



Published on STAT 510 (<https://onlinecourses.science.psu.edu/stat510>)

[Home](#) > Lesson 4: Seasonal Models

# Lesson 4: Seasonal Models

## Assignments:

- Read Section 3.9 of your text.
- Read through the Lesson 4 online notes that follow.
- Complete Lesson 4 Assignment.

## Overview:

This week we'll cover models for seasonal data and continue to study non-seasonal models too.

## Learning Objectives:

After successfully completing this lesson, you should be able to:

- Difference for trend and seasonality
- Identify and interpret a seasonal ARIMA model
- Distinguish seasonal ARIMA terms from simultaneously exploring an ACF and PACF
- Create and interpret diagnostic plots

## 4.1 Seasonal ARIMA models

Seasonality in a time series is a regular pattern of changes that repeats over S time periods, where S defines the number of time periods until the pattern repeats again.

For example, there is seasonality in monthly data for which high values tend always to occur in some particular months and low values tend always to occur in other particular months. In this case, S = 12 (months per year) is the span of the periodic seasonal behavior. For quarterly data, S = 4 time periods per year.

In a seasonal ARIMA model, seasonal AR and MA terms predict  $x_t$  using data values and errors at times with lags that are multiples of S (the span of the seasonality).

- With monthly data (and S = 12), a seasonal first order autoregressive model would use  $x_{t-12}$  to predict  $x_t$ . For instance, if we were selling cooling fans we might predict this August's sales using last August's sales. (This relationship of predicting using last year's data would hold for any month of the year.)
- A seasonal second order autoregressive model would use  $x_{t-12}$  and  $x_{t-24}$  to predict  $x_t$ .

Here we would predict this August's values from the past two Augusts.

- A seasonal first order MA(1) model (with  $S = 12$ ) would use  $w_{t-12}$  as a predictor. A seasonal second order MA(2) model would use  $w_{t-12}$  and  $w_{t-24}$ .

## Differencing

Almost by definition, it may be necessary to examine differenced data when we have seasonality. Seasonality usually causes the series to be nonstationary because the average values at some particular times within the seasonal span (months, for example) may be different than the average values at other times. For instance, our sales of cooling fans will always be higher in the summer months.

**Seasonal differencing** is defined as a difference between a value and a value with lag that is a multiple of  $S$ .

- With  $S = 12$ , which may occur with monthly data, a seasonal difference is  $(1-B^{12})x_t = x_t - x_{t-12}$ .

The differences (from the previous year) may be about the same for each month of the year giving us a stationary series.

- With  $S = 4$ , which may occur with quarterly data, a seasonal difference is  $(1-B^4)x_t = x_t - x_{t-4}$ .

Seasonal differencing removes seasonal trend and can also get rid of a seasonal random walk type of nonstationarity.

**Non-seasonal differencing:** If trend is present in the data, we may also need non-seasonal differencing. Often (not always) a first difference (non-seasonal) will “detrend” the data. That is, we use  $(1-B)x_t = x_t - x_{t-1}$  in the presence of trend.

**Differencing for Trend and Seasonality:** When both trend and seasonality are present, we may need to apply both a non-seasonal first difference and a seasonal difference.

That is, we may need to examine the ACF and PACF of  $(1-B^{12})(1-B)x_t = (x_t - x_{t-1}) - (x_{t-12} - x_{t-13})$ .

Removing trend doesn't mean that we have removed the dependency. We may have removed the mean,  $\mu_t$ , part of which may include a periodic component. In some ways we are breaking the dependency down into recent things that have happened and long-range things that have happened.

## Non-seasonal Behavior Will Still Matter

With seasonal data, it is likely that short run non-seasonal components will still contribute to the model. In the monthly sales of cooling fans mentioned above, for instance, sales in the previous month or two, along with the sales from the same month a year ago, may help predict this month's sales.

We'll have to look at the ACF and PACF behavior over the first few lags (less than  $S$ ) to

assess what non-seasonal terms might work in the model.

## Seasonal ARIMA Model

The seasonal ARIMA model incorporates both non-seasonal and seasonal factors in a multiplicative model. One shorthand notation for the model is

$$\text{ARIMA}(p, d, q) \times (P, D, Q)_S,$$

with  $p$  = non-seasonal AR order,  $d$  = non-seasonal differencing,  $q$  = non-seasonal MA order,  $P$  = seasonal AR order,  $D$  = seasonal differencing,  $Q$  = seasonal MA order, and  $S$  = time span of repeating seasonal pattern.

Without differencing operations, the model could be written more formally as

$$(1) \quad \Phi(B^S)\varphi(B)(x_t - \mu) = \Theta(B^S)\theta(B)w_t$$

The non-seasonal components are:

$$\text{AR: } \varphi(B) = 1 - \varphi_1B - \dots - \varphi_pB^p$$

$$\text{MA: } \theta(B) = 1 + \theta_1B + \dots + \theta_qB^q$$

The seasonal components are:

$$\text{Seasonal AR: } \Phi(B^S) = 1 - \Phi_1B^S - \dots - \Phi_PB^{PS}$$

$$\text{Seasonal MA: } \Theta(B^S) = 1 + \Theta_1B^S + \dots + \Theta_QB^{QS}$$

Note that on the left side of equation (1) the seasonal and non-seasonal AR components multiply each other, and on the right side of equation (1) the seasonal and non-seasonal MA components multiply each other.

