Great time series analysis example using the "Ages at Death of the Kings of England" Dataset

Posted on Jul 23, Posted by Tom Reilly Category Forecasting

This is a great example of how ignoring outliers can make you analysis can go very wrong. We will show you the wrong way and then the right way. A quote comes to mind that said "A good forecaster is not smarter than everyone else, he merely has his ignorance better organized".

A fun dataset to explore is the "age of the death of kings of England". The data comes form the 1977 book from McNeill called "Interactive Data Analysis" as is an example used by some to perform time series analysis. We intend on showing you the right way and the wrong way(we have seen examples of this!). Here is the data so you can you can try this out yourself: 60,43,67,50,56,42,50,65,68,43,65,34,47,34,49,41,13,35,53,56,16,43,69,59,48,59,86,55,68,51,33,49,67,77,81,67,71,81,68,70,77,56

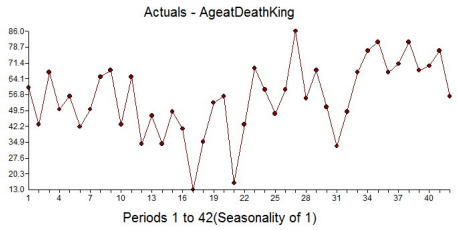
It begins at William the Conqueror from the year 1028 to present(excluding the current Queen Elizabeth II) and shows the ages at death for 42 kings. It is an interesting example in that there is an underlying variable where life expectancy gets larger over time due to better health, eating, medicine, cyrogenic chambers???, etc and that is ignored in the "wrong way" example. We have seen the wrong way example as they are not looking for deterministic approaches to modeling and forecasting. Box-Jenkins ignored deterministic aspects of modeling when they formulated the ARIMA modeling process in 1976. The world has changed since then with research done by Tsay , Chatfield/Prothero (Box-Jenkins seasonal forecasting: Problems in a case study(with discussion)" J. Roy Statist soc., A, 136, 295-352)
, I. Chang, Fox that showed how important it is to consider deterministic options to achieve at a better model and forecast.

As for this dataset, there could be an argument that there would be no autocorrelation in the age between each king, but an argument could be made that heredity/genetics could have an autocorrelative impact or that if there were periods of stability or instability of the government would also matters. There could be an argument that there is an upper limit to how long we can live so there should be a cap on the maximum life span.

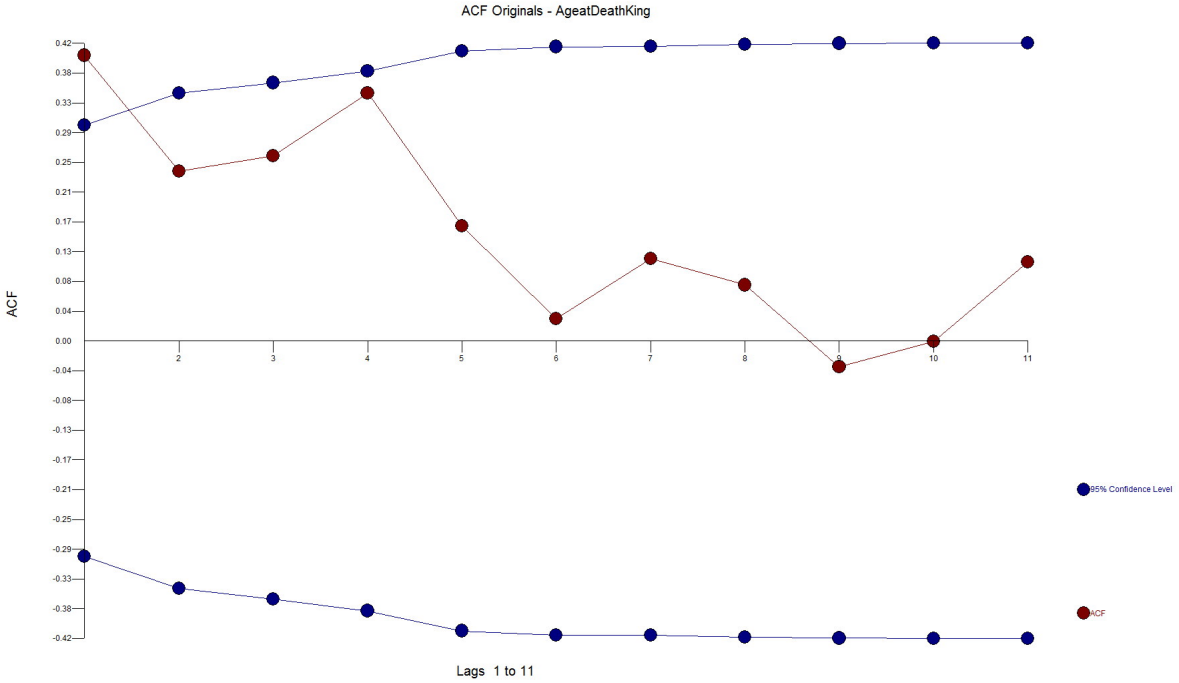
If you look at the dataset knew nothing about statistics, you might say that the first dozen obervations look stable and see that there is a trend up with some occasional real low values. If you ignored the outliers you might say there has been a change to a new higher mean, but that

