994	
3	3.2106	
4	4.3960	
5	5.4982	
6	6.5810	
7	7.5690	
8	8.5332	
9	9.4704	
10	10.4419	
11	11.4154	
12	12.3452	
13	13.3759	
14	14.4411	
15	15.6184	
16	16.9601	
17	18.2442	
18	19.3782	
19	20.3934	
20	21.3618	
21	22.4010	
22	23.4788	
23	24.5450	
24	25.5906	
25	26.6620	

FIGURE 2.18 JMP output for the sample variogram of the Whole Foods Market stock price data from Figure 1.7 and Appendix Table B.7.

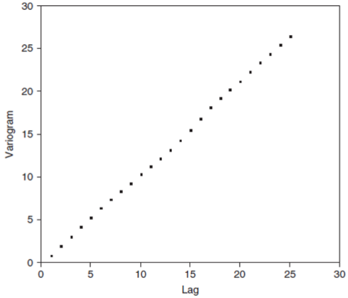


FIGURE 2.19 Sample variogram of the Whole Foods Market stock price data from Figure 1.7 and Appendix Table B.7.

2.3 Numerical Description of Time Series Data- Get Variogram by R

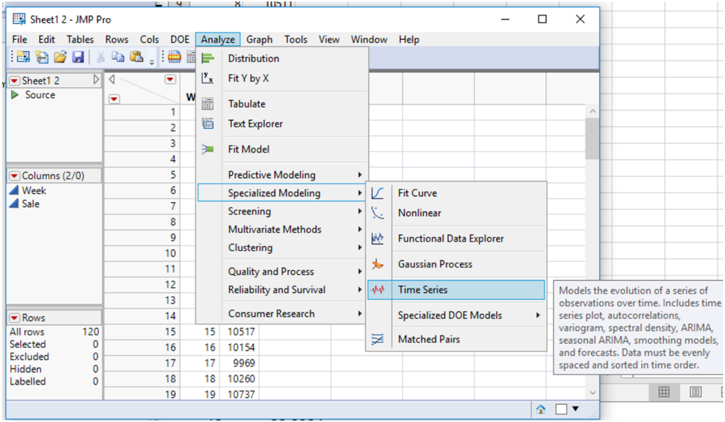
```
#Define the variogram function
variogram <- function(x, lag) {
  x <- as.matrix(x) # Set x as a "matrix".
  Lag <- NULL
  var_k <- NULL
  vario <- NULL
  for (k in 1:lag) {
    Lag[k] <- k
    var_k[k] <- sd(diff(x, k))^2
    vario[k] <- var_k[k] / var_k[1]
  }
  return(as.data.frame(cbind(Lag, vario)))
}
# Let y=pharma.data[,2] from Example 2.17.
# make sure the class(y) returns a "matrix".
# If not, let y=as.matrix(pharma.data[,2]).
variogram(y,length(y)/4)
```

2.3 Numerical Description of Time Series Data- Get Variogram by JMP I

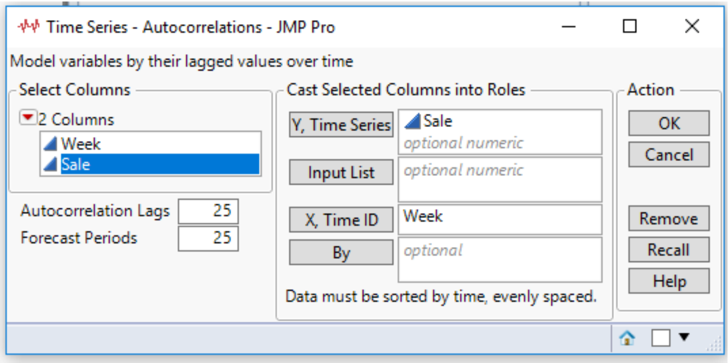
How to obtain Variogram by JMP:

1. JMP $\rightarrow$ File $\rightarrow$ Open (choose the excel file in your PC).
2. In the new opening window, select options based on your data, then click "Import".
3. See the three figures below.

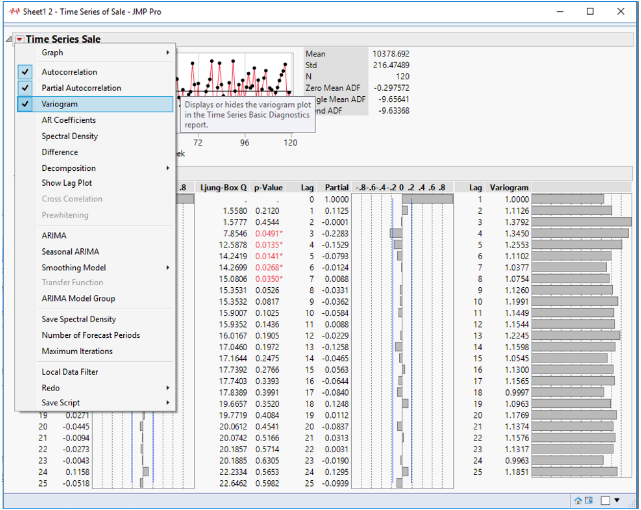
2.3 Numerical Description of Time Series Data- Get Variogram by JMP II



2.3 Numerical Description of Time Series Data- Get Variogram by JMP III



2.3 Numerical Description of Time Series Data- Get Variogram by JMP IV



In the sunspot example, we can see the variability changes with respect to time.

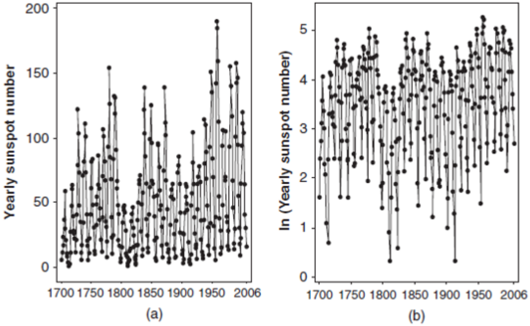


FIGURE 2.20 Yearly International Sunspot Number, (a) untransformed and (b) natural logarithm transformation. *Source*: SIDC.

This is usually due to **nonconstant variance** in the time series data.

2.4.1 Transformations I

To deal with the problem, we can apply transformations to the data. One commonly used type is the power family of transformations

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}, \lambda \neq 0 \\ \dot{y} \ln y, \lambda = 0 \end{cases} \tag{18}$$

where $\dot{y} = exp[(1/T) \sum_{t=1}^{T} \ln y_t]$ is the geometric mean of the observations (constant). The divisor $\dot{y}^{\lambda-1}$ is simply a **scale factor** that ensures that when different models are fit to investigate the utility of different transformations (values of λ), the residual sum of squares for these models can be meaningfully compared.