### Example 1: ARIMA (0, 0, 1) $\times$ (0, 0, 1)<sub>12</sub>

The model includes a non-seasonal MA(1) term, a seasonal MA(1) term, no differencing, no AR terms and the seasonal period is  $S = 12$ .

The non-seasonal MA(1) polynomial is  $\theta(B) = 1 + \theta_1B$ .

The seasonal MA(1) polynomial is  $\Theta(B^{12}) = 1 + \Theta_1B^{12}$ .

The model is  $(x_t - \mu) = \Theta_1(B^{12})\theta_1(B)w_t = (1 + \Theta_1B^{12})(1 + \theta_1B)w_t$ .

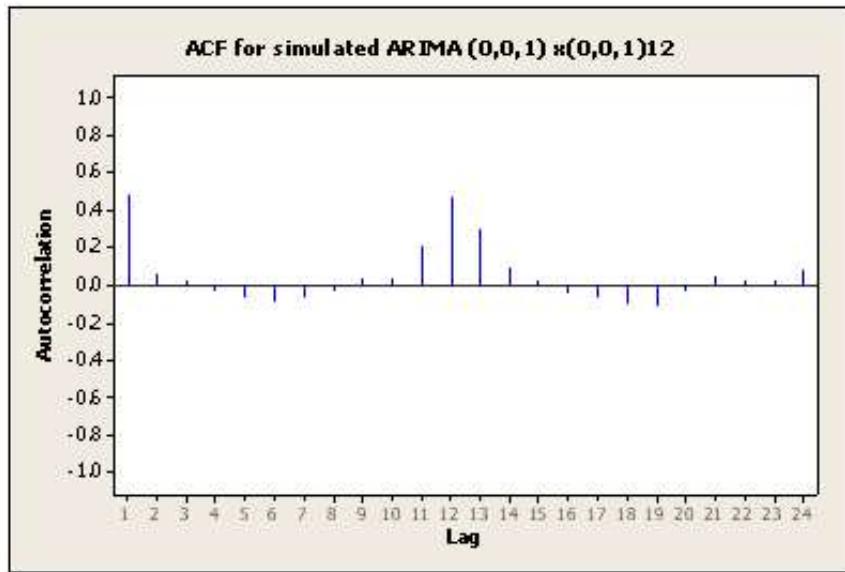
When we multiply the two polynomials on the right side, we get

$$\begin{aligned} (x_t - \mu) &= (1 + \theta_1B + \Theta_1B^{12} + \theta_1\Theta_1B^{13})w_t \\ &= w_t + \theta_1w_{t-1} + \Theta_1w_{t-12} + \theta_1\Theta_1w_{t-13}. \end{aligned}$$

Thus the model has MA terms at lags 1, 12, and 13. This leads many to think that the identifying ACF for the model will have non-zero autocorrelations only at lags 1, 12, and 13. There's a slight surprise here. There will also be a non-zero autocorrelation at lag 11. We supply a proof in Appendix 1 for this document.

### Example 1 Continued:

We used Minitab to simulate  $n = 1000$  values from an ARIMA  $(0, 0, 1) \times (0, 0, 1)_{12}$ . The non-seasonal MA (1) coefficient was  $\theta_1=.7$ . The seasonal MA(1) coefficient was  $\Theta_1=.6$ . The sample ACF for the simulated series was as follows:



Note the spikes at lags 1, 11, 12, and 13. This is characteristic of the ACF for the ARIMA  $(0, 0, 1) \times (0, 0, 1)_{12}$ .

### Example 2 ARIMA $(1, 0, 0) \times (1, 0, 0)_{12}$

The model includes a non-seasonal AR(1) term, a seasonal AR(1) term, no differencing, no MA terms and the seasonal period is  $S = 12$ .

The non-seasonal AR(1) polynomial is  $\varphi(B) = 1 - \varphi_1 B$ .

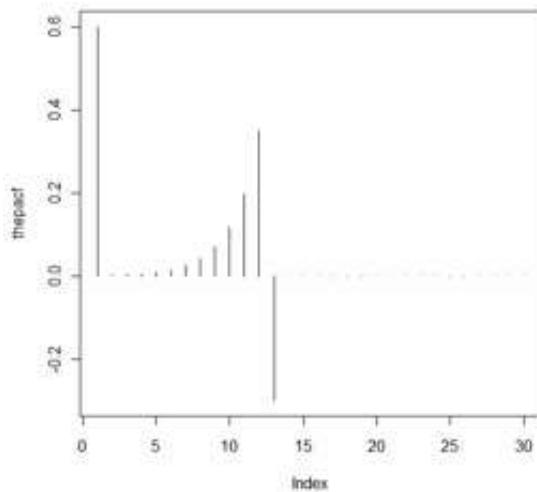
The seasonal AR(1) polynomial is  $\Phi(B^{12}) = 1 - \Phi_1 B^{12}$ .

The model is  $(1 - \Phi_1 B^{12})(1 - \varphi_1 B)(x_t - \mu) = w_t$ .

If we let  $z_t = x_t - \mu$  (for simplicity), multiply the two AR components and push all but  $z_t$  to the right side we get  $z_t = \varphi_1 z_{t-1} + \Phi_1 z_{t-12} + (-\Phi_1 \varphi_1) z_{t-13} + w_t$ .

This is an AR model with predictors at lags 1, 12, and 13.