is when you ignore outliers and fall prey to Simpson's paradox or simply put "local vs global" inferences.

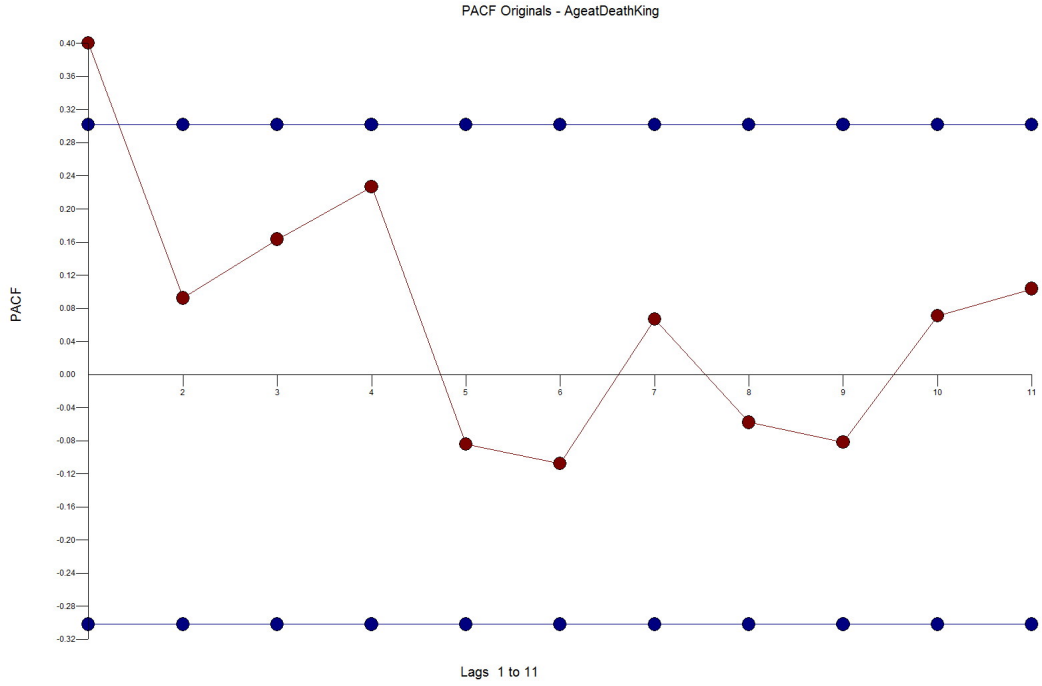
If you have some knowledge about time series analysis and were using your "rule book"on how to model, you might look at the ACF and PACF and say the series has no need for differencing and an AR1 model would suit it just fine. We have seen examples on the web where these experts use their brain and see the need for differencing and an AR1 as they like the forecast.