2.4.1 Transformations II

In particular, $\lambda = 0$ implies the log transformation. Because

$$\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \lim_{\lambda \to 0} \frac{y^\lambda lny}{1} = lny \tag{19}$$

$\lambda = 0.5$ imples a square root transformation.
$\lambda = -1$ implies an inverse transformation.
$\lambda = 1$ implies no transformation.

When the data variability increases with the average level of the series, the log transformation is often considered. The log transformation also has a very nice interpretation as percentage change.

Let us apply log transformation to the sunspot data, as shown in previous picture part (b). The variance is more stable now.

2.4.2 Trend and Seasonal Adjustment I

Two other widely used adjustment are **trend** adjustment and **seasonal** adjustment.

A time series with a trend is not stationary. Modeling and forecasting will be simplified if we can eliminate the trend. To do this, we can fit a regression model in terms of time and then subtract it from the original observations. After all, the residuals are free of trend.

The most common trend model is the linear trend

$$E(y_t) = \beta_0 + \beta_1 t.$$

It could also be a quadratic one

$$E(y_t) = \beta_0 + \beta_1 t + \beta_2 t^2$$

or an exponential one $E(y_t) = \beta e^{\beta_t t}$.

2.4.2 Trend and Seasonal Adjustment II

Consider the blue and gorgonzola cheese example, where there is a clear linear trend. The picture shows a fitted line that is constructed by least squares estimate, followed by pictures about residuals.

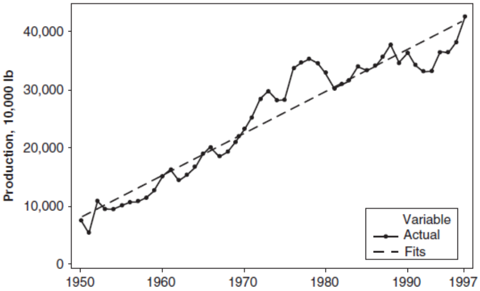


FIGURE 2.21 Blue and gorgonzola cheese production, with fitted regression line. *Source*: USDA–NASS.

2.4.2 Trend and Seasonal Adjustment III

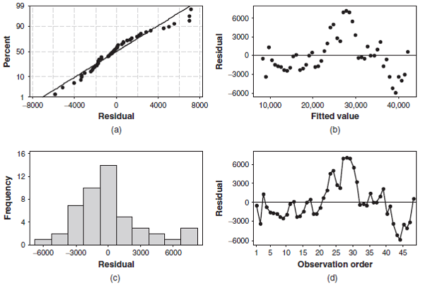


FIGURE 2.22 Residual plots for simple linear regression model of blue and gorgonzola cheese production.

Part (a) is normal probability plot from which we are looking for a straight line. Part (c) is a histogram where we are looking for normal shape. Part (b) is the scatter plot of residuals versus fitted values and Part (d) is residuals versus observations order, where we are looking for no patterns. In this case, (b) and (d) are the same. Why?

2.4.2 Trend and Seasonal Adjustment IV

Differencing Another approach to is to obtain a new time series, say x_t by differencing the data, i.e.

$$x_t = y_t - y_{t-1} = \nabla y_t, \tag{20}$$

where ∇ (read as *nabla*) is the difference operator.

An equivalent way to write the operator is to use the backshift operator $\mathbf{B}$ defined by

$$\mathbf{B}y_t = y_{t-1}. \tag{21}$$

Therefore we have

$$x_t = y_t - y_{t-1} = y_t - \mathbf{B}y_t = (1 - \mathbf{B})y_t \tag{22}$$

from which we can see $\nabla = (1 - \mathbf{B})$.

2.4.2 Trend and Seasonal Adjustment V

The operators can be performed successively if needed to remove the trend. For example,

$$x_t = \nabla^2 y_t = \nabla(\nabla y_t) = (1 - \mathbf{B})^2 y_t = (1 - 2\mathbf{B} + \mathbf{B}^2) y_t = y_t - 2y_{t-1} + y_{t-2} \tag{23}$$

where we used the fact that $\mathbf{B}^d y_t = y_{t-d}$.

2.4.2 Trend and Seasonal Adjustment VI

Features of differencing method:

1. It does not require estimation of any parameter.
2. It can allow the trend component to change through time, while regression assumes the trend is fixed.
3. Usually, one or two differences are sufficient.

We apply the differencing method to the blue and gorgonzola cheese example, and obtain the plot:

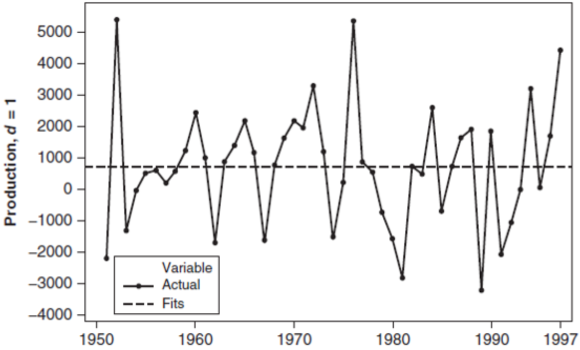


FIGURE 2.23 Blue and gorgonzola cheese production, with one difference. *Source*: USDA–NASS.

We can see the trend which was increasing has been removed.

2.4.2 Trend and Seasonal Adjustment VIII

The appearance of the residuals is improved when a linear model is fitted to detrend the time series. This illustrates that differencing may be a good alternative to detrending a time series by using a regression model.

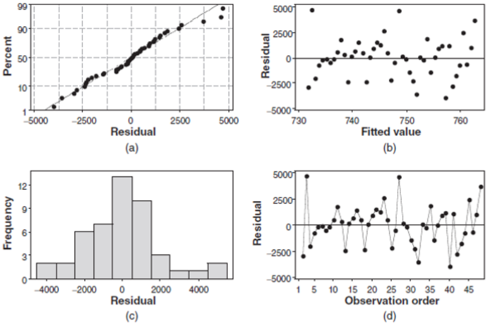


FIGURE 2.24 Residual plots for one difference of blue and gorgonzola cheese production.

2.4.2 Trend and Seasonal Adjustment IX

Seasonal components are present in many time series. We can use a **seasonal difference** to eliminate seasonality. Similarly, define:

$$\nabla_d y_t = (1 - \mathbf{B}^d) y_t = y_t - y_{t-d} \tag{24}$$

For example, $d = 12$ if we had monthly data with an annual season.

2.4.2 Trend and Seasonal Adjustment X

When **both trend and seasonality** are present, we can sequentially difference to eliminate them.

That is, apply ∇_d to remove seasonality then apply ∇ once or several times to remove the trend.

Consider the beverage shipment example which shows both trend and monthly cycles. The seasonal differencing result is shown in Figure 2.25 (a). Then differencing removes the trend, as shown in Figure 2.25 (b).