R can be used to determine and plot the PACF for this model, with  $\varphi_1=.6$  and  $\Phi_1=.5$ . That PACF (partial autocorrelation function) is:



It's not quite what you might expect for an AR, but it almost is. There are distinct spikes at lags 1, 12, and 13 with a bit of action coming before lag 12. Then, it cuts off after lag 13.

Note: R commands were

```
thepacf=ARMAacf (ar = c(.6,0,0,0,0,0,0,0,0,.5,-.30),lag.max=30,pacf=T)
plot (thepacf,type="h")
```

## Identifying a Seasonal Model

**Step 1:** Do a time series plot of the data. Examine it for features such as trend and seasonality. You'll know that you've gathered seasonal data (months, quarters, etc.) so look at the pattern across those time units (months, etc.) to see if there is indeed a seasonal pattern.

**Step 2:** Do any necessary differencing. The general guidelines are:

- If there is seasonality and no trend take a difference of lag S. For instance, take a 12<sup>th</sup> difference for monthly data with seasonality.
- If there is linear trend and no obvious seasonality, take a first difference. If there is a curved trend, consider a transformation of the data before differencing.
- If there is both trend and seasonality, apply both a non-seasonal and seasonal difference to the data, as two successive operations (in either order). For instance, if the series is called *x*, a successive first and 12<sup>th</sup> difference in R would be:

```
diff1 = diff(x, 1)
diff1and12=diff(diff1, 12)
```

- If there is neither obvious trend nor seasonality, don't take any differences.

**Step 3:** Examine the ACF and PACF of the differenced data (if differencing is necessary).

We're using this information to determine possible models. This can be tricky going involving some (educated) guessing. Some basic guidance:

**Non-seasonal terms:** Examine the early lags (1, 2, 3, ...) to judge non-seasonal terms. Spikes in the ACF (at low lags) indicate non-seasonal MA terms. Spikes in the PACF (at low lags) indicated possible non-seasonal AR terms.

**Seasonal terms:** Examine the patterns across lags that are multiples of S. For example, for monthly data, look at lags 12, 24, 36, and so on (probably won't need to look at much more than the first two or three seasonal multiples). Judge the ACF and PACF at the seasonal lags in the same way you do for the earlier lags.

**Step 4:** Estimate the model(s) that might be reasonable on the basis of Step 3. Don't forget to include any differencing that you did before looking at the ACF and PACF. In the software, specify the original series as the data and then indicate the desired differencing when specifying parameters in the `arima` command that you're using.

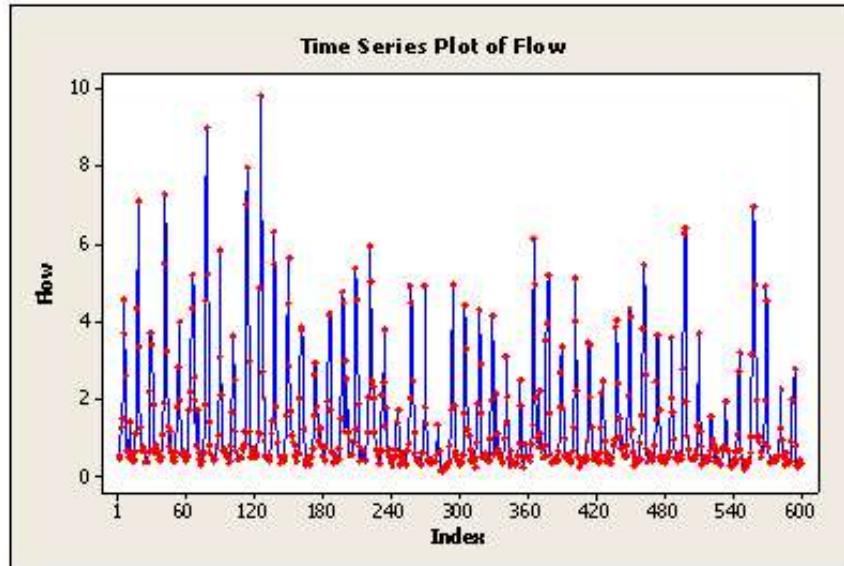
**Step 5:** Examine the residuals (with ACF, Box-Pierce, and any other means) to see if the model seems good. Compare AIC or BIC values if you tried several models.

If things don't look good here, it's back to Step 3 (or maybe even Step 2).

### Example 3

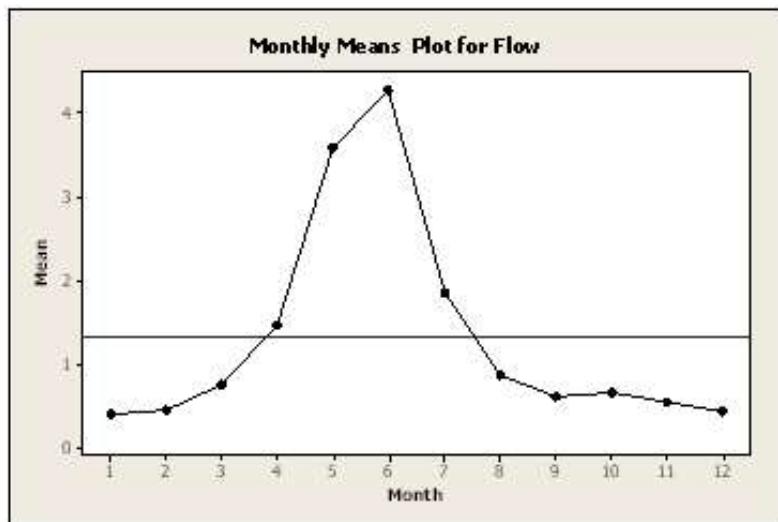
The data series are a monthly series of a measure of the flow of the Colorado River, at a particular site, for  $n = 600$  consecutive months.