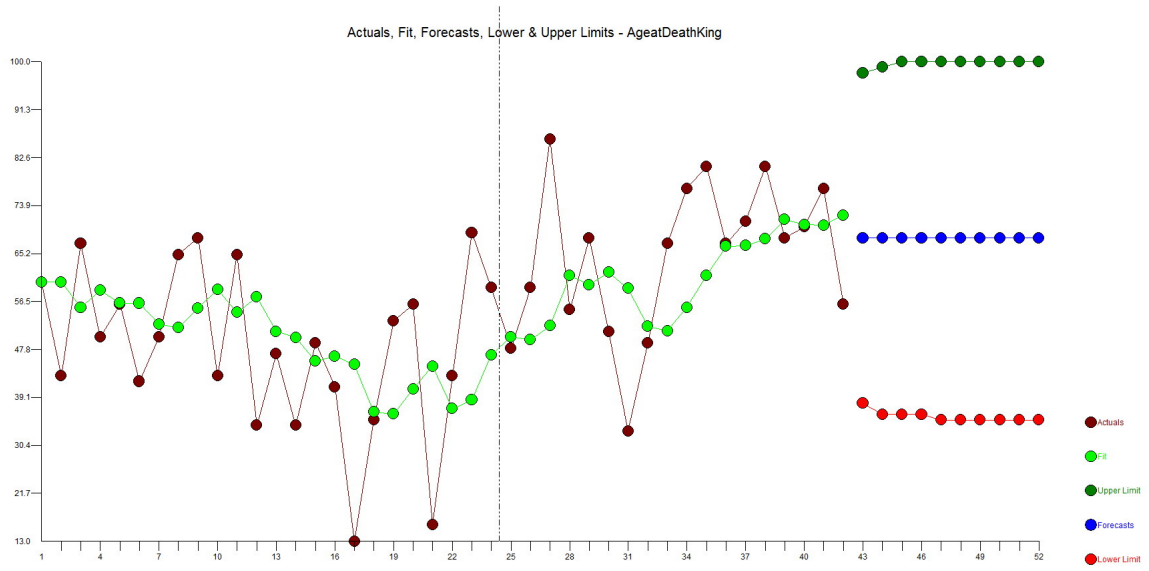
You might (incorrectly), look at the Autocorrelation function and Partial Autocorrelation and see a spike at Lag 1 and conclude that there is autocorrelation at lag 1 and then should then include an AR1 component to the model. Not shown here, but if you calculate the ACF on the first 10 observations the sign is negative and if you do the same on the last 32 observations they are positive supporting the "two trend" theory.

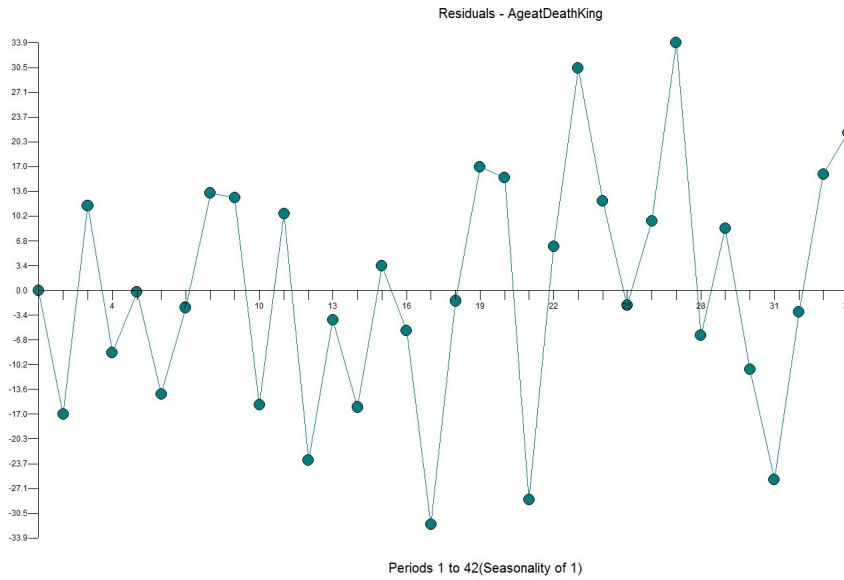


The PACF looks as follows:



Here is the forecast when using differencing and an AR1 model.