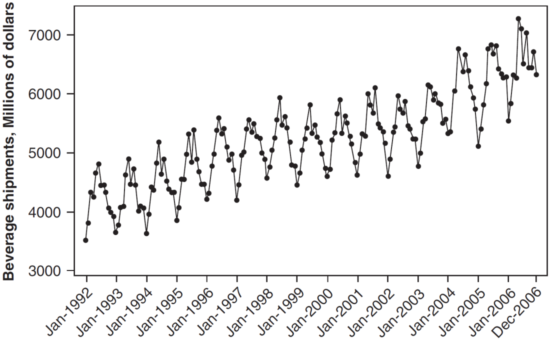


FIGURE 1.5 The US beverage manufacturer monthly product shipments, unadjusted. *Source:* US Census Bureau.

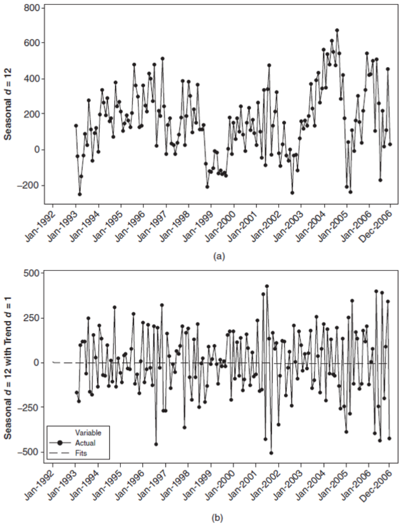


FIGURE 2.25 Time series plots of seasonal- and trend-differenced beverage data.

The residual plots seem to be acceptable.

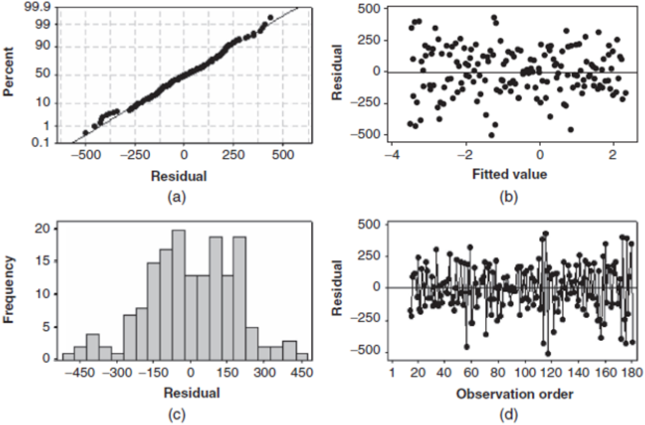


FIGURE 2.26 Residual plots for linear trend model of differenced beverage shipments.

2.4.2 Trend and Seasonal Adjustment XIV

Regression can also be used to eliminate seasonality from time series data. A useful model is

$$E(y_t) = \beta_0 + \beta_1 \sin \frac{2\pi}{d} t + \beta_2 \cos \frac{2\pi}{d} t, \tag{25}$$

where d is the length of the season (e.g. $d = 12$ for monthly data and annual season). The model can be written as a sine wave after some algebra.

A more general version of the model is obtained by adding more terms.

$$E(y_t) = \beta_0 + \sum_{j=1}^{K} \left(\beta_{1,j} \sin \frac{2\pi j}{d} t + \beta_{2,j} \cos \frac{2\pi j}{d} t \right) \tag{26}$$

These terms are called harmonics.

2.4.2 Trend and Seasonal Adjustment XV

It is also beneficial to introduce a "classical" approach to the **decomposition**. The general model for this is

$$y_t = f(S_t, T_t, \epsilon_t), \tag{27}$$

where S_t is the seasonal component, T_t is the trend effect, and ϵ_t is the random error. Two common forms are:

1. Additive Model

   $$y_t = S_t + T_t + \epsilon_t,$$

   which is used if the magnitude of the seasonal variation does not vary.

2. Multiplicative Model

   $$y_t = S_t T_t \epsilon_t,$$

   which is used if the magnitude of the seasonal fluctuations changes with respect to the average level of the time series.

2.4.2 Trend and Seasonal Adjustment XVI

For the **additive** model, the procedure includes:

1. Model trend, e.g. a simple linear model : $T_t = \beta_0 + \beta_1 t$.
2. Remove the trend (**detrend**):

$$y_t - T_t = S_t + \epsilon_t, \tag{28}$$

   notice $\{y_t - T_t\}$ is another time series **without** trend.
3. Compute the seasonal components by taking the average of all of the seasonal factors for each period in the season (**de-seasonality**). For example, subtract the mean or median of all Januaries from each observation in Januaries.

$$y_{\text{each obs in Jan}} - \bar{y}_{\text{all Januaries}}$$
$$y_{\text{each obs in Feb}} - \bar{y}_{\text{all Februaries}}$$
$$\vdots$$
$$y_{\text{each obs in Dec}} - \bar{y}_{\text{all Decembers}}$$

2.4.2 Trend and Seasonal Adjustment XVII

For **multiplicative** model, replace subtraction by division in step 2.

$$y_t / T_t = S_t \epsilon_t \tag{29}$$

Reconsider the beverage shipment example by using **decomposition** approach. The following pictures are the original data, detrended data, seasonally adjusted data, and the data when both adjustment are made.

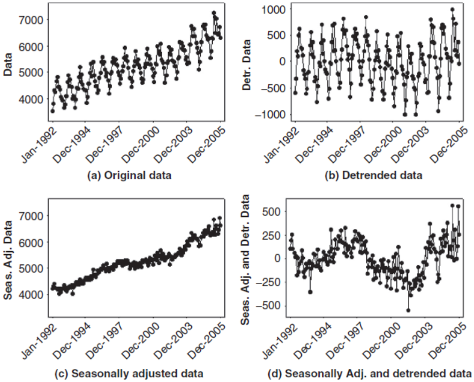


FIGURE 2.29 Component analysis of beverage shipments.

Notice that (d) is essentially the residual plot versus time. There could be a potential constant variance issue over time. In addition, let us look at a diagnose of the residuals:

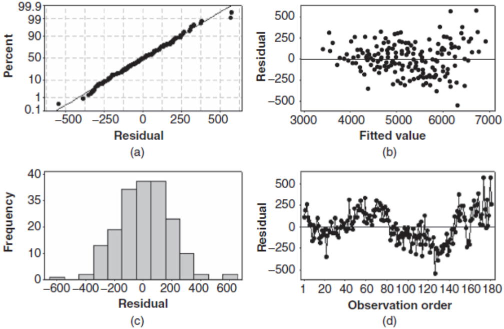


FIGURE 2.30 Residual plots for additive model of beverage shipments.

(d) is identical to the previous (d). Part (b) shows a **funnel** shape which indicates that a **log transformation** may be needed to stabilize the variance.

After applying a log transformation to the data, we repeat the previous process.

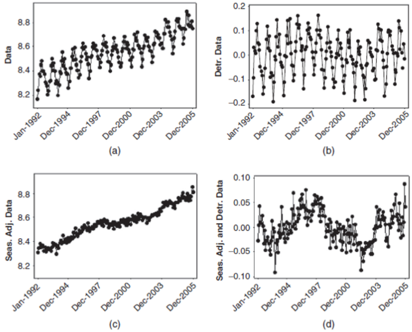


FIGURE 2.32 Component analysis of transformed beverage data.

And the diagnose:

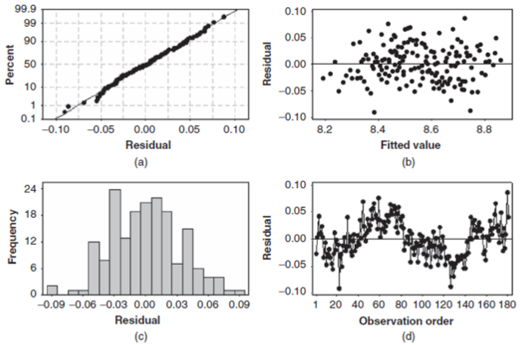


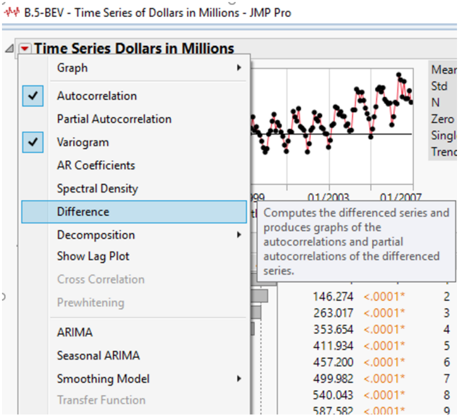
FIGURE 2.33 Residual plots from decomposition model for transformed beverage data.

Part (b) now indicates the stable variance. Part (d) still has a pattern over time. Notice that this was not a problem in Figure 2.26.

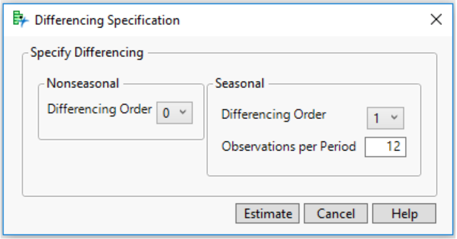
2.5 General Approach to Time Series Modeling and Forecasting

1. Plot the time series and visually determine its features, such as trend and seasonality.
2. Eliminate the trend or seasonal components, as well as consider potential data transformations.
3. Develop a forecasting model for the residuals. It is common that there are more than one plausible models.
4. Validate the performance of the model.
5. For forecasts, if transformations are made in step 2, remember to do the reverse transformations. Such as, log vs exponent, detrending vs adding trend back and etc.
6. Monitoring the forecast to ensure that deterioration of the model will be detected quickly, usually by evaluating the forecast errors.

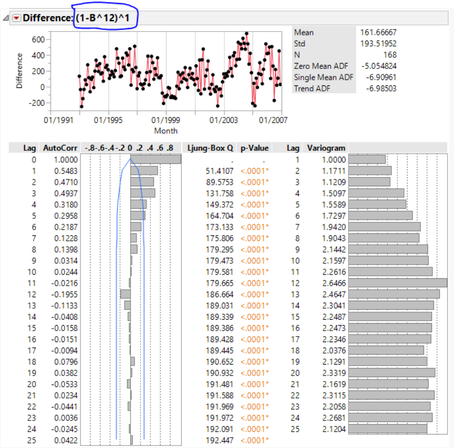
1. Click the red inverted triangle on the top-left corner, then select the **Difference**.



2. For nonseasinal data, select a number for Differencing Order from the drop-down menu in the Nonseasonal block. For seasonal data, select a number for Differencing order and Observations per Period in the block of Seasonal. For example, set the Differencing Order be 1 and the length of a period is 12.



3.The notation $(1 - B^\wedge 12)^\wedge 1$ implies that both differencing (d=1) and de-seasonality (d=12) have been applied to the data.New ACF and variogram were generated and presented.



Lag	AutoCorr	Ljung-Box Q	p-Value
0	1.0000	.	.
1	0.5483	51.4107	<.0001*
2	0.4710	89.5753	<.0001*
3	0.4937	131.758	<.0001*
4	0.3180	149.372	<.0001*
5	0.2958	164.704	<.0001*
6	0.2187	173.133	<.0001*
7	0.1228	175.806	<.0001*
8	0.1398	179.295	<.0001*
9	0.0314	179.473	<.0001*
10	0.0244	179.581	<.0001*
11	-0.0216	179.665	<.0001*
12	-0.1955	186.664	<.0001*
13	-0.1133	189.031	<.0001*
14	-0.0408	189.339	<.0001*
15	-0.0158	189.386	<.0001*
16	-0.0151	189.428	<.0001*
17	-0.0094	189.445	<.0001*
18	0.0796	190.652	<.0001*
19	0.0382	190.932	<.0001*
20	-0.0533	191.481	<.0001*
21	0.0234	191.588	<.0001*
22	-0.0441	191.969	<.0001*
23	0.0036	191.972	<.0001*
24	-0.0245	192.091	<.0001*
25	0.0422	192.447	<.0001*

Lag	Variogram
1	1.0000
2	1.1711
3	1.1209
4	1.5097
5	1.5589
6	1.7297
7	1.9420
8	1.9043
9	2.1442
10	2.1597
11	2.2616
12	2.6466
13	2.4647
14	2.3041
15	2.2487
16	2.2473
17	2.2346
18	2.0376
19	2.1291
20	2.3319
21	2.1619
22	2.3115
23	2.2058
24	2.2681
25	2.1204

2.6 Evaluating and Monitoring Forecasting Model Performance

There are many measures about how well a model is fitted. They often use the residual and do not reflect the capability of the forecast. Measures of forecast accuracy should always be evaluated.

2.6.1 Forecasting Model Evaluation I

Recall the one-step-ahead forecast error, (Error = Observation - Forecast)

$$e_t(1) = y_t - \hat{y}_t(t-1)$$

mean error or average error:

$$ME = \frac{1}{n}\sum_{t=1}^{n} e_t(1)$$

mean absolute deviation:

$$MAD = \frac{1}{n}\sum_{t=1}^{n} |e_t(1)|$$

mean squared error:

$$MSE = \frac{1}{n}\sum_{t=1}^{n} [e_t(1)]^2$$

2.6.1 Forecasting Model Evaluation II

For ME, we are looking for zero, that is the forecast produces **unbiased** forecasts. If the ME differs from zero significantly, **bias** is indicated. It is also possible that the ME is closed to zero in the beginning, then drifts away after the technique is used for a while. This can be an indication that the series has changed in some fashion.

The MSE is an estimator of the variance of the one-step-ahead forecast errors:

$$\hat{\sigma}^2_{e(1)} = MSE.$$

If the forecast errors are normally distributed, which is the assumption we usually want them to satisfy, MAD and standard deviation of forecast errors satisfy

$$\hat{\sigma}_{e(1)} = \sqrt{\frac{\pi}{2}} MAD \cong 1.25 MAD$$

2.6.1 Forecasting Model Evaluation III

Notice that the formulas indicate that these measures have the same unit as data, so we might not know whether a value is relatively large or small. To fix this, we need a measure of **relative forecast error** (in percent):

$$re_t(1) = 100\frac{y_t - \hat{y}_t(t-1)}{y_t} = 100\frac{e_t(1)}{y_t}$$

mean percent forecast error

$$MPE = \frac{1}{n}\sum_{i=1}^{n} re_t(1)$$

mean absolute percent forecast error.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} |re_t(1)|$$

Example 2.10

Table 2.2 illustrates the calculation of the one-step-ahead forecast error, the absolute errors, the squared errors, the relative (percent) error, and the absolute percent error from a forecasting model for 20 time periods.

The last row of columns (3) through (7) display the sums required to Calculate the ME, MAD, MSE, MPE and MAPE.

TABLE 2.2 Calculation of Forecast Accuracy Measures

Time Period	(1) Observed Value y_t	(2) Forecast $\hat{y}_t(t-1)$	(3) Forecast Error $e_t(1)$	(4) Absolute Error $\lvert e_t(1)\rvert$	(5) Squared Error $[e_t(1)]^2$	(6) Relative (%) Error $(e_t(1)/y_t)\,100$	(6) Absolute (%) Error $\lvert(e_t(1)/y_t)\,100\rvert$
1	47	51.1	−4.1	4.1	16.81	−8.7234	8.723404
2	46	52.9	−6.9	6.9	47.61	−15	15
3	51	48.8	2.2	2.2	4.84	4.313725	4.313725
4	44	48.1	−4.1	4.1	16.81	−9.31818	9.318182
5	54	49.7	4.3	4.3	18.49	7.962963	7.962963
6	47	47.5	−0.5	0.5	0.25	−1.06383	1.06383
7	52	51.2	0.8	0.8	0.64	1.538462	1.538462
8	45	53.1	−8.1	8.1	65.61	−18	18
9	50	54.4	−4.4	4.4	19.36	−8.8	8.8
0	51	51.2	−0.2	0.2	0.04	−0.39216	0.392157
11	49	53.3	−4.3	4.3	18.49	−8.77551	8.77551
12	41	46.5	−5.5	5.5	30.25	−13.4146	13.41463
13	48	53.1	−5.1	5.1	26.01	−10.625	10.625
14	50	52.1	−2.1	2.1	4.41	−4.2	4.2
15	51	46.8	4.2	4.2	17.64	8.235294	8.235294
16	55	47.7	7.3	7.3	53.29	13.27273	13.27273
17	52	45.4	6.6	6.6	43.56	12.69231	12.69231
18	53	47.1	5.9	5.9	34.81	11.13208	11.13208
19	48	51.8	−3.8	3.8	14.44	−7.91667	7.916667
20	52	45.8	6.2	6.2	38.44	11.92308	11.92308
	Totals		*−11.6*	*86.6*	*471.8*	*−35.1588*	*177.3*

2.6.1 Forecasting Model Evaluation VI

$$ME = \frac{1}{n}\sum_{t=1}^{n} e_t(1) = \frac{1}{20}(-11.6) = -0.58$$

$$MAD = \frac{1}{n}\sum_{t=1}^{n} |e_t(1)| = \frac{1}{20}(86.6) = 4.33$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t(1)^2 = \frac{1}{20}(471.8) = 23.59$$

$$\hat{\sigma}^2_{e(1)} = MSE = 23.59, \ \hat{\sigma}_{e(1)} = \sqrt{MSE} = 4.86$$

$$\hat{\sigma}_{e(1)} \cong 1.25 MAD = 1.25(4.33) = 5.41$$

$$MPE = \frac{1}{n}\sum_{t=1}^{n} re_t(1) = \frac{1}{20}(-35.1588) = -1.76\%$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} |re_t(1)| = \frac{1}{20}(177.3) = 8.87\%$$

One-Step-Ahead forecast errors and the sample ACF of the forecast errors are shown below:

TABLE 2.3 One-Step-Ahead Forecast Errors

Period, t	$e_t(1)$	Period, t	$e_t(1)$	Period, t	$e_t(1)$	Period, t	$e_t(1)$	Period, t	$e_t(1)$
1	−0.62	11	−0.49	21	2.90	31	−1.88	41	−3.98
2	−2.99	12	4.13	22	0.86	32	−4.46	42	−4.28
3	0.65	13	−3.39	23	5.80	33	−1.93	43	1.06
4	0.81	14	2.81	24	4.66	34	−2.86	44	0.18
5	−2.25	15	−1.59	25	3.99	35	0.23	45	3.56
6	−2.63	16	−2.69	26	−1.76	36	−1.82	46	−0.24
7	3.57	17	3.41	27	2.31	37	0.64	47	−2.98
8	0.11	18	4.35	28	−2.24	38	−1.55	48	2.47
9	0.59	19	−4.37	29	2.95	39	0.78	49	0.66
10	−0.63	20	2.79	30	6.30	40	2.84	50	0.32

TABLE 2.4 Sample ACF of the One-Step-Ahead Forecast Errors in Table 2.3

Lag	Sample ACF, r_k	Z-Statistic	Ljung–Box Statistic, Q_{LB}
1	0.004656	0.03292	0.0012
2	−0.102647	−0.72581	0.5719
3	0.136810	0.95734	1.6073
4	−0.033988	−0.23359	1.6726
5	0.118876	0.81611	2.4891
6	0.181508	1.22982	4.4358
7	−0.039223	−0.25807	4.5288
8	−0.118989	−0.78185	5.4053
9	0.003400	0.02207	5.4061
10	0.034631	0.22482	5.4840
11	−0.151935	−0.98533	7.0230
12	−0.207710	−1.32163	9.9749
13	0.089387	0.54987	10.5363

No ACF value is significantly away from 0.

If the forecast errors are structureless, the sample ACF of them should look like the ACF of random data. If there are spikes on the sample ACF at low lag, then the forecasts can be improved to take the structure out of the errors.

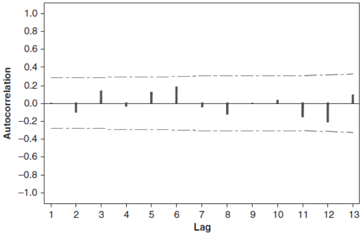


FIGURE 2.36 Sample ACF of forecast errors from Table 2.4.

2.6.1 Forecasting Model Evaluation IX

If a time series consists of uncorrelated observations and has constant variance, we say it is white noise. Notice the ACF of white noise vanishes when $k \neq 0$. In addition, if a white noise has normal distribution, we say it is Gaussian white noise.

There is much empirical evidence that the distribution of forecast errors is approximately normal. To check the normality, one thing we can do is to plot the Normal Probability plot.

The normal probability plot is shown below. The horizontal axis is the ordered errors, and the vertical axis is the ranks of the errors. Also notice that the scales are adjusted according to normal distribution.

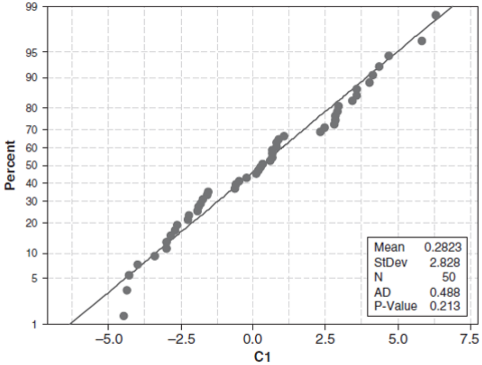


FIGURE 2.37 Normal probability plot of forecast errors from Table 2.3.

from which we are hoping for a line.

2.6.1 Forecasting Model Evaluation XI

If a time series is white noise, the distribution of the sample autocorrelation coefficient at lag k in large sample is approximately normal with mean zero and variance $\frac{1}{T}$,

$$r_k \sim N(0, \frac{1}{T}) \tag{30}$$

Therefore we could test the hypothesis

$$H_0 : \rho_k = 0 \qquad H_a : \rho_k \neq 0$$

using the test statistic

$$Z_0 = \frac{r_k}{\sqrt{\frac{1}{T}}} = r_k\sqrt{T}. \tag{31}$$

2.6.1 Forecasting Model Evaluation XII

Go back to Table 2.4, column three and four show two sets of statistics.

The Z-Statistic is for the hypothesis $\mathbf{H}_0 : \rho_k = 0$. At significance level $\alpha = 0.05$, none of the Zs is better than the critical value $|Z| = 1.96$ (R code: qnorm(1-0.05/2)). So we can not reject any individual ACF value is 0.

2.6.1 Forecasting Model Evaluation XIII

The last column of Table 2.4 is a Ljung-Box goodness-of-fit statistic

$$Q_{LB} = T(T+2)\sum_{i=1}^{K}\left(\frac{1}{T-i}\right) r_i^2 \sim \chi^2_{\alpha,K-p}. \tag{32}$$

where p is the number of parameters in the model, so the number of degrees of freedom in the chi-square distribution becomes $k-p$. Q_{LB} gives the group test of ACF.

If the value at row j is bigger than the critical value $\chi^2_{0.05,j}$, then we reject the null hypothesis that all ACF values up to j are 0.
In particular,at lag $k = 13$, $Q_{LB,13} = 10.5363$,

$$\chi^2_{0.05,13} = qchisq(1-0.05, 13) = 22.3620 > 10.5363. \tag{33}$$

There is no strong evidence to indicate that the first 13 autocorrelations of the forecast errors considered jointly differ from zero.

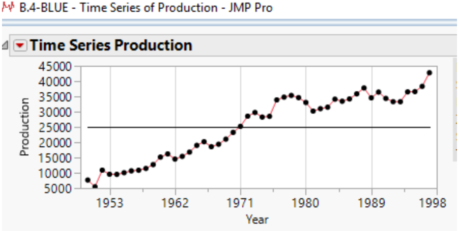
2.6.1 Forecasting Model Evaluation XIV

If we calculate the p-value for this test statistic, we find that

$$p = P(\chi^2 > 10.5363) = 1 - pchisq(10.5363, 13) = 0.65. \tag{34}$$

This is a good indication that the forecast errors are white noise.

The Ljung-Box statistic Q and p-vaules come with the ACF in JMP. Small *p*-values indicate significant autocorrelation.

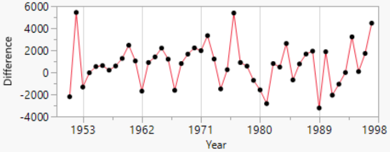


Time Series Basic Diagnostics

Lag	AutoCorr	Ljung-Box Q	p-Value
0	1.0000	.	.
1	0.9226	43.4611	<.0001*
2	0.8502	81.1785	<.0001*
3	0.7997	115.285	<.0001*
4	0.7433	145.421	<.0001*
5	0.6926	172.197	<.0001*
6	0.6466	196.091	<.0001*
7	0.5921	216.613	<.0001*
8	0.5265	233.244	<.0001*
9	0.4652	246.562	<.0001*
10	0.3987	256.601	<.0001*
11	0.3439	264.271	<.0001*

2.6.1 Forecasting Model Evaluation XVI

Difference: (1-B)^1



Mean	747.14894
Std	1902.6693
N	47
Zero Mean ADF	-5.963179
Single Mean ADF	-7.019626
Trend ADF	-6.963266

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Variogram
0	1.0000	.	.	1	1.0000
1	-0.0683	0.2335	0.6290	2	0.9880
2	-0.0555	0.3910	0.8224	3	1.0071
3	-0.0759	0.6926	0.8749	4	0.9022
4	0.0362	0.7627	0.9434	5	1.1325
5	-0.2098	3.1771	0.6727	6	0.8490
6	0.0930	3.6628	0.7222	7	0.9243
7	0.0126	3.6719	0.8167	8	0.9436
8	-0.0080	3.6757	0.8851	9	1.1069
9	-0.1825	5.6938	0.7701	10	0.9680
10	-0.0341	5.7663	0.8345	11	0.8058
11	0.1392	7.0061	0.7986	12	0.7773
12	0.1697	8.9004	0.7114	13	0.9765
13	-0.0431	9.0264	0.7709	14	0.9697
14	-0.0359	9.1165	0.8235	15	1.1011
15	-0.1762	11.3521	0.7273	16	1.0344
16	-0.1050	12.1714	0.7321	17	1.0084
17	-0.0773	12.6297	0.7606	18	0.8411
18	0.1015	13.4480	0.7643	19	1.0109
19	-0.0799	13.9731	0.7853	20	0.9953
20	-0.0633	14.3149	0.8142	21	0.9302
21	0.0063	14.3184	0.8555	22	1.0057
22	-0.0744	14.8279	0.8695	23	0.9989
23	0.0671	15.2600	0.8851	24	0.7338

2.6.2 Choosing Between Competing Models I

Here we discuss some principles of model selection from goodness of fit point of view.

Danger of **overfitting**. Overfitting is "the production of an analysis that corresponds too closely or exactly to a particular set of data (such as historical data), and may therefore fail to fit additional data or predict future observations reliably".

In general, the best approach is to select the model that results in the smallest standard deviation (or mean squared error) of the one-step-ahead forecast errors when the model is applied to data that were not used in the fitting process. Some authors refer to this as an **out- of-sample** forecast error standard deviation (or mean squared error).

2.6.2 Choosing Between Competing Models II

A standard way to measure this out-of-sample performance is by utilizing some form of **data splitting**; that is, divide the time series data into two segments:

- Model Fitting
- Performance Testing

Sometimes data splitting is called **cross-validation**. It is somewhat arbitrary as to how the data splitting is accomplished. However, a good rule of thumb is to have at least 20 or 25 observations in the performance testing data set.

2.6.2 Choosing Between Competing Models III

When evaluating the fit of the model to **historical** data, there are several criteria: MSE, R^2, Adjusted R^2, AIC, BIC, corrected AIC (AIC_c), etc.

2.6.2 Choosing Between Competing Models IV

- Mean Squared Error (MSE) of the residuals:

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T - p},$$

T is the number of periods.
p is the number of parameters.
e_t is the residual from the model-fitting process in period t.

The MSE is the sample variance of the residuals and it is an estimator of the variance of the model errors. We are looking for a small MSE in general.

2.6.2 Choosing Between Competing Models V

- R-Square

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

Notice that the denominator is constant (not model dependent) when the data is given regardless the model, and the numerator is just the numerator of MSE. Therefore, when sum of the squared residuals is minimized, the R-squared attains maximum. Because of this, we are looking for a large R-square in general.

However, a problem with R-square is that it increases when parameters(even bad ones) are added to the model. So **relying on R-square brings overfitting**. In addition, a large R-square does not ensure that the out-of-sample one-step-ahead forecast errors will be small.

2.6.2 Choosing Between Competing Models VI

- Adjusted R-Square

$$R^2_{Adj} = 1 - \frac{\sum_{t=1}^{T} e_t^2/(T-p)}{\sum_{t=1}^{T}(y_t - \bar{y})^2/(T-1)}.$$

The adjustment is a size adjustmentthat is, adjust for the number of parameters in the model. Note that a model that maximizes the adjusted R^2 statistic is also the model that minimizes the residual mean square. Here the "adjusted" means that the number of parameter is considered meaning that a large p will result in a small R^2_{Adj} if other quantities remain the same.

Notice the numerator is just MSE, so maximizing the adjusted R-square is equivalent to minimizing the MSE.

- Akaike Information Criterion (AIC), Schwarz Bayesian Information Criterion (BIC) and corrected AIC (AICc)

$$AIC = \ln\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right) + \frac{2p}{T},$$

$$BIC = \ln\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right) + \frac{p\ln(T)}{T}$$

and

$$AICc = \ln\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right) + \frac{2T(p+1)}{T-p-2}$$

These three criteria penalize the sum of squared residuals for including additional parameters in the model. Models that have small values of the AIC or BIC are considered good models.

2.6.2 Choosing Between Competing Models VIII

One way to evaluate model selection criteria is in terms of **consistency**. A model selection criterion is consistent if it selects the true model when the true model is among those considered with probability approaching unity as the sample size becomes large, and if the true model is not among those considered, it selects the best approximation with probability approaching unity as the sample size becomes large. It turns out that s^2, the adjusted R^2, and the AIC are all inconsistent, because they do not penalize for adding parameters heavily enough. Relying on these criteria tends to result in overfitting. The BIC, which caries a heavier size adjustment penalty, is consistent.

2.6.2 Choosing Between Competing Models IX

An **asymptotically efficient** model selection criterion chooses a sequence of models as T(the amount of data available) gets large for which the one-step-ahead forecast error variances approach the one-step-ahead forecast error variance for the true model at least as fast as any other criterion. The AIC is asymptotically efficient but the BIC is not.

The AIC is a biased estimator of the discrepancy between all candidate models and the true model. This has led to developing a corrected version of AIC_c.

2.6.2 Choosing Between Competing Models X

When both AIC and BIC are available, we prefer using BIC. It generally results in smaller, and hence simpler, models, and so its use is consistent with the time-honored model-building principle of parsimony.

2.6.3 Monitoring a Forecast Model I

The one-step-ahead forecast error $e_t(1)$ are typically used for forecast monitoring.

Shewhart Control Chart: a plot of the forecast errors versus time containing a center line and a set of control limits. The center line is usually taken as either zero or the average forecast error and the control limits are placed at three standard deviations away from the center line.

- Individual Control Chart
- Moving Range Control Chart

2.6.3 Monitoring a Forecast Model II

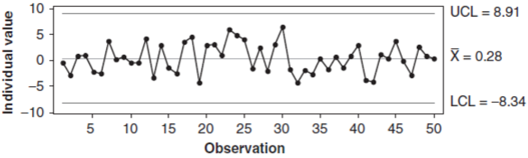
Two other types of control charts are:

- **Cumulative Sum (CUSUM)**
- **Exponentially Weighted Moving Average (EWMA)**

These charts are more effective at detecting smaller changers or disturbances in the forecasting model performance than the individuals control chart.

2.6.3 Monitoring a Forecast Model III

For **Individual Control Chart**, the center line is at the average of the forecast errors, $ME = \frac{1}{n}\sum_{t=1}^{n} e_t(1)$. The upper and lower control limits are $3\hat{\sigma}_{e(1)}$ away from the center line.



2.6.3 Monitoring a Forecast Model IV

Define the Moving Range as $|e_t(1) - e_{t-1}(1)|$ and the Moving Range for n observations as

$$MR = \sum_{t=2}^{n} |e_t(1) - e_{t-1}(1)|.$$

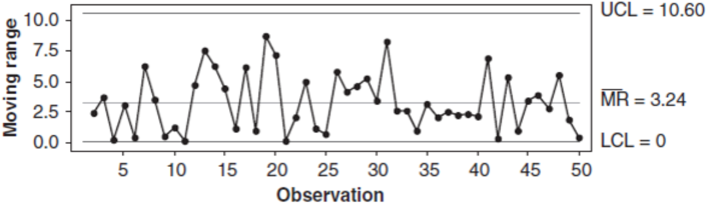
The estimate of the standard deviation of the one-step-ahead forecast errors is

$$\hat{\sigma}_{e(1)} = \frac{\sqrt{\pi}}{2} \cdot \frac{MR}{n-1} = \frac{0.8865 MR}{n-1} = 0.8865 \overline{MR}$$

where $\overline{MR}$ is the average of the moving ranges.

2.6.3 Monitoring a Forecast Model V

For the **Moving Range Control Chart**, the center line is at $\overline{MR}$, the upper control limit is at $3.267MR/(n-1)$, and the lower control limit is at zero.



The data for this example is in Table 2.3. All of the one-step-ahead forecast errors plot within the limits. We would conclude that there is no strong evidence of statistical inadequacy in the forecasting model.

