http://ucanalytics.com/blogs/step-by-step-graphic-guide-to-forecasting-through-arima-modeling-in-r-manufacturing-case-study-example/



|  |
| --- |
| * [**Home**](http://ucanalytics.com/blogs)
* [**Blog-Navigation**](http://ucanalytics.com/blogs/navigation/)
* [**Art**](http://ucanalytics.com/blogs/art-gallery/)
* [**About**](http://ucanalytics.com/blogs/about-me/)
* [**Contact**](http://ucanalytics.com/blogs/contact/)
 |

[**Step-by-Step Graphic Guide to Forecasting through ARIMA Modeling using R – Manufacturing Case Study Example (Part 4)**](http://ucanalytics.com/blogs/step-by-step-graphic-guide-to-forecasting-through-arima-modeling-in-r-manufacturing-case-study-example/)

[28th June, 2015](http://ucanalytics.com/blogs/step-by-step-graphic-guide-to-forecasting-through-arima-modeling-in-r-manufacturing-case-study-example/) · [Roopam Upadhyay](http://ucanalytics.com/blogs/author/roopam/%22%20%5Co%20%22View%20all%20posts%20by%20Roopam%20Upadhyay) [91 Comments](http://ucanalytics.com/blogs/step-by-step-graphic-guide-to-forecasting-through-arima-modeling-in-r-manufacturing-case-study-example/#comments)

This article is a continuation of our manufacturing case study example to forecast tractor sales through time series and ARIMA models. You can find the previous parts at the following links:

[**Part 1**](http://ucanalytics.com/blogs/forecasting-time-series-analysis-manufacturing-case-study-part-1/) : Introduction to time series modeling & forecasting

[**Part 2**](http://ucanalytics.com/blogs/time-series-decomposition-manufacturing-case-study-part-2/): Time series decomposition to decipher patterns and trends before forecasting

[**Part 3**](http://ucanalytics.com/blogs/arima-models-manufacturing-case-study-example-part-3/): Introduction to ARIMA models for forecasting

In this part, we will use plots and graphs to forecast tractor sales for PowerHorse tractors through ARIMA. We will use ARIMA modeling concepts learned in the previous article for our case study example. But before we start our analysis, let’s have a quick discussion on forecasting:

**Trouble with Nostradamus**



Human Obsession with Future & ARIMA – by Roopam

Humans are obsessed about their future – so much so that they worry more about their future than enjoying the present. This is precisely the reason why horoscopists, soothsayers, and fortune tellers are always in high-demand.  Michel de Nostredame (a.k.a Nostradamus)was a French soothsayer who lived in the 16th century. In his book *Les Propheties* (The Prophecies) he made predictions about important events to follow till the end of time. Nostradamus’ followers believe that his predictions are irrevocably accurate about major events including the World Wars and the end of the world. For instance in one of the prophecies in his book, which later became one of his most debated and popular prophesies, he wrote the following

*“Beasts ferocious with hunger will cross the rivers*
*The greater part of the battlefield will be against****Hister****.*
*Into a cage of iron will the great one be drawn,*
*When the child of Germany observes nothing.”*

His followers claim that *Hister* is an allusion to *Adolf Hitlor* where Nostradamus misspelled Hitlor’s name. One of the conspicuous thing about Nostradamus’ prophecies is that he never tagged these events to any date or time period. Detractors of Nostradamus believe that his book is full of cryptic pros (like the one above) and his followers try to force fit events to his writing. To dissuade detractors, one of his avid followers (based on his writing) predicted the month and the year for the end of the world as July 1999 – quite dramatic, isn’t it? Ok so of course nothing earth-shattering happened in that month of 1999 otherwise you would not be reading this article. However, Nostradamus will continue to be a topic of discussion because of the eternal human obsession to predict the future.

Time series modelling and ARIMA forecasting are scientific ways to predict the future. However, you must keep in mind that these scientific techniques are also not immune to force fitting and human biases. On this note let us return to our manufacturing case study example.

**ARIMA Model – Manufacturing Case Study Example**

Back to our manufacturing case study example where you are helping PowerHorse Tractors with sales forecasting for them to manage their inventories and suppliers. The following sections in this article represent your analysis in the form of a graphic guide.

|  |
| --- |
| You could find the data shared by PowerHorse’s MIS team at the following link [Tractor Sales](http://ucanalytics.com/blogs/wp-content/uploads/2015/06/Tractor-Sales.csv). You may want to analyze this data to revalidate the analysis you will carry-out in the following sections. |

Now you are ready to start with your analysis to forecast tractors sales for the next 3 years.

**Step 1: Plot tractor sales data as time series**

To begin with you have prepared a time series plot for the data. The following is the R code you have used to read the data in R and plot a time series chart.

|  |
| --- |
| data<-read.csv(“[location of data]“)data<-ts(data[,2],start = c(2003,1),frequency = 12)plot(data, xlab=”Years”, ylab = “Tractor Sales”) |



Clearly the above chart has an upward trend for tractors sales and there is also a seasonal component that we have already analysed an earlier article on [time series decomposition](http://ucanalytics.com/blogs/time-series-decomposition-manufacturing-case-study-example-part-2/).

**Step 2: Difference data to make data stationary on mean (remove trend)**

The next thing to do is to make the series stationary as learned in [the previous article](http://ucanalytics.com/blogs/arima-models-manufacturing-case-study-example-part-3/). This to remove the upward trend through 1st order differencing the series using the following formula:

|  |  |
| --- | --- |
| 1st Differencing (d=1) |  Y_{t}^{'}=Y_t -Y_{t-1}  |

The R code and output for plotting the differenced series is displayed below:

|  |
| --- |
| plot(diff(data),ylab=”Differenced Tractor Sales”) |



Okay so the above series is not stationary on variance i.e. variation in the plot is increasing as we move towards the right of the chart. We need to make the series stationary on variance to produce reliable forecasts through ARIMA models.

**Step 3: log transform data to make data stationary on variance**

One of the best ways to make a series stationary on variance is through transforming the original series through log transform. We will go back to our original tractor sales series and log transform it to make it stationary on variance. The following equation represents the process of log transformation mathematically:

|  |  |
| --- | --- |
| Log of sales |  Y_{t}^{new}=log_{10}(Y_t)  |

The following is the R code for the same with the output plot. Notice, this series is not stationary on mean since we are using the original data without differencing.

|  |
| --- |
| plot(log10(data),ylab=”Log (Tractor Sales)”) |



Now the series looks stationary on variance.

**Step 4: Difference log transform data to make data stationary on both mean and variance**

Let us look at the differenced plot for log transformed series to reconfirm if the series is actually stationary on both mean and variance.

|  |  |
| --- | --- |
| 1st Differencing (d=1) of log of sales |  Y_{t}^{new'}=log_{10}(Y_t) -log_{10}(Y_{t-1})  |

The following is the R code to plot the above mathematical equation.

|  |
| --- |
| plot(diff(log10(data)),ylab=”Differenced Log (Tractor Sales)”) |



Yes, now this series looks stationary on both mean and variance. This also gives us the clue that I or integrated part of our ARIMA model will be equal to 1 as 1st difference is making the series stationary.

**Step 5: Plot ACF and PACF to identify potential AR and MA model**

Now, let us create autocorrelation factor (ACF) and partial autocorrelation factor (PACF) plots to identify patterns in the above data which is stationary on both mean and variance. The idea is to identify presence of AR and MA components in the residuals. The following is the R code to produce ACF and PACF plots.

|  |
| --- |
| par(mfrow = c(1,2))acf(ts(diff(log10(data))),main=”ACF Tractor Sales”)pacf(ts(diff(log10(data))),main=”PACF Tractor Sales”)) |



Since, there are enough spikes in the plots outside the insignificant zone (dotted horizontal lines) we can conclude that the residuals are not random. This implies that there is juice or information available in residuals to be extracted by AR and MA models. Also, there is a seasonal component available in the residuals at the lag 12 (represented by spikes at lag 12). This makes sense since we are analyzing monthly data that tends to have seasonality of 12 months because of patterns in tractor sales.

**Step 6: Identification of best fit ARIMA model**

Auto arima function in forecast package in R helps us identify the best fit ARIMA model on the fly. The following is the code for the same. Please install the required ‘forecast’ package in R before executing this code.

|  |
| --- |
| require(forecast)ARIMAfit <- auto.arima(log10(data), approximation=FALSE,trace=FALSE)Summary(ARIMAfit) |

The best fit model is selected based on Akaike Information Criterion (AIC) , and Bayesian Information Criterion (BIC) values. The idea is to choose a model with minimum AIC and BIC values. We will explore more about AIC and BIC in the next article. The values of AIC and BIC for our best fit model developed in R are displayed at the bottom of the following results:

|  |  |
| --- | --- |
| **Time series:** | log10(Tractor Sales) |
| Best fit Model:**ARIMA(0,1,1)(0,1,1)[12]** |
|  | **ma1** | **sma1** |
| **Coefficients:** | -0.4047 | -0.5529 |
| **s.e.** | 0.0885 | 0.0734 |
| log likelihood=354.4 |
| AIC=-702.79 | AICc=-702.6 | BIC=-694.17 |

As expected, our model has I (or integrated) component equal to 1. This represents differencing of order 1. There is additional differencing of lag 12 in the above best fit model. Moreover, the best fit model has MA value of order 1. Also, there is seasonal MA with lag 12 of order 1.

**Step 6: Forecast sales using the best fit ARIMA model**

The next step is to predict tractor sales for next 3 years i.e. for 2015, 2016, and 2017 through the above model. The following R code does this job for us.

|  |
| --- |
| pred <- predict(ARIMAfit, n.ahead = 36)predplot(data,type=”l”,xlim=c(2004,2018),ylim=c(1,1600),xlab = “Year”,ylab = “Tractor Sales”)lines(10^(pred$pred),col=”blue”)lines(10^(pred$pred+2\*pred$se),col=”orange”)lines(10^(pred$pred-2\*pred$se),col=”orange”) |

The following is the output with forecasted values of tractor sales in blue. Also, the range of expected error (i.e. 2 times standard deviation) is displayed with orange lines on either side of predicted blue line.



Now, forecasts for a long period like 3 years is an ambitious task. The major assumption here is that the underlining patterns in the time series will continue to stay the same as predicted in the model. A short term forecasting model, say a couple of business quarters or a year, is usually a good idea to forecast with reasonable accuracy. A long term model like the one above needs to evaluated on a regular interval of time (say 6 months). The idea is to incorporate the new information available with the passage of time in the model.

**Step 7: Plot ACF and PACF for residuals of ARIMA model to ensure no more information is left for extraction**

Finally, let’s create an ACF and PACF plot of the residuals of our best fit ARIMA model i.e. ARIMA(0,1,1)(0,1,1)[12]. The following is the R code for the same.

|  |
| --- |
| par(mfrow=c(1,2))acf(ts(fit$residuals),main=”ACF Residual”)pacf(ts(fit$residuals),main=”PACF Residual”) |



Since there are no spikes outside the insignificant zone for both ACF and PACF plots we can conclude that residuals are random with no information or juice in them. Hence our ARIMA model is working fine.

However, I must warn you before concluding this article that randomness is a funny thing and can be extremely confusing. We will discover this aspect about randomness and patterns in the epilogue of this forecasting case study example.

**Sign-off Note**

I must say Nostradamus was extremely clever since he had not tagged his prophecies to any time period. So he left the world with a book containing some cryptic sets of words to be analysed by the human imagination. This is where randomness becomes interesting. A prophesy written in  cryptic words without a defined time-period is almost 100% likely to come true since humans are the perfect machine to make patterns out of randomness.

Let me put my own prophesy for a major event in the future. If someone will track this for the next 1000 years I am sure this will make me go in the books next to Nostradamus.

*A boy of strength will rise from the home of the poor*

*Will rule the world and have both strong friends and enemies*

*His presence will divide the world into half*

*The man of God will be the key figure in resolution of this conflict*

|  |  |
| --- | --- |
| Month-Year,Number of Tractor Sold |  |
| Jan-03,141 |  |  |  |
| Feb-03,157 |  |  |  |
| Mar-03,185 |  |  |  |
| Apr-03,199 |  |  |  |
| May-03,203 |  |  |  |
| Jun-03,189 |  |  |  |
| Jul-03,207 |  |  |  |  |
| Aug-03,207 |  |  |  |
| Sep-03,171 |  |  |  |
| Oct-03,150 |  |  |  |
| Nov-03,138 |  |  |  |
| Dec-03,165 |  |  |  |
| Jan-04,145 |  |  |  |
| Feb-04,168 |  |  |  |
| Mar-04,197 |  |  |  |
| Apr-04,208 |  |  |  |
| May-04,210 |  |  |  |
| Jun-04,209 |  |  |  |
| Jul-04,238 |  |  |  |  |
| Aug-04,238 |  |  |  |
| Sep-04,199 |  |  |  |
| Oct-04,168 |  |  |  |
| Nov-04,152 |  |  |  |
| Dec-04,196 |  |  |  |
| Jan-05,183 |  |  |  |
| Feb-05,200 |  |  |  |
| Mar-05,249 |  |  |  |
| Apr-05,251 |  |  |  |
| May-05,289 |  |  |  |
| Jun-05,249 |  |  |  |
| Jul-05,279 |  |  |  |  |
| Aug-05,279 |  |  |  |
| Sep-05,232 |  |  |  |
| Oct-05,204 |  |  |  |
|  |  |  |  |
|

|  |
| --- |
| Aug-14,848 |
| Sep-14,640 |
| Oct-14,581 |
| Nov-14,519 |
| Dec-14,605 |

 |  |  |  |
|  |  |  |  |
|  |  |  |  |