**Step 1:** A time series plot is



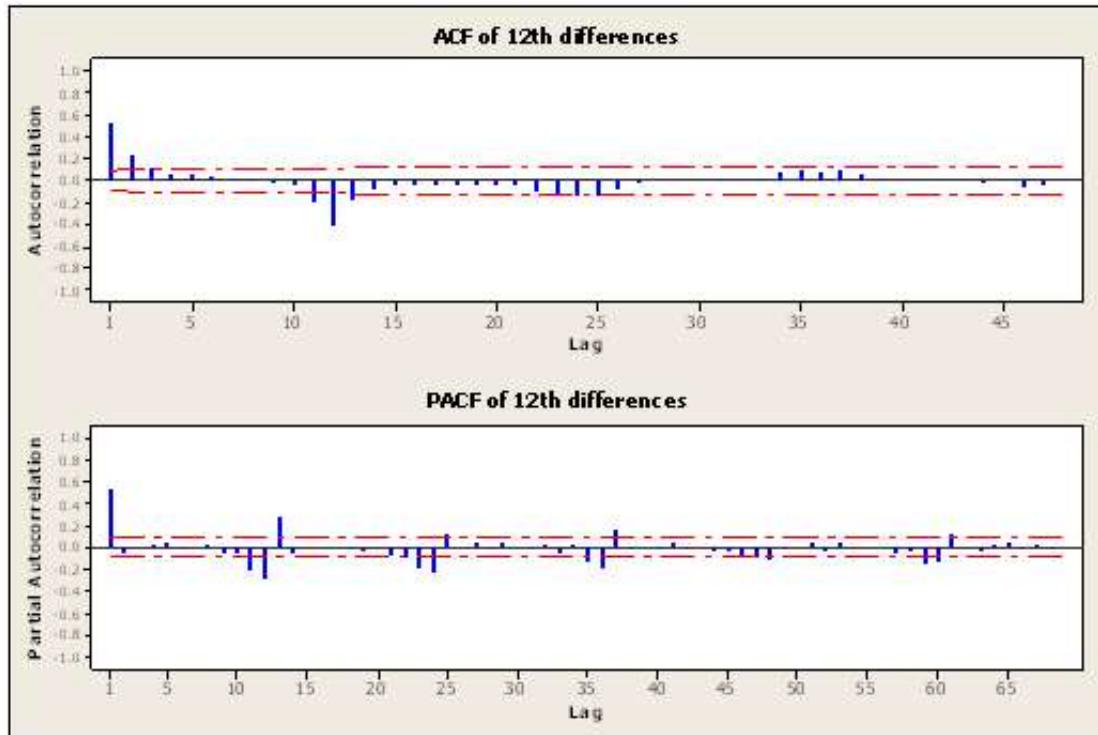
With so many data points, it's difficult to judge whether there is seasonality. If it was your job to work on data like this, you probably would know that river flow is seasonal – perhaps likely to be higher in the late spring and early summer, due to snow runoff.

Without this knowledge, we might determine means by month of the year. Below is a plot of means for the 12 months of the year. It's clear that there are monthly differences (seasonality).



Looking back at the time series plot, it's hard to judge whether there's any long run trend. If there is, it's slight.

*Steps 2 and 3:* We might try the idea that there is seasonality, but no trend. To do this, we can create a variable that gives the 12<sup>th</sup> differences (seasonal differences), calculated as  $x_t - x_{t-12}$ . Then, we look at the ACF and the PACF for the 12<sup>th</sup> difference series (not the original data). Here they are:



*Non-seasonal behavior:* The PACF shows a clear spike at lag 1 and not much else until about lag 11. This is accompanied by a tapering pattern in the early lags of the ACF. A non-seasonal AR(1) may be a useful part of the model.

*Seasonal behavior:* We look at what's going on around lags 12, 24, and so on. In the ACF, there's a cluster of (negative) spikes around lag 12 and then not much else. The PACF tapers in multiples of S; that is the PACF has significant lags at 12, 24, 36 and so on. This is

similar to what we saw for a seasonal MA(1) component in Example 1 of this lesson.

Remembering that we're looking at 12<sup>th</sup> differences, the model we might try for the original series is ARIMA (1,0,0)×(0,1,1)<sub>12</sub>.

**Step 4:** Minitab results for the ARIMA (1,0,0)× (0,1,1)<sub>12</sub>:

#### Final Estimates of Parameters

Type	Coef	SE Coef	T	P
AR 1	0.5162	0.0354	14.58	0.000
SMA 12	0.9140	0.0169	53.95	0.000
Constant	-0.006502	0.002884	-2.25	0.025

Differencing: 0 regular, 1 seasonal of order 12

Number of observations: Original series 600, after differencing 588

Residuals: SS = 272.886 (backforecasts excluded)

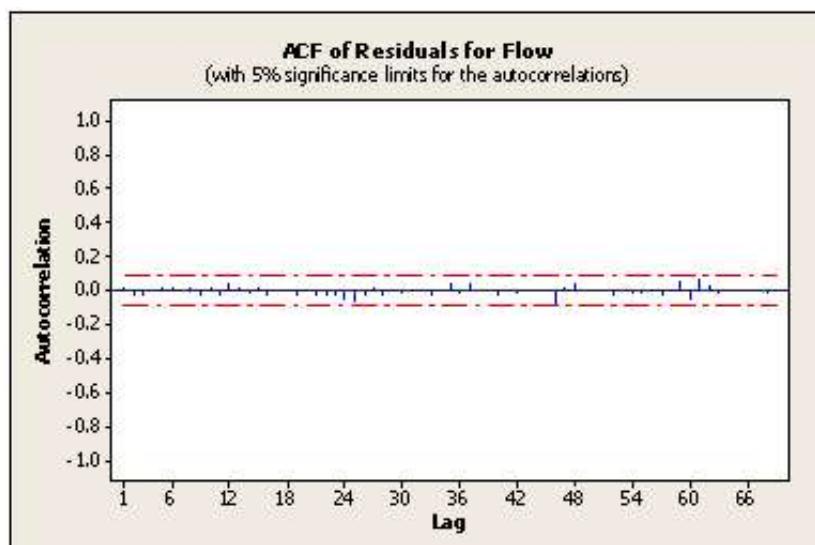
MS = 0.466 DF = 585

#### Modified Box-Pierce (Ljung-Box) Chi-Square statistic

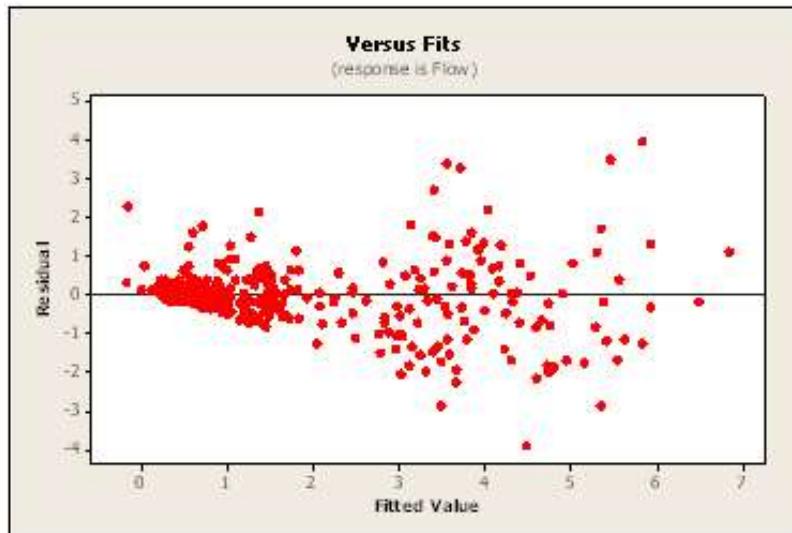
Lag	12	24	36	48
Chi-Square	5.6	10.0	16.3	24.0
DF	9	21	33	45
P-Value	0.781	0.979	0.994	0.996

Things look good. The Box-Pierce statistics are all non-significant and the estimated coefficients (previous page) are statistically significant.

**Step 5 (diagnostics):** We've already looked at the Box-Pierce tests. The ACF of the residuals looks good too:



What doesn't look perfect is a plot of residuals versus fits. There's non-constant variance.



We've got three choices for what to do about the non-constant variance: (1) ignore it, (2) go back to step 1 and try a variance stabilizing transformation like log or square root, or (3) use an ARCH model that includes a component for changing variances. We'll get to ARCH models later in the course.

Lesson 4.2 for this week will give R guidance and an additional example or two.

### Appendix (Optional reading):

Only those interested in theory things need to look at the following.

In Example 1, we promised a proof that  $\rho_{11} \neq 0$  for ARIMA  $(0, 0, 1) \times (0, 0, 1)_{12}$ .

A correlation is defined as Covariance/ product of standard deviations.

The covariance between  $x_t$  and  $x_{t-11}$  =  $E(x_t - \mu)(x_{t-11} - \mu)$ .

For the model in Example 1,

$$x_t - \mu = w_t + \theta_1 w_{t-1} + \Theta_1 w_{t-12} + \theta_1 \Theta_1 w_{t-13}$$

$$x_{t-11} - \mu = w_{t-11} + \theta_1 w_{t-12} + \Theta_1 w_{t-23} + \theta_1 \Theta_1 w_{t-24}$$

The covariance between  $x_t$  and  $x_{t-11}$

$$(2) E(w_t + \theta_1 w_{t-1} + \Theta_1 w_{t-12} + \theta_1 \Theta_1 w_{t-13})(w_{t-11} + \theta_1 w_{t-12} + \Theta_1 w_{t-23} + \theta_1 \Theta_1 w_{t-24})$$

The  $w$ 's are independent errors. The expected value of any product involving  $w$ 's with different subscripts will be 0. A covariance between  $w$ 's with the same subscripts will be the variance of  $w$ .

If you inspect all possible products in expression 2, there will be one product with matching subscripts. They have lag  $t - 12$ . Thus this expected value (covariance) will be different from 0.

This shows that the lag 11 autocorrelation will be different from 0. If you look at the more general problem, you can find that only lags 1, 11, 12, and 13 have non-zero autocorrelations for the ARIMA(0, 0, 1)  $\times$  (0, 0, 1)<sub>12</sub>.

A seasonal ARIMA model incorporates both non-seasonal and seasonal factors in a multiplicative fashion.

## 4.2 Identifying Seasonal Models and R Code

In Lesson 4.1, Example 3 described the analysis of monthly flow data for a Colorado River location. An ARIMA(1,0,0)  $\times$  (0,1,1)<sub>12</sub> was identified and estimated. In the first part of this lesson, you'll see the R code and output for that analysis. (Lesson 4.1 gave Minitab output.)

### Example 3 Revisited: R code for the Colorado River Analysis

The data are in the file coloradoflow.dat. We used scripts in the astsa library (Stoffer's routines as described in Weeks 1 & 3), so the first two lines of code are:

```
library(astsa)
flow <- ts(scan("coloradoflow.dat"))
```

The first step in the analysis is to plot the time series. The command is:

```
plot(flow, type="b")
```

The plot showed clear monthly effects and no obvious trend, so we examined the ACF and PACF of the 12<sup>th</sup> differences (seasonal differencing). Commands are:

```
diff12 = diff(flow, 12)
acf2(diff12, 48)
```

The `acf2` command asks for information about 48 lags. On the basis of the ACF and PACF of the 12<sup>th</sup> differences, we identified an ARIMA(1,0,0)  $\times$  (0,1,1)<sub>12</sub> model as a possibility (See Lesson 4.1). The command for fitting this model is

```
sarima(flow, 1,0,0,0,1,1,12)
```

The parameters of the command just given are the data series, the non-seasonal specification of AR, differencing, and MA, and then the seasonal specification of seasonal AR, seasonal differencing, seasonal MA, and period or span for the seasonality.

Output from the `sarima` command is:

Coefficients:

arl	smal	constant
0.5149	-0.8828	-0.0011
s.e.	0.0356	0.0239

0.0007

$\sigma^2$  estimated as 0.4778: log likelihood = -607.72, aic = 1223.44

\$AIC

[1] 0.2715198

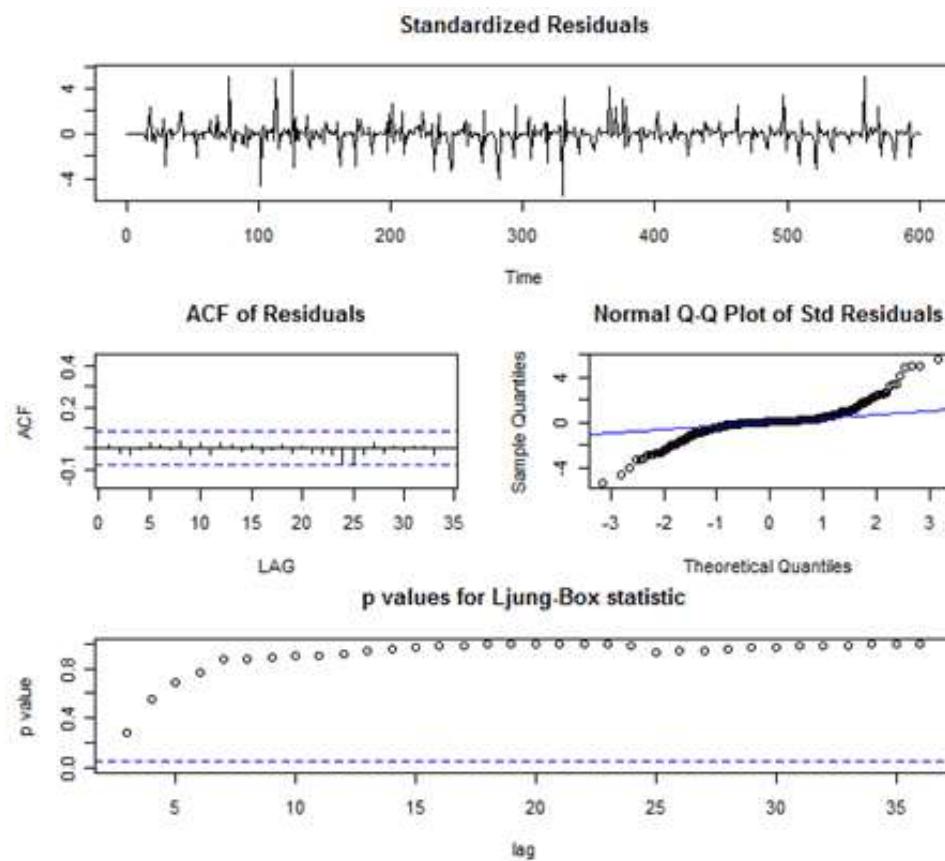
\$AICc

[1] 0.2749652

\$BIC

[1] -0.7064956

The output included these residual plots. The only difficulty we see is in the normal probability plot. The extreme standardized sample residuals (on both ends) are larger than they would be for normally distributed data. This could be related to the non-constant variance noted in Lesson 4.1.