2.6.3 Monitoring a Forecast Model VI

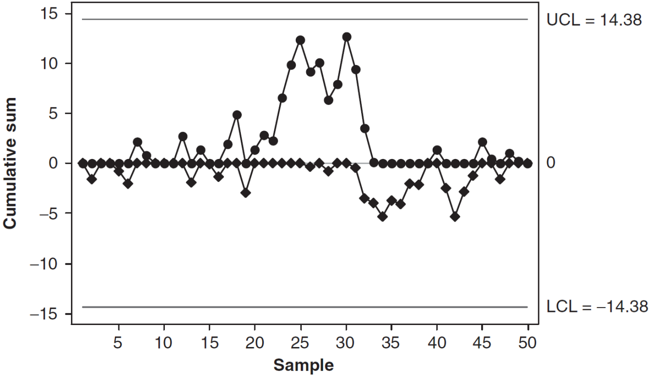


FIGURE 2.39 CUSUM control chart of the one-step-ahead forecast errors in Table 2.3.

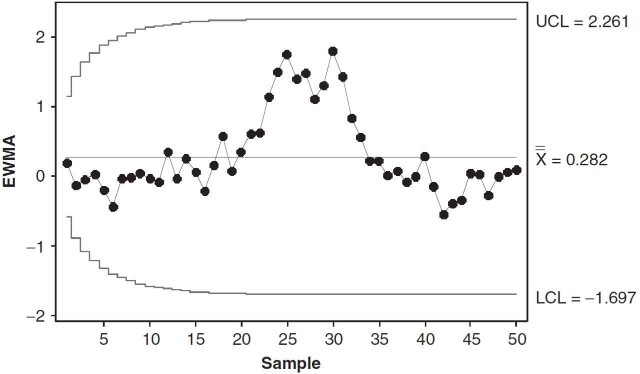
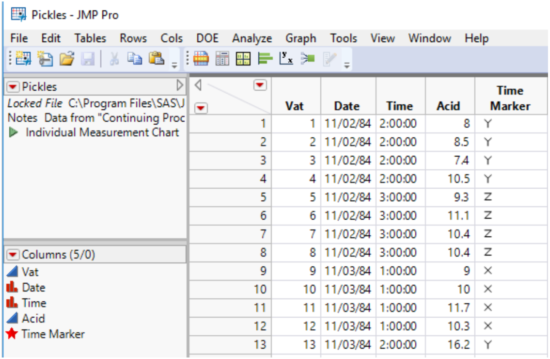


FIGURE 2.40 EWMA control chart of the one-step-ahead forecast errors in Table 2.3.

Control Chart in JMP I

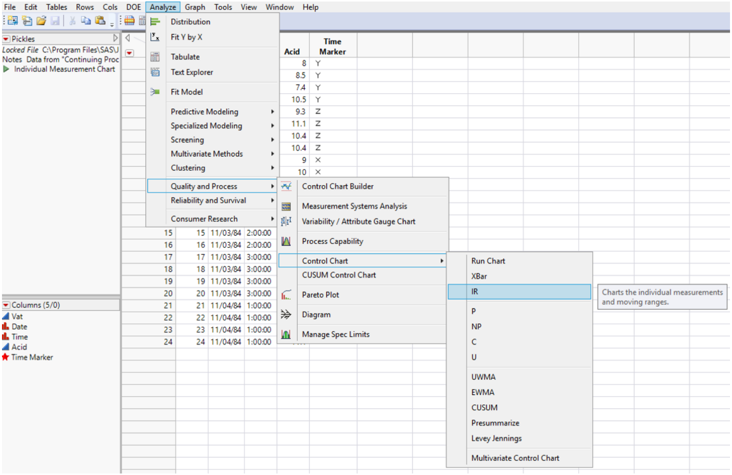
1. **Help** $\rightarrow$ **Sample Data Library** and open the file **Quality Control**, then select **Pickles.jmp**.Or import your own data.



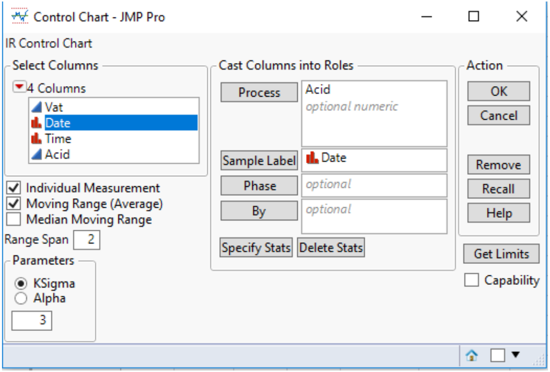
	Vat	Date	Time	Acid	Time Marker
1	1	11/02/84	2:00:00	8	Y
2	2	11/02/84	2:00:00	8.5	Y
3	3	11/02/84	2:00:00	7.4	Y
4	4	11/02/84	2:00:00	10.5	Y
5	5	11/02/84	3:00:00	9.3	Z
6	6	11/02/84	3:00:00	11.1	Z
7	7	11/02/84	3:00:00	10.4	Z
8	8	11/02/84	3:00:00	10.4	Z
9	9	11/03/84	1:00:00	9	X
10	10	11/03/84	1:00:00	10	X
11	11	11/03/84	1:00:00	11.7	X
12	12	11/03/84	1:00:00	10.3	X
13	13	11/03/84	2:00:00	16.2	Y

Control Chart in JMP II

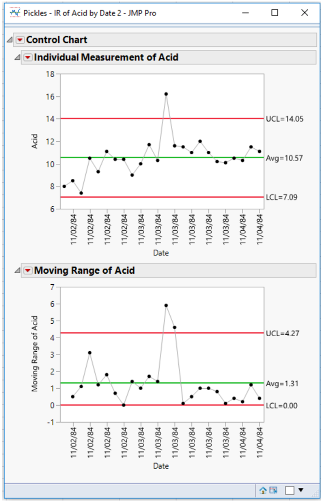
2. **Analyze → Quality and Process → Control Chart → IR.**



3. Select both **Individual Measurement** and **Moving Range chart** types.
4. Select Acid and click **Process**.
5. Select Date and click **Sample Label**.



6. Click **OK**.


Pickles - IR of Acid by Date 2 - JMP Pro
Control Chart
Individual Measurement of Acid
Acid
UCL=14.05
Avg=10.57
LCL=7.09
Date
Moving Range of Acid
Moving Range of Acid
UCL=4.27
Avg=1.31
LCL=0.00
Date

Control Charts in R

The individuals chart requires a simple vector of data. The moving range chart needs a two-column matrix arranged so that qcc() can calculate the moving range from each row. Let's use the forecast errors data in Example 2.25 as an example.

```
library(qcc)
# Generating the chart for individuals
#fe2.data[,2] is the forecast errors e_t(1), not the original y_t observations
ind_chart <- qcc(fe2.data[,2], type = "xbar.one",
                 title = "Individuals Chart for the Forecast Error", plot = TRUE)
# Creating the moving range chart and qcc object.
# qcc takes a two-column matrix that is used to calculate the moving range.
MA_obj <- matrix(cbind(fe2.data[,1:(length(fe2.data[,2])-1)], fe2.data[,2:length(fe2.data[,2])]), ncol=2)
MA_chart <- qcc(MA_obj, type="R", plot = TRUE)
```