#	MODEL COMPONENT	LAG (BOP)	COEFF	STANDARD ERROR	P VALUE	T VALUE
1	CONSTANT		33.0	8.25	.0003	4.01
2	Autoregressive-Factor # 1	1	.401	.143	.0078	2.80

DIAGNOSTIC CHECK #3: THE TIAO TEST FOR CONSTANCY OF THE MEAN OF THE RESIDUALS

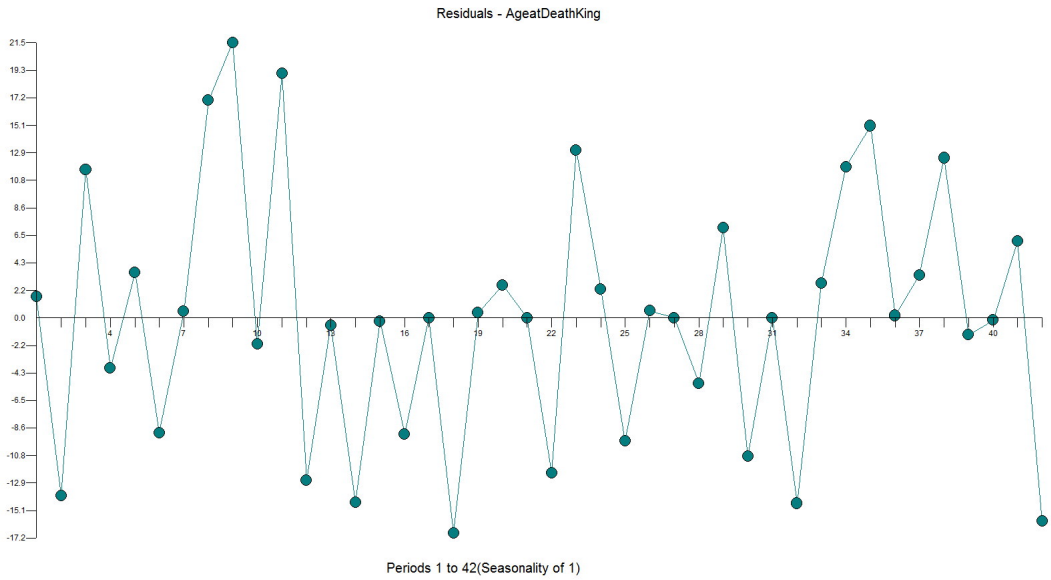
(ASSUMING THAT THE ORIGINAL SERIES FOLLOWS AN ARMAX PROCESS)

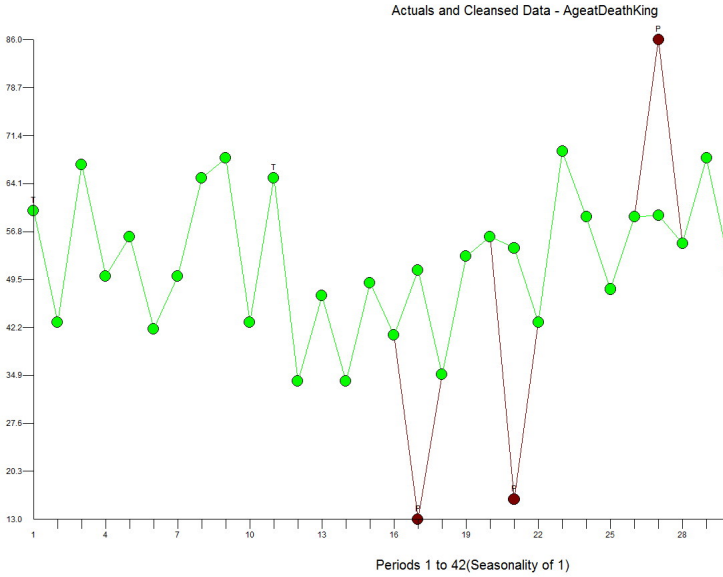
The Critical Value used for this test : .12

TYPE OF INTERVENTION (OUTLIER)	PATTERN	CYCLE	TIME (T)	DATE	REGRESSION WEIGHT	P VALUE
Additive	Trend	NA	3	3	-2.2897	.0001
Additive	Trend	NA	11	11	3.2309	.0001
Additive	Pulse	NA	21	21	-37.388	.0532
Additive	Pulse	NA	17	17	-36.721	.0350
Additive	Pulