Chapter 470

The Box-Jenkins Method

Introduction

Box - Jenkins Analysis refers to a systematic method of identifying, fitting, checking, and using integrated autoregressive, moving average (ARIMA) time series models. The method is appropriate for time series of medium to long length (at least 50 observations).

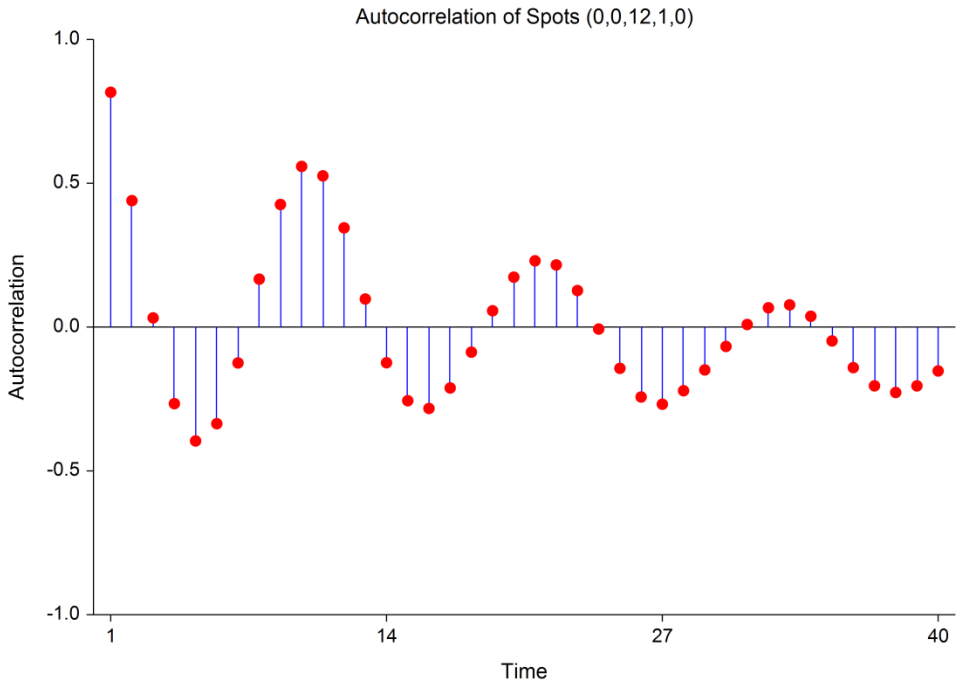
In this chapter we will present an overview of the Box-Jenkins method, concentrating on the how-to parts rather than on the theory. Most of what is presented here is summarized from the landmark book on time series analysis written by George Box and Gwilym Jenkins (1976).

A time series is a set of values observed sequentially through time. The series may be denoted by $X_1, X_2, \cdots, X_t$, where *t* refers to the time period and *X* refers to the value. If the X's are exactly determined by a mathematical formula, the series is said to be *deterministic*. If future values can be described only by their probability distribution, the series is said to be a *statistical* or *stochastic* process.

A special class of stochastic processes is a *stationary stochastic process*. A statistical process is stationary if the probability distribution is the same for all starting values of *t*. This implies that the mean and variance are constant for all values of *t*. A series that exhibits a simple trend is not stationary because the values of the series depend on *t*. A stationary stochastic process is completely defined by its mean, variance, and autocorrelation function. One of the steps in the Box - Jenkins method is to transform a non-stationary series into a stationary one.

Autocorrelation Function

The stationary assumption allows us to make simple statements about the correlation between two successive values, X_t and X_{t+k}. This correlation is called the *autocorrelation of lag k* of the series. The autocorrelation function displays the autocorrelation on the vertical axis for successive values of k on the horizontal axis. The following figure shows the autocorrelation function of the sunspot data.



Since a stationary series is completely specified by its mean, variance, and autocorrelation function, one of the major (and most subjective) tasks in Box-Jenkins analysis is to identify an appropriate model from the sample autocorrelation function. Although the sample autocorrelations contains random fluctuations, for moderate sample sizes they are fairly accurate in signaling the order of the ARIMA model.

The ARMA Model

The ARMA (autoregressive, moving average) model is defined as follows:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

where the ϕ's (phis) are the autoregressive parameters to be estimated, the $\theta's$ θ's (thetas) are the moving average parameters to be estimated, the *X's* are the original series, and the *a's* are a series of unknown random errors (or residuals) which are assumed to follow the normal probability distribution.

Box-Jenkins use the backshift operator to make writing these models easier. The backshift operator, *B*, has the effect of changing time period *t* to time period *t-1*. Thus $BX_t = X_{t-1}$ and $B^2 X_t = X_{t-2}$. Using this backshift notation, the above model may be rewritten as:

$$\left(1 - \phi_t B - \cdots - \phi_p B^p\right) X_t = \left(1 - \theta_1 B - \cdots - \phi_q B^q\right) a_t$$

This may be abbreviated even further by writing:

$$\phi_p(B)X_t = \theta_q(B)a_t$$

where

$$\phi_p(B) = \left(1 - \phi_1 B - \cdots - \phi_p B^p\right)$$

$$\theta_q(B) = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right)$$

These formulas show that the operators $\phi_p(B)$ and $\theta_q(B)$ are polynomials in *B* of orders *p* and *q* respectively. One of the benefits of writing models in this fashion is that we can see why several models may be equivalent.

For example, consider the model

$$X_t = 0.8X_{t-1} - 0.15X_{t-2} + a_t - 0.3a_{t-1}$$

This could be rewritten in the form of (8.3) as:

$$(1 - 0.8B + 0.15B^2)X_t = (1 - 0.3B)a_t$$

Notice that the polynomial on the left may be factored, so that we can rewrite the model as

$$(1 - 0.5B)(1 - 0.3B)X_t = (1 - 0.3B)a_t$$

Finally, canceling the *(1 - 0.3B)* from both sides leaves the simpler, but equivalent, model

$$(1 - 0.5B)X_t = a_t$$

or

$$X_t = 0.5X_{t-1} + a_t$$

Note that this is a much simpler model!

This type of model rearrangement is used by experienced Box-Jenkins forecasters to obtain the simplest models possible. The Theoretical ARIMA program displays the roots of the two polynomials, $\phi_p(B)$ and $\theta_q(B)$, so you can see possible model simplifications.

Nonstationary Models

Many time series encountered in practice exhibit nonstationary behavior. Usually, the nonstationarity is due to a trend, a change in the local mean, or seasonal variation. Since the Box-Jenkins methodology is for stationary models only, we have to make some adjustments before we can model these nonstationary series.

We use one of two methods for reducing a nonstationary series with trend to a stationary series (without trend):

1. Use the first differences of the series, $W_t = X_t - X_{t-1}$. Note that this can be rewritten as $W_t = (1 - B)X_t$. A more general form of this equation is:

$$\phi_p(B)(1 - B)^d X_t = \theta_q(B)a_t$$

   where *d* is the order of differencing. This is known as the *ARIMA(p,d,q)* model.

2. Fit a least squares trend and fit the Box-Jenkins model to the residuals.

If the model exhibits an occasional change of mean, first differences will result in a stationary model.

For seasonal series, Box-Jenkins provided a modification to this equation that will be the subject of the next section.

Seasonal Time Series

To deal with series containing seasonal fluctuations, Box-Jenkins recommend the following general model:

$$\phi_p(B)\Phi_P(B)(1 - B)^d(1 - B^s)^D X_t = \theta_q(B)\Theta_Q(B^s)a_t$$

where *d* is the order of differencing, *s* is the number of seasons per year, and *D* is the order of seasonal differencing. The operator polynomials are

$$\phi_p(B) = \left(1 - \phi_1 B - \cdots - \phi_p B^p\right)$$

$$\theta_q(B) = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right)$$

$$\Phi_P(B^s) = \left(1 - \Phi_1 B^s - \cdots - \Phi_p B^{sp}\right)$$

$$\Theta_Q(B^s) = \left(1 - \Theta_1 B^s - \cdots - \Theta_Q B^{sQ}\right)$$

Note that $(1 - B^s)X_t = X_t - X_{t-s}$.

Box-Jenkins explain that the maximum value of *d, D, p, q, P,* and *Q* is two. Hence, these operator polynomials are usually simple expressions.

Partial Autocorrelation Function

We previously discussed the autocorrelation function, which gives the correlations between different lags of a series. The Partial Autocorrelation Function is a second function that expresses information useful in determining the order of an ARIMA model.

This function is constructed by calculating the partial correlation between X_t and X_{t-1}, X_t and X_{t-2}, and so on, statistically adjusting out the influence of intermediate lags. For example, the partial autocorrelation of lag four is the partial correlation between X_t and X_{t-4} after statistically removing the influence of X_{t-1}, X_{t-2}, and X_{t-3} from both X_t and X_{t-4}.

The autoregressive order, *p*, is estimated as the lag of the last large partial autocorrelation. For example, suppose the partial autocorrelations were

Lag	Partial Autocorrelation
1	0.55
2	0.21
3	0.11
4	0.72
5	0.06
6	0.09
7	0.13

We would conclude that a reasonable value for *p* is four, since the partial autocorrelations are relatively small after the fourth lag.

Box-Jenkins Methodology – An Overview

The Box-Jenkins method refers to the iterative application of the following three steps:

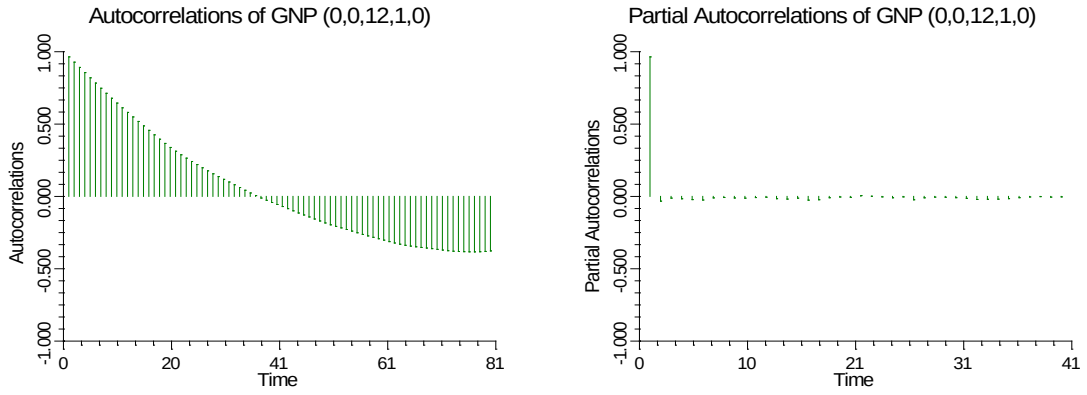
1. ***Identification***. Using plots of the data, autocorrelations, partial autocorrelations, and other information, a class of simple ARIMA models is selected. This amounts to estimating appropriate values for *p*, *d*, and *q*.
2. ***Estimation***. The phis and thetas of the selected model are estimated using maximum likelihood techniques, backcasting, etc., as outlined in Box-Jenkins (1976).
3. ***Diagnostic Checking***. The fitted model is checked for inadequacies by considering the autocorrelations of the residual series (the series of residual, or error, values).

These steps are applied iteratively until step three does not produce any improvement in the model. We will now go over these steps in detail.

Model Identification

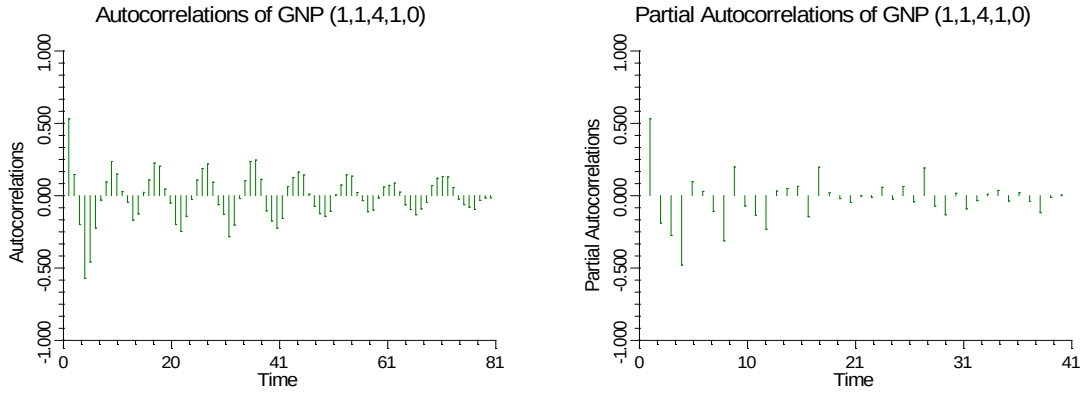
Assuming for the moment that there is no seasonal variation, the objective of the model identification step is to select values of *d* and then *p* and *q* in the *ARIMA(p,d,q)* model. When the series exhibits a trend, we may either fit and remove a deterministic trend or difference the series. Box-Jenkins seem to prefer differencing, while several other authors prefer the deterministic trend removal.

The first step, in either case, is to look at the plots of the autocorrelations and partial autocorrelations. A series with a trend will have an autocorrelation patterns similar to the following:



We notice that the large autocorrelations persist even after several lags. This indicates that either a trend should be removed or that the series should be differenced. The next step would be to difference the series.

When the series is differenced, the autocorrelation plots might appear as follows:



Differencing usually reduces the number of large autocorrelations considerably. If the differenced series still does not appear stationary, we would have to difference it again.

It is often useful to determine the magnitude of a large autocorrelation and partial autocorrelation coefficient. An autocorrelation must be at least $2/\sqrt{N}$, in absolute value to be statistically significant. The following list gives some common values of significant autocorrelations for various sample sizes. Note that even though an autocorrelation is statistically significant, it may not be large enough to worry about.

N	Large Autocorrelation
25	0.40
50	0.28
75	0.23
100	0.23
200	0.14
500	0.09
1000	0.06

By considering the patterns of the autocorrelations and the partial autocorrelations, we can guess a reasonable model for the data. The following chart shows the autocorrelation patterns that are produced by various types of ARMA models.

Model	Autocorrelations	Partial Autocorrelations
ARIMA(p,d,0)	Infinite. Tails off.	Finite. Cuts off after *p* lags.
ARIMA(0,d,q)	Finite. Cuts off after *q* lags.	Infinite. Tails off.
ARIMA(p,d,q)	Infinite. Tails off.	Infinite. Tails off.

The identification phase determines the values of *d* (differencing), *p* (autoregressive order), and *q* (moving average order). By studying the two autocorrelation plots, you estimate these values.

Differencing

The level of differencing is estimated by considering the autocorrelation plots. When the autocorrelations die out quickly, the appropriate value of *d* has been found.

Value of p

The value of *p* is determined from the partial autocorrelations of the appropriately differenced series. If the partial autocorrelations cut off after a few lags, the last lag with a large value would be the estimated value of *p*. If the partial autocorrelations do not cut off, you either have a moving average model *(p=0)* or an ARIMA model with positive *p* and *q*.

Value of q

The value of *q* is found from the autocorrelations of the appropriately differenced series. If the autocorrelations cut off after a few lags, the last lag with a large value would be the estimated value of *q*. If the autocorrelations do not cut off, you either have an autoregressive model *(q=0)* or an ARIMA model with a positive *p* and *q*.

Mixed Model

When neither the autocorrelations nor the partial autocorrelations cut off, a *mixed model* is suggested. In an *ARIMA(p,d,q)* model, the autocorrelation function will be a mixture of exponential decay and damped sine waves after the first *q-p* lags. The partial autocorrelation function have the same pattern after *p-q* lags. By studying the first few correlations of each plot, you may be able to obtain reasonable guesses for *p* and *q*.

Our experience has been that directly identifying the values of *p* and *q* in mixed models is very difficult. Instead, we use a trial and error approach in which successively more complex models are fit until the residuals show no further structure (large autocorrelations). Usually, we try fitting an *ARIMA(1,d,0),* an *ARIMA(2,d,1),* and an *ARMA(4,3).* We would select the simplest model that had a reasonably good fit. (We have found that the *ARIMA(2,d,1)* often works well and we usually begin with it.)

Identification of a seasonal series is much more difficult. Box-Jenkins describe methods for model identification, but the user must be very skilled and experienced to successfully identify the model order. We have found that trial and error must usually be used. Usually, you want to keep the number of parameters to a minimum, so the values of *p, P, q, Q, d*, and *D* that you select should be less than or equal to two.

As you can see, the identification step is subjective. One of the frequent objections about the Box-Jenkins method is that two trained forecasters will arrive at different forecasting models, even though they are using the same software. However, as we showed earlier, often models that appear to be very different on the surface are actually quite similar.

Model Estimation and Diagnostic Checking

Maximum Likelihood Estimation

Once you have guesstimated values of *p, d*, and *q*, you are ready to estimate the phis and thetas. This program follows the maximum likelihood estimation process outlined in Box-Jenkins (1976). The maximum likelihood equation is solved by nonlinear function maximization. Backcasting is used to obtain estimates of the initial residuals. The estimation process is calculation intensive and iterative, so it often takes a few seconds to obtain a solution.

Diagnostic Checking

Once a model has been fit, the final step is the diagnostic checking of the model. The checking is carried out by studying the autocorrelation plots of the residuals to see if further structure (large correlation values) can be found. If all the autocorrelations and partial autocorrelations are small, the model is considered adequate and forecasts are generated. If some of the autocorrelations are large, the values of *p* and/or *q* are adjusted and the model is re-estimated.

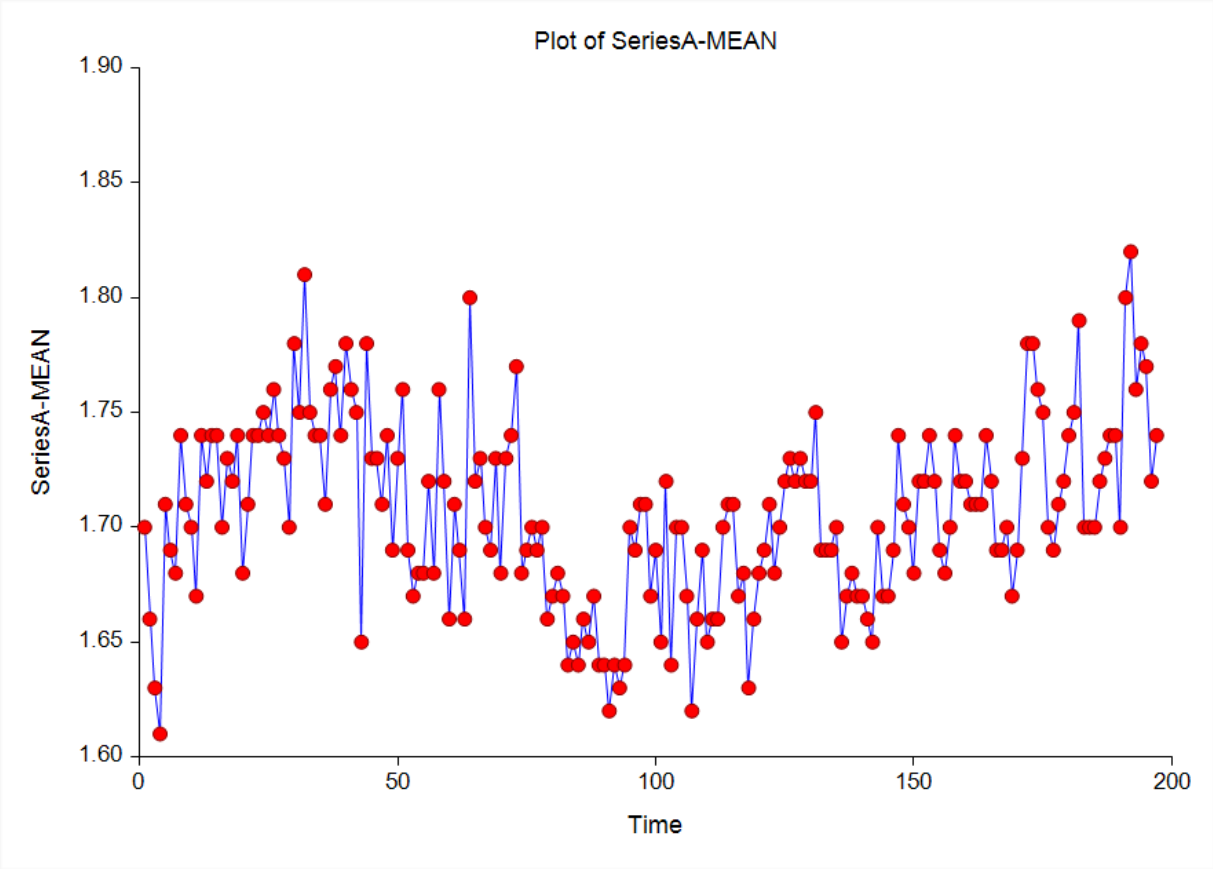
This process of checking the residuals and adjusting the values of *p* and *q* continues until the resulting residuals contain no additional structure. Once a suitable model is selected, the program may be used to generate forecasts and associated probability limits.

Example 1 – Chemical Process Concentrations

To complete this chapter, we will construct forecasts for two example problems. The first example we consider is called Series A by Box-Jenkins and is from their book. This is a set of 197 concentration values from a chemical process taken at two-hour intervals. The data are stored in the SeriesA dataset. If you want to follow along, you should open this dataset now. The following figure shows a plot of the data.

Time Series Data Plot

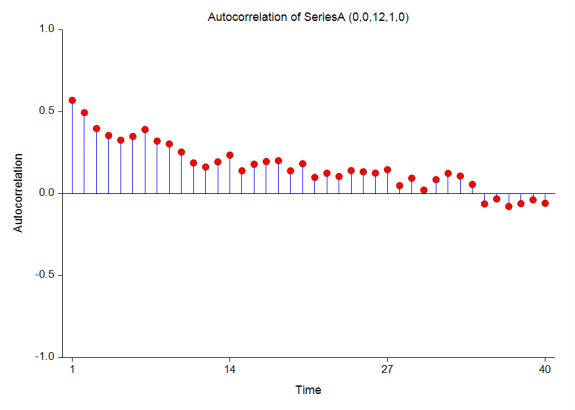
Data Plot

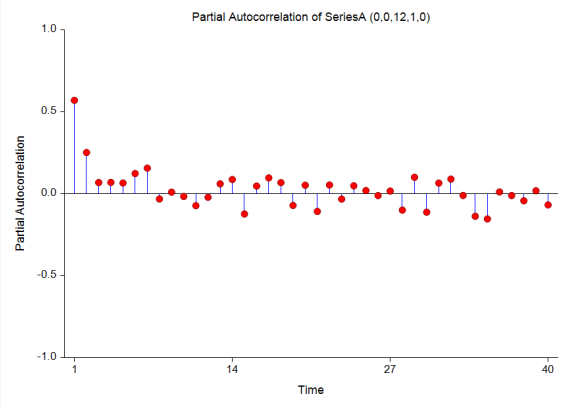


Notice that although the series moves around, it does not seem to follow a definite trend. The autocorrelation charts are shown next.

Series Autocorrelation Plots

Autocorrelation Plot Section





The autocorrelations seem to die down fairly regularly after lag 1. The partial autocorrelations seem to be small after the first one, so we decide to fit an *ARIMA(1,0,1)* to these data.

Model Estimation Reports

The following output shows the results of fitting the model.

Model Description Section

Series	SeriesA-MEAN	
Model	Regular(1,0,1)	Seasonal(No seasonal parameters)
Mean	1.706244	
Observations	197	
Missing Values	None	
Iterations	11	
Pseudo R-Squared	38.477242	
Residual Sum of Squares	0.1922096	
Mean Square Error	0.0009856902	
Root Mean Square	0.0313957	

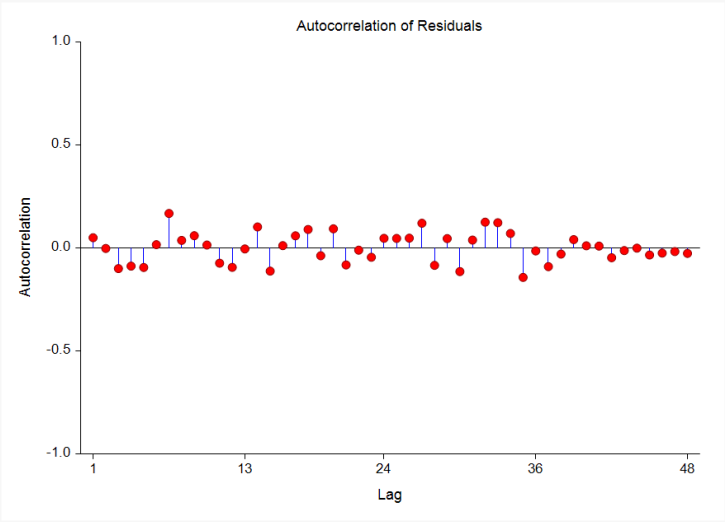
Model Estimation Section

Parameter Name	Parameter Estimate	Standard Error	T-Value	Prob Level
AR(1)	0.9208993	0.04111259	22.3994	0.000000
MA(1)	0.5958619	0.08240521	7.2309	0.000000

The final step is to make the diagnostic checks of our model. The autocorrelation plot of the residuals are shown next.

Autocorrelation of Residuals Plot

Autocorrelation Plot Section



No action here. Finally, we take a look at the Portmanteau test results.

Portmanteau Test Report

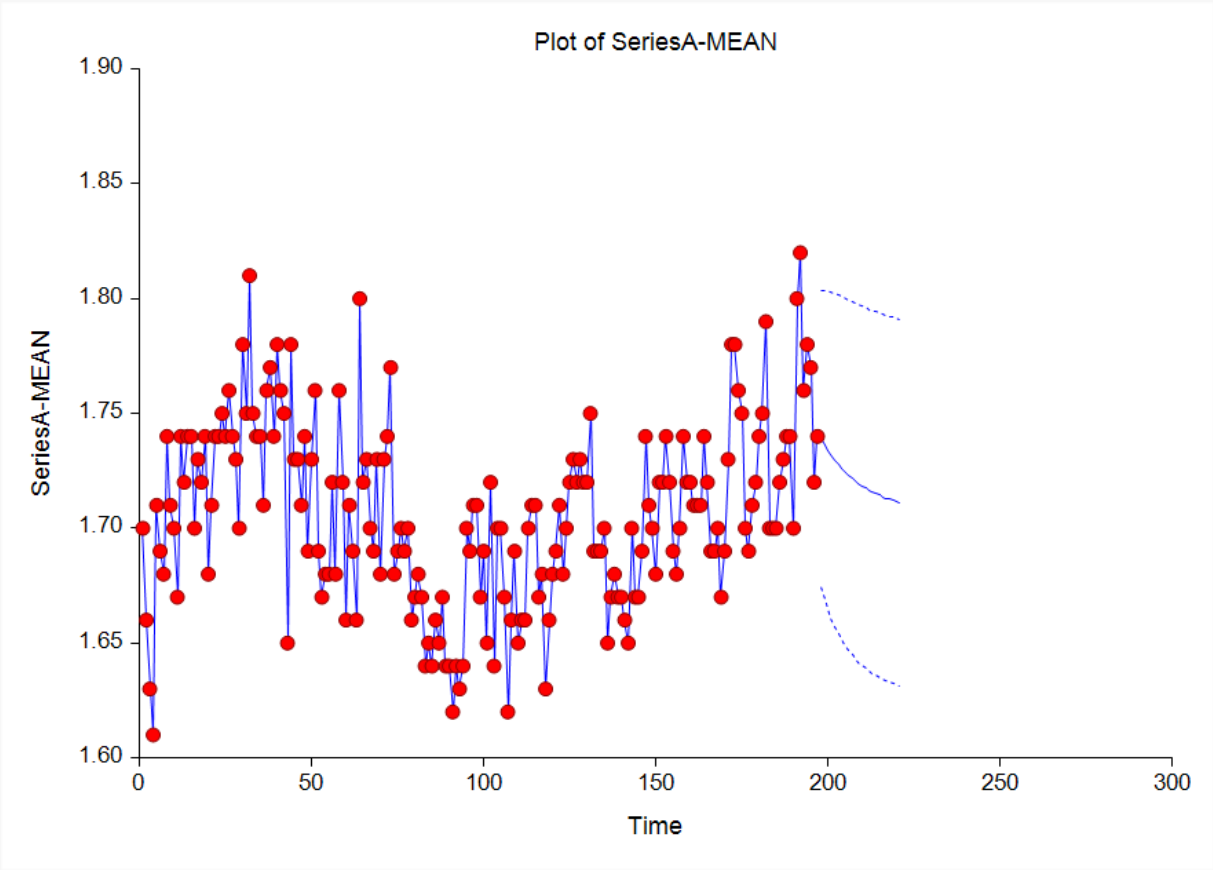
Portmanteau Test Section SeriesA-MEAN

Lag	DF	Portmanteau Test Value	Prob Level	Decision (0.05)
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
13	11	15.87	0.146054	Adequate Model
14	12	18.12	0.112215	Adequate Model
15	13	20.82	0.076500	Adequate Model
16	14	20.85	0.105477	Adequate Model
17	15	21.62	0.118122	Adequate Model
18	16	23.40	0.103375	Adequate Model
19	17	23.71	0.127450	Adequate Model
20	18	25.64	0.108222	Adequate Model
21	19	27.14	0.101334	Adequate Model
22	20	27.17	0.130594	Adequate Model
23	21	27.62	0.151180	Adequate Model
24	22	28.13	0.171548	Adequate Model
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

The diagnostic checking reveals no new patterns, so we can assume that our model is adequate. We generate the forecasts for the next few periods. These are shown next.

Time Series Plot Including Forecasts

Forecast and Data Plot

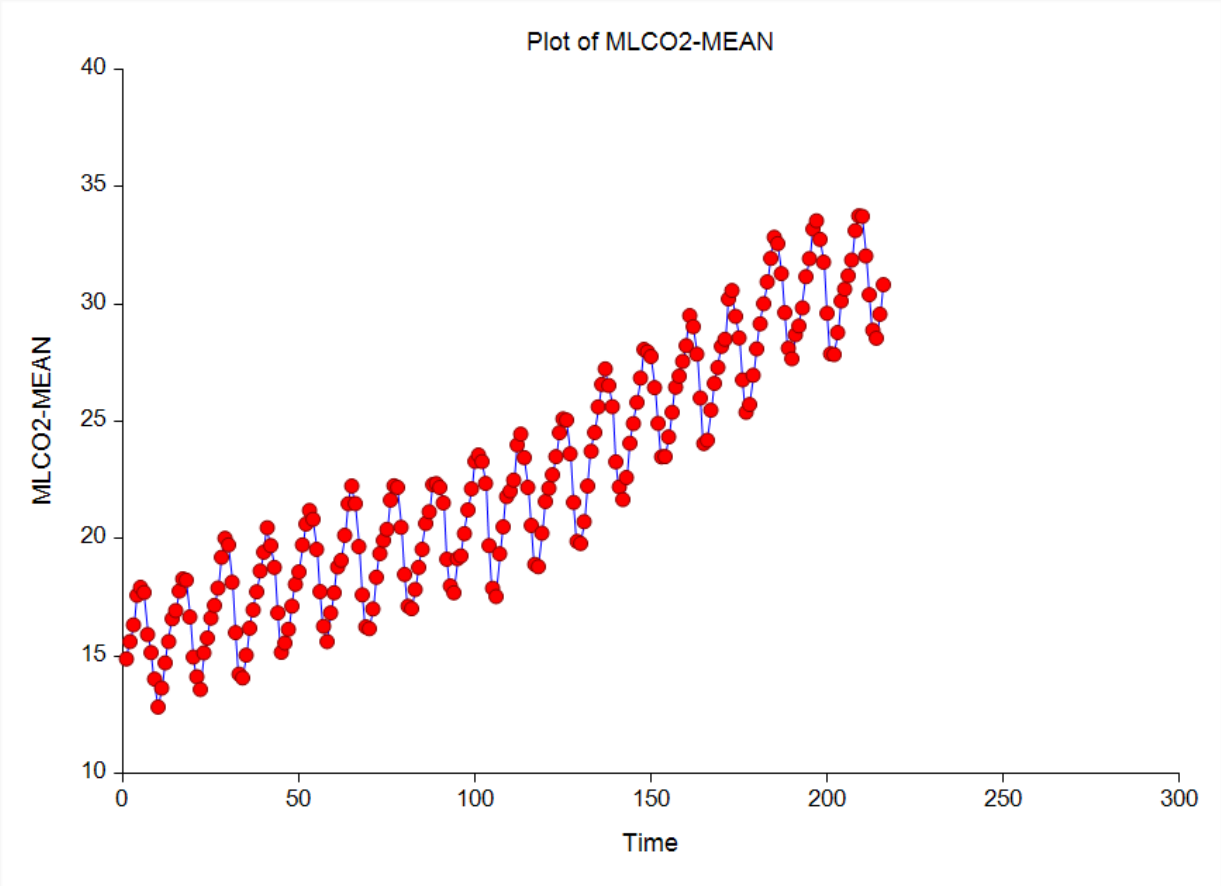


Example 2 – Carbon Dioxide Above Mauna Loa, Hawaii

This example will an approach to data with a linear trend and seasonal variation. We will consider 216 monthly carbon dioxide measurements above Mauna Loa, Hawaii. The data was obtained from Newton (1988). It is stored in the dataset named MLCO2.

Time Series Data Plot

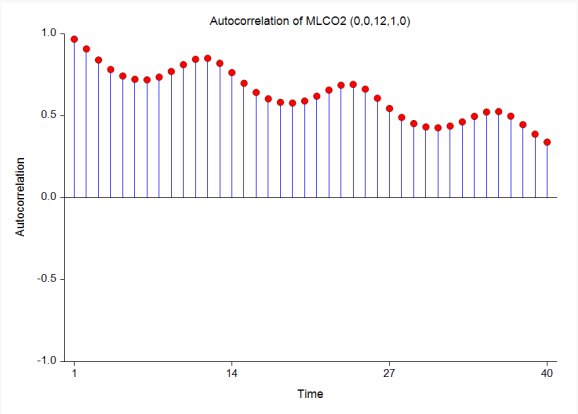
Forecast and Data Plot

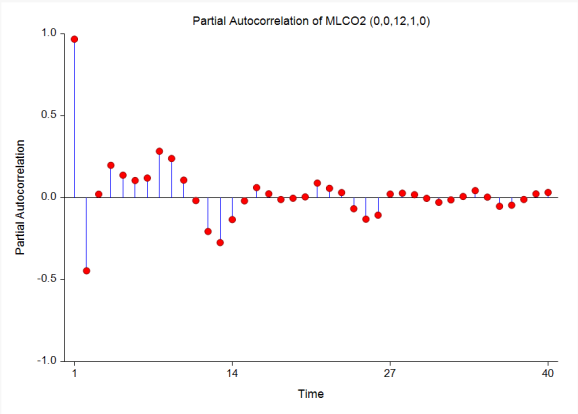


Note that the data are nonstationary on two counts: they show a trend and an annual cycle. The next step is to study the autocorrelations. The autocorrelation charts are shown next.

Series Autocorrelation Plots

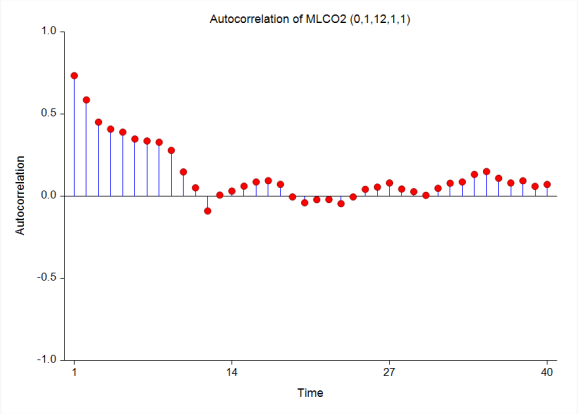
Autocorrelation Plot Section

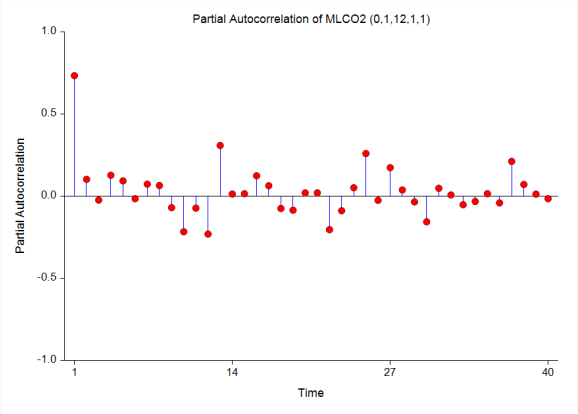




Notice that the autocorrelations do not die out and they show a cyclical pattern. This points to nonstationarity in the data. The partial autocorrelations point to a value of 2 for p. However, because of the obvious nonstationarity, we first want to look at the autocorrelation functions of the first differences. Because these are monthly data, we use seasonal differences of length twelve. We also remove the trend in the data.

Autocorrelation Plot Section





The autocorrelations die out fairly quickly. The partial autocorrelations are large around lags one and twelve. This suggests the multiplicative seasonal model: ARIMA(1,0,0) x $(1,1,0)_{12}$.

Model Estimation Reports

Following are the results of fitting this model.

Model Description Section

Series	MLCO2-TREND
Model	Regular(1,0,1) Seasonal(1,1,0) Seasons =12
Trend Equation	(14.07418)+(0.07830546)x(date)
Observations	216
Missing Values	None
Iterations	13
Pseudo R-Squared	99.500042
Residual Sum of Squares	30.3262
Mean Square Error	0.1508766
Root Mean Square	0.3884284

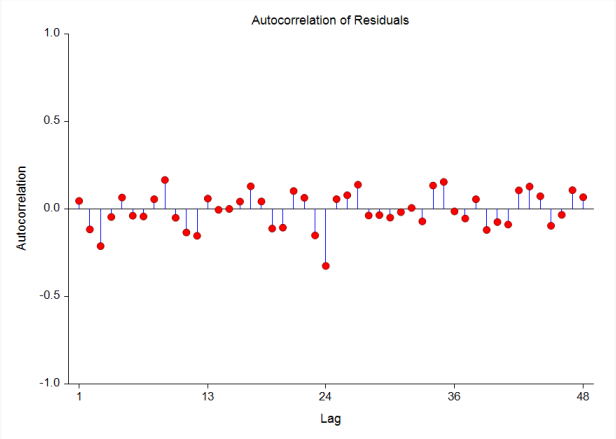
Model Estimation Section

Parameter Name	Parameter Estimate	Standard Error	T-Value	Prob Level
AR(1)	0.9836381	0.01274416	77.1834	0.000000
SAR(1)	-0.4927093	0.05991305	-8.2237	0.000000
MA(1)	0.3183001	0.06915411	4.6028	0.000004

Everything appears fine here. The final step is to make the diagnostic checks of our model. The autocorrelation plot of the residuals is shown next.

Autocorrelation of Residuals Plot

Autocorrelation Plot Section



There appear to be some persistent autocorrelations at lag 25. We take a look at the Portmanteau test results.

Portmanteau Test Report

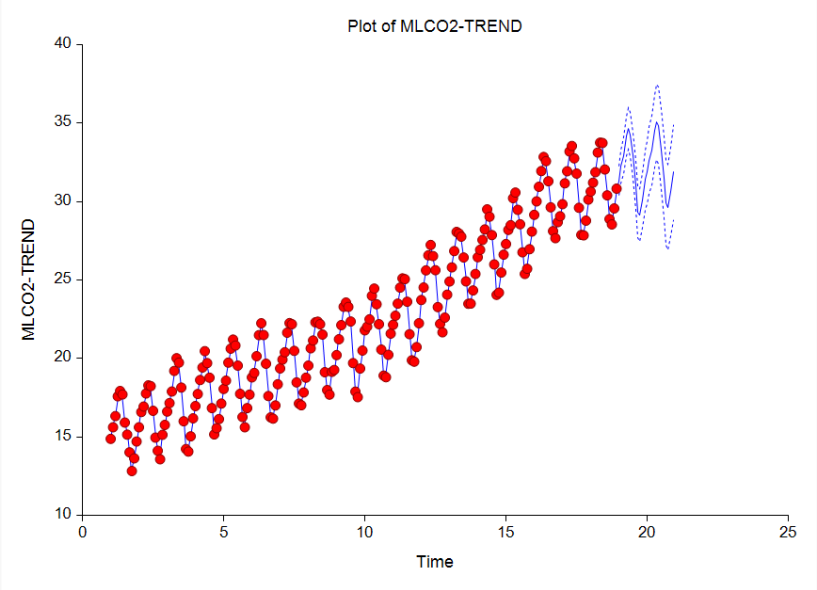
Portmanteau Test Section MLCO2-TREND

Lag	DF	Portmanteau Test Value	Prob Level	Decision (0.05)
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
13	10	33.45	0.000229	Inadequate Model
14	11	33.45	0.000445	Inadequate Model
15	12	33.45	0.000823	Inadequate Model
16	13	33.88	0.001255	Inadequate Model
17	14	37.87	0.000544	Inadequate Model
18	15	38.32	0.000810	Inadequate Model
19	16	41.28	0.000504	Inadequate Model
20	17	44.00	0.000342	Inadequate Model
21	18	46.57	0.000245	Inadequate Model
22	19	47.58	0.000295	Inadequate Model
23	20	53.11	0.000078	Inadequate Model
24	21	79.02	0.000000	Inadequate Model
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

The test points to additional information in the residual autocorrelations. We should refine our model further. We tried several other models but could not find one that worked a lot better. Finally, we generate the forecasts from this model.

Time Series Plot Including Forecasts

Forecast and Data Plot



As an exercise, you might try fitting this data with the Winters exponential smoothing algorithm.