We didn't generate any forecasts in Lesson 4.1, but if we were to use R to generate forecasts for the next 24 months in R, one command that could be used is

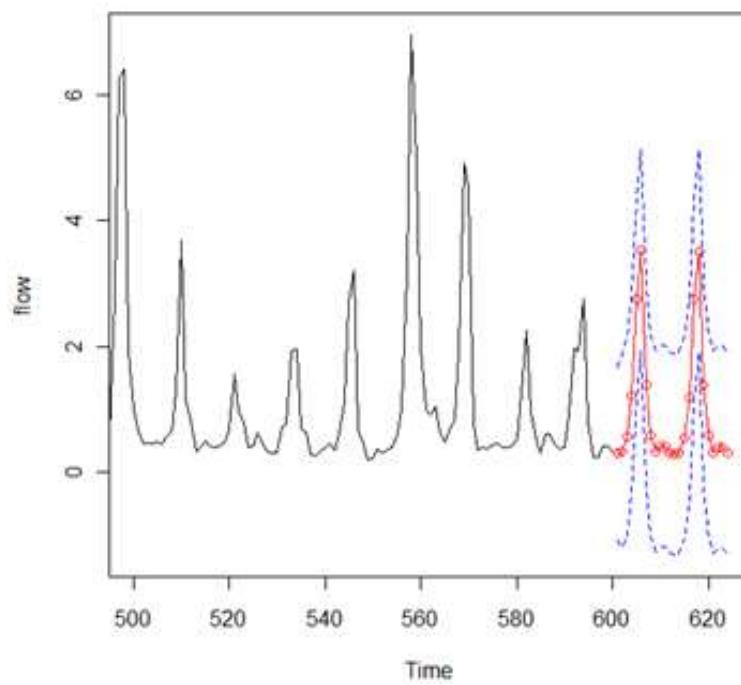
```
sarima.for(flow, 24, 1, 0, 0, 0, 1, 1, 12)
```

Note that the order of parameters in the command is the name of the data series, the number of times for which we want forecasts, followed by the parameters of the ARIMA model.

Partial output for the forecast command follows (We skipped giving the standard errors.)

```
$pred
Time Series:
Start = 601
End = 624
Frequency = 1

[1] 0.3013677 0.3243119 0.5676440 1.2113713 2.7631409 3.5326281
[7] 1.3931886 0.5971816 0.3294376 0.4161899 0.4180578 0.3186012
[13] 0.2838731 0.3088276 0.5531948 1.1974552 2.7494992 3.5191278
[19] 1.3797610 0.5837915 0.3160668 0.4028290 0.4047020 0.3052481
```



Note that the lower limits of some prediction intervals are negative - impossible for the flow of a river. In practice, we might truncate these lower limits to 0 when presenting them.

If you were to use R's native commands to do the fit and forecasts, the commands might be:

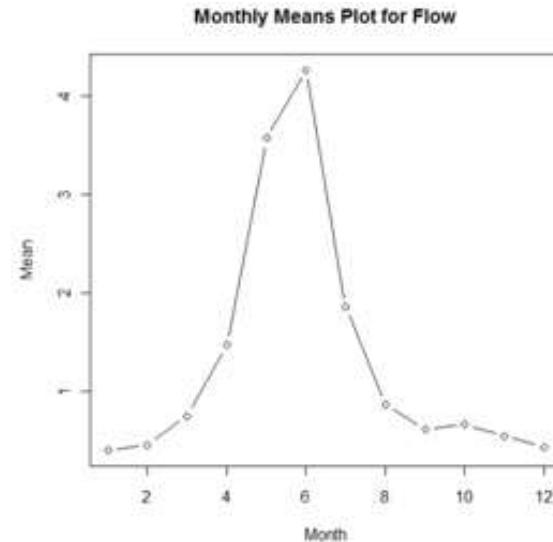
```
themodel = arima(flow, order = c(1,0,0), seasonal = list(order = c(0,1,1),
period = 12))
themodel
predict(themodel, n.ahead=24)
```

The first command does the arima and stores results in an “object” called “themodel.” The second command, which is simply themodel, lists the results and the final command generates forecasts for 24 times ahead.

In Lesson 4.1, we presented a graph of the monthly means to make the case that the data are seasonal. The R commands are:

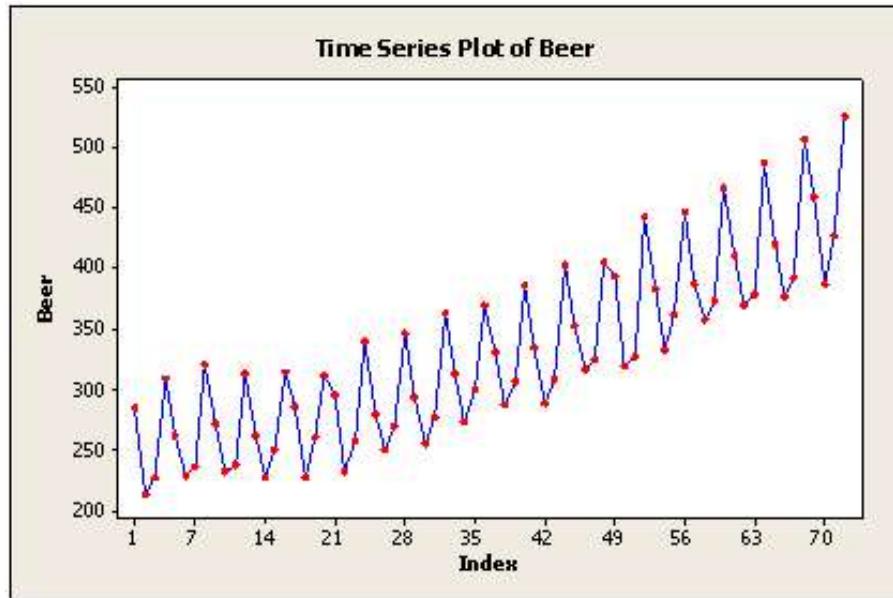
```
flowm = matrix(flow, ncol=12, byrow=TRUE)
col.means=apply(flowm, 2, mean)
```

```
plot(col.means, type="b", main="Monthly Means Plot for Flow", xlab="Month",
ylab="Mean")
```



### Example 4 (for the week): Beer Production in Australia

In Lesson 4.1, this plot for quarterly beer production in Australia was given.



There is an upward trend and seasonality. The regular pattern of ups and downs is an indicator of seasonality. If you count points, you'll see that the 4<sup>th</sup> quarter of a year is the high point, the 2<sup>nd</sup> and 3<sup>rd</sup> quarters are low points, and the 1<sup>st</sup> quarter value falls in between.

With trend and quarterly seasonality we should take both a first and a fourth difference (in either order). The plot above was produced in Minitab, but we'll use R for the rest of this. The commands for creating the differences and graphing the ACF and PACF of these differences are:

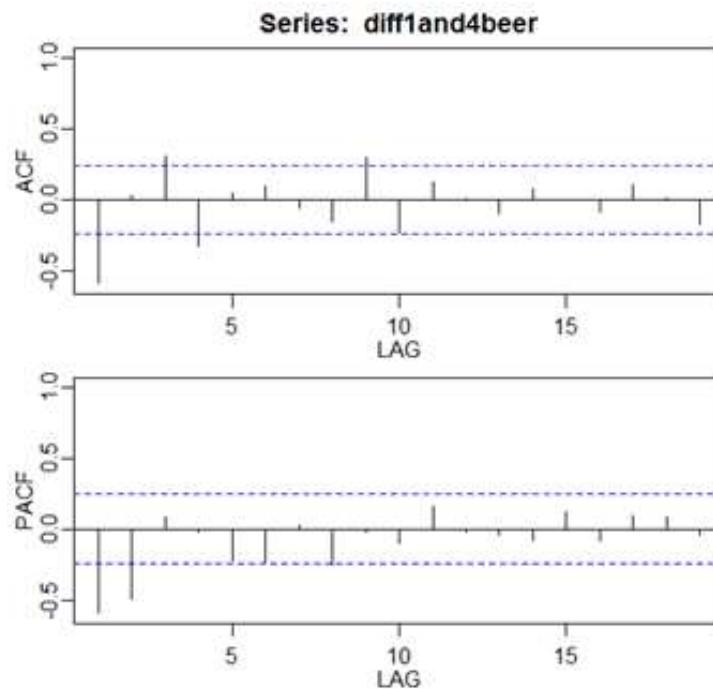
```
diff1beer = diff(beer,1)
diff1and4beer = diff(diff1beer,4)
```

```
acf2(diff1and4beer, 24)
```

Note: When we read the data, we called the series “beer” and also loaded the astsa library.

The ACF and PACF of the differenced data follow. The interpretation:

- **Non-seasonal:** Looking at just the first 2 or 3 lags, it seems possible that a MA(1) might work based on the single spike in the ACF and the PACF tapering to 0. With S=4 the nonseasonal aspect is sometimes difficult to interpret in such a narrow window. Because the PACF only has one additional significant lag that might lead to an ARMA(1,1) interpretation.
- **Seasonal:** Look at lags that are multiples of 4 (we have quarterly data). Not much is going on there, although there is a (barely) significant spike in the ACF at lag 4 and a somewhat confusing spike at lag 9 (in ACF). Nothing significant is happening at the higher lags. Maybe a seasonal MA(1) or MA(2) might work.



Fitting the model – We tried a few models. An initial guess was ARIMA(0,1,1) $\times$ (0,1,1)<sub>4</sub>, which wasn't a bad guess. We also tried ARIMA(1,1,1) $\times$ (0,1,1)<sub>4</sub>, ARIMA(0,1,1) $\times$ (0,1,2)<sub>4</sub>, and ARIMA(1,1,1) $\times$ (0,1,2)<sub>4</sub>.

The commands for the 4 models are:

```
sarima (beer, 0,1,1,0,1,1,4)
sarima (beer, 1,1,1,0,1,1,4)
sarima (beer, 0,1,1,0,1,2,4)
sarima (beer, 1,1,1,0,1,2,4)
```

A summary of the results is:

---

Model	MSE	Sig. of coefficients	AIC <sub>C</sub>	ACF of residuals
-------	-----	----------------------	------------------	------------------

ARIMA(0,1,1)×(0,1,1) <sub>4</sub>	116.1	All significant	5.810	Sig. spikes at lags 1 and 9
ARIMA(1,1,1)×(0,1,1) <sub>4</sub>	107.3	All significant	5.759	OK, except a barely sig. spike at lag 9
ARIMA(0,1,1)×(0,1,2) <sub>4</sub>	116.1	The Seasonal MA for lag 2 was not sig.	5.838	Sig. spikes at lags 1 and 9
ARIMA(1,1,1)×(0,1,2) <sub>4</sub>	107.3	The Seasonal MA for lag 2 was not sig.	5.787	OK, except a barely sig. spike at lag 9

The second model is the winner so far. The difficulty at lag 9 probably will have to be ignored. It's probably just a sampling oddity.

*Datasets:* You might use these to work through this lesson.

The beer production data is in the Week 4 folder as beerprod.dat

The Colorado River flow is in the Week 4 folder as coloradolflow.dat

**Source URL:** <https://onlinecourses.science.psu.edu/stat510/node/50>