Build or Make your own ARIMA forecasting model?

Posted on Jan 28, Posted by Tom Reilly Category Forecasting

It is the standard Economic Question. Should I make or build? The question here is: Should I build my own ARIMA modeling system or use software that can do this automatically? We lay out the general approach how to consider approaches to building a robust model. Easier said than done does apply here.

Overview

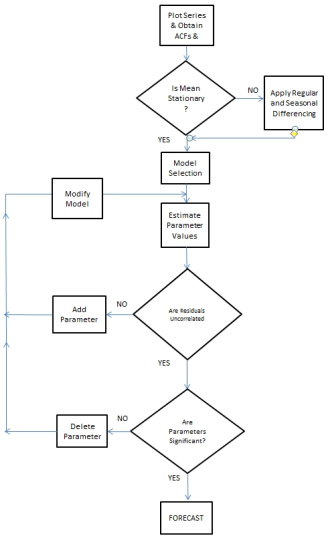
Autoregressive Integrated Moving Average (ARIMA) is a process designed to identify a weighted moving-average model specifically tailored to the individual dataset by using time series data to identify a suitable model. It is a rear-window approach that doesn't use user-specified helping variables; such as price and promotion. It uses correlations within the history to identify patterns that can be statistically tested and then used to forecast. Often we are limited to using only the history and no causals whereas the general class of Box-Jenkins models can efficiently incorporate causal/exogenous variables (Transfer Functions or ARIMAX).

This post will introduce the steps and concepts used to identify the model, estimate the model, and perform diagnostic checking to revise the model. We will also list the assumptions and how to incorporate remedies when faced with potential violations.

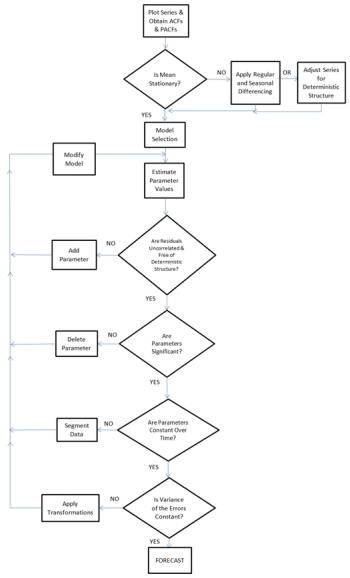
Background

Our understanding of how to build an ARIMA model has grown since it was introduced in 1976 (1). Properly formed ARIMA models are a general class that includes all well-known models except some state space and multiplicative Holt-Winters models. As originally formulated, classical ARIMA modeling attempted to capture stochastic structure in the data; little was done about incorporating deterministic structure other than a possible constant and/or identifying change points in parameters or error variance.

We will highlight procedures relevant to suggested augmentation strategies that were not part of the original ARIMA approach suggested in but are now standard. This step is often ignored as it is necessary that the mean of the residuals is invariant over time and that the variance of the final model's errors is constant over time. Here is the classic circa 1970.



Here is the flowchart revised for additions by Tsay, Tiao, Bell, Reilly & Gregory Chow (ie chow test)



The idea of modeling is to characterize the pattern in the data and the goal is to identify an underlying model that is generating and influencing that pattern. The model that you will build should match the history which can then be extrapolated into the future. The actual minus the fitted values are called the residuals. The residuals should be random around zero (i.e. Gaussian) signifying that the pattern has been captured by the model.



• For example, an AR model for monthly data may contain information from lag 12, lag 24, etc.

– i.e. Yt = A1Yt-12 +A2 Yt-24 + at

– This is referred to as an ARIMA(0,0,0)x(2,0,0)12 model

• General form is ARIMA(p,d,q)x(ps,ds,qs)s

Tools

The ARIMA process uses regression/correlation statistics to identify the stochastic patterns in the data. Regressions are run to find correlations based on different lags in the data. The correlation between successive months would be the lag 1 correlation or in ARIMA terms, the ACF of lag 1. We then examine if this month is related to one year ago at this time would then be apparent from evaluating the lag 12 correlation or in ARIMA terms, the ACF of lag 12. By studying the autocorrelations in the history, we can determine if there are any relationships and then take action by adding parameters to the model to account for that relationship. The different autocorrelations for the different lags are arranged together in what is known as a correlogram and are often presented using a plot. They are sometimes presented as a bar chart. We present it as a line chart showing 95% confidence limits around 0.0. The autocorrelation is referred to as the autocorrelation function (ACF).

- The key statistic in time series analysis is the autocorrelation coefficient (the correlation of the time series with itself, lagged 1, 2, or more periods).

The Partial Autocorrelation Function (PACF). The PACF of lag 12 for example is a regression using a lag of 12, but also uses all of the lags from 1 to 11 as well, hence the name partial. It is complex to compute and we won't bother with that here.

Now that we have explained the ACF and the PACF, let's discuss the components of ARIMA. There are three pieces to the model. The "I" means Integrated, but it simply means that you took differencing on the Y variable during the modeling process. The "AR" means that you have a model parameter that explicitly uses the history of the series. The "MA" means that you have a model parameter that explicitly uses the previous forecast errors. Not all models have all parts of the ARIMA model. All models can be re-expressed as pure AR models or pure MA models. The reason we attempt to mix and match has to do with attempting to use as few parameters as possible.

Identifying the order of differencing starts with the following initial assumptions, which are ultimately need to be verified:

1) The sequence of errors (a's) are assumed to have a constant mean of zero and a constant variance for all sub-intervals of time.

2) The sequence of errors (a's) are assumed to be normally distributed where the a's are independent of each other.

3) Finally the model parameters and error variance are assumed to be fixed over all sub-intervals.

We study the ACF and PACF and identify an initial model. If this initial model is significant, the residuals will be free of structure and we are done. If not, we identify that structure and add it to the current model until a subsequent set of residuals is free of structure. One could consider this iterative approach as moving structure currently in the errors to the model until there is no structure in the errors to relocate.

The following are some simplified guidelines to apply when identifying an appropriate ARIMA model with the following assumptions:

• Guideline 1: If the series has a large number of positive autocorrelations then differencing should be introduced. The order of the differencing is suggested by the significant spikes in the PACF based upon the standard deviation of the differenced series. This needs to be tempered with the understanding that a series with a mean change or a trend change can also have these characteristics.

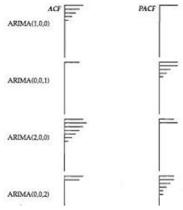
• Guideline 2: Include a constant if your model has no differencing; include a constant elsewhere if it is statistically significant.

• Guideline 3: Domination of the ACF over the PACF suggests an AR model while the reverse suggests an MA model. The order of the model is suggested by the number of significant values in the subordinate.

• Guideline 4: Parsimony: Keep the model as simple as you can, but not too simple as overpopulation often leads to redundant structure.

• Guideline 5: Evaluate the statistical properties of the residual (at) series and identify the additional structure (step-forward) required

• Guideline 6: Reduce the model via step-down procedures to end up with a minimally sufficient model that has effectively deconstructed the original series to signal and noise.
Over-differencing leads to unnecessary MA structure while under-differencing leads to overly complicated AR structure.



Guidelines for model form

Process	MA(q)	AR(p)	ARMA(p,q)
Autocorrelation function(ACF)	Cuts off	Doesn't cut off	Doesn't cut off
Partial Autocorrelation function(PACF)	Doesn't cut off	Cuts off	Doesn't cut off

If a tentative model exhibits errors that have a mean change this can be remedied in a number of ways;

1) Identify the need to validate that the errors have constant mean via Intervention Detection (2,3) yielding pulse, seasonal pulse/level shift/local time trends

2) Confirming that the parameters of the model are constant over time

3) Confirming that the error variance has had no deterministic change points or stochastic change points.

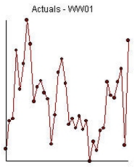
The tool to identify omitted deterministic structure is fully explained in references 2 and 3 as follows:

1) use the model to generate residuals

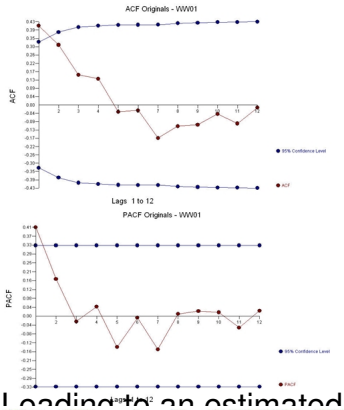
2) identify the intervention variable needed following the procedure defined in reference

3) Re-estimate the residuals incorporating the effect into the model and then go back to Step 1 until no additional interventions are found.

Example 1) 36 annual values:

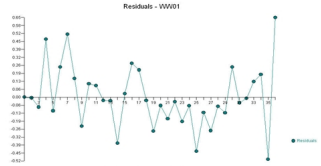


The ACF and the PACF suggest an AR(1) model (1,0,0)(0,0,0).

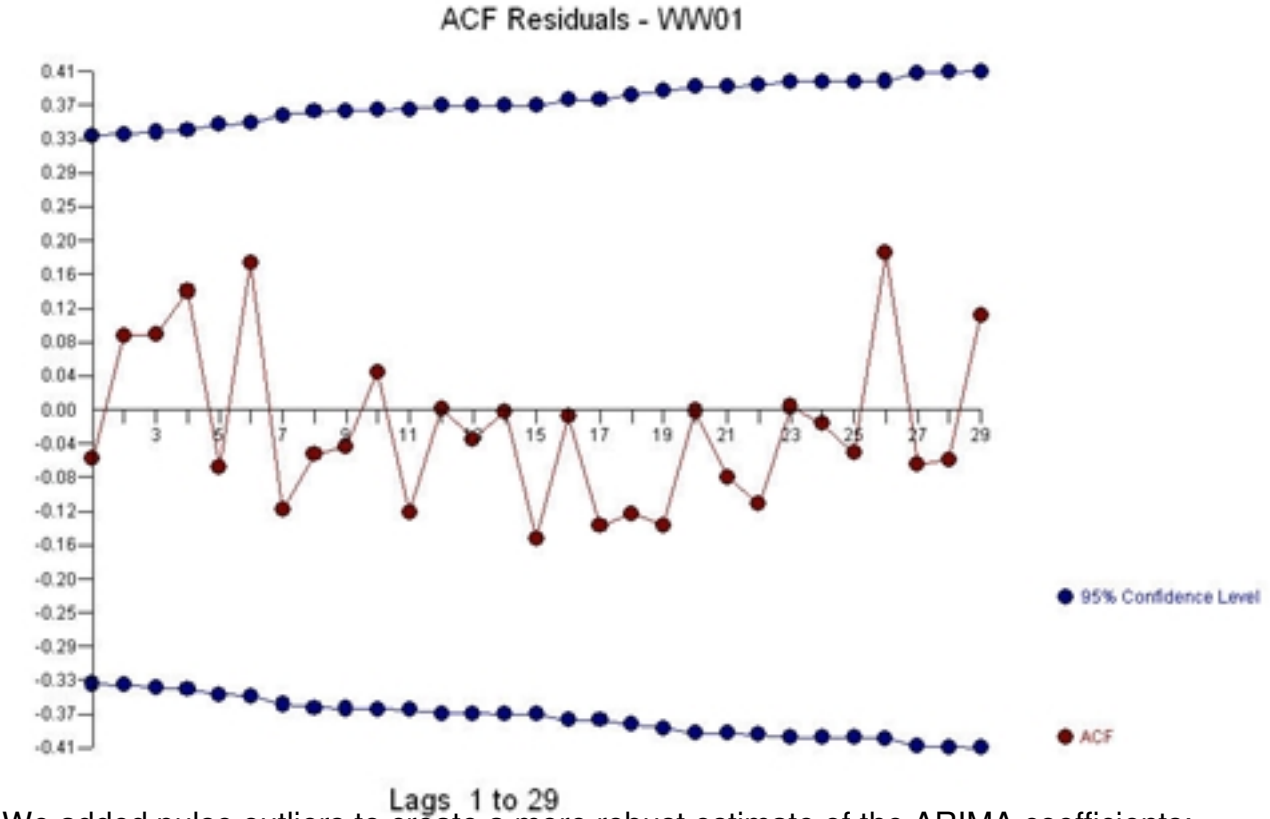


Leading to an estimated model (1,0,0)(0,0,0).

With the following residual plot, suggesting some “unusual values”:



The ACF and PACF of the residuals suggests no stochastic structure as the anomalies

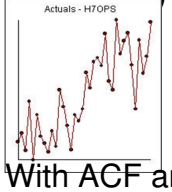


We added pulse outliers to create a more robust estimate of the ARIMA coefficients:

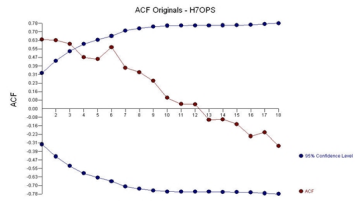


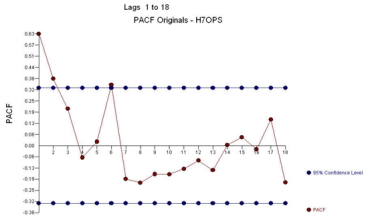
Y(T) = .75857		WW01	
+[X1(T)][(+ .280)]		:PULSE	7
+[X2(T)][(+ .441)]		:PULSE	4
+[X3(T)][(- .746)]		:PULSE	35
+[X4(T)][(- .254)]		:PULSE	25
+[X5(T)][(- .302)]		:PULSE	14
+[X6(T)][(- .300)]		:PULSE	9
+[X7(T)][(- .198)]		:PULSE	19
+[X8(T)][(+ .246)]		:PULSE	30

Example 2) 36 monthly observations:



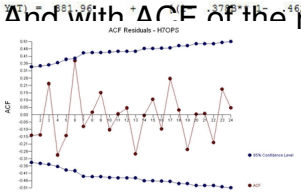
With ACF and PACF:



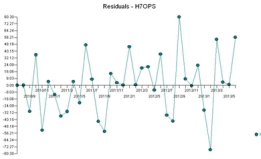


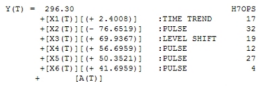
Loading [an estimated model: AR(2) (2,0,0)(0,0,0)

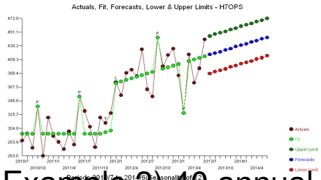
And with ACF of the residuals:



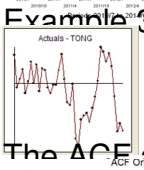
With the following residual plot:



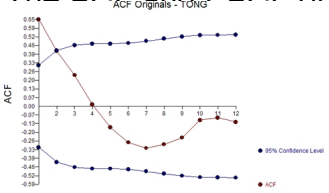




Example 3) 40 annual values:

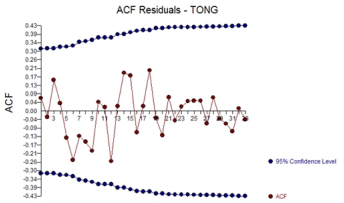


The ACF and PACF of the original series are:

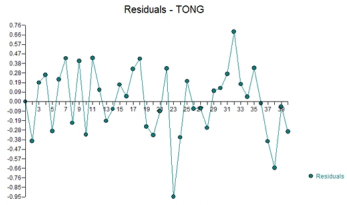


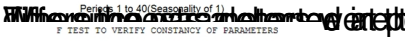
Suggesting a model (1,0,0,(0,0,0)

With a residual ACF of:



And residual plot:





F TEST TO VERIFY CONSTANCY OF PARAMETERS

CANDIDATE BREAKPOINT	F VALUE	P VALUE
18	6.0905	.0054861917
19	5.6979	.0073431038
20	6.2987	.0047097652
21	5.5801	.0080203805

This model for period 1-20 is:

MODEL COMPONENT	LAG (BOP)	COEFF	STANDARD ERROR	P VALUE	T VALUE
1CONSTANT		.963E-01	.587E-01	.1183	1.64
2Autoregressive-Factor # 1	1	.122	.228	.5911	.54

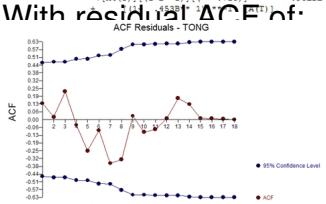
This model for period 21-40 is:

MODEL COMPONENT	LAG (BOP)	COEFF	STANDARD ERROR	P VALUE	T VALUE
1CONSTANT		-.668E-01	.805E-01	.4175	-.83
2Autoregressive-Factor # 1	1	.861	.128	.0000	6.72

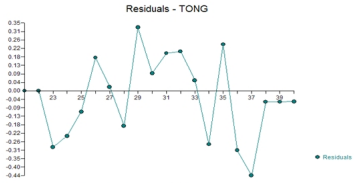
A final model using the last 20 values was:

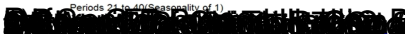
```
[(1-B**1)]Y(T) = +[X1(T)][(1-B**1)][(-  .886)]        :PULSE          29
                 +[X2(T)][(1-B**1)][(+  .286)]        :PULSE          32
                 +[X3(T)][(1-B**1)][(-  .345)]        :PULSE          38
                 +[X4(T)][(1-B**1)][(-  .715)]        :PULSE          24
```

With residual ACF of:



And residual plot of:





Tags: Tagged in: arima box-jenkins acf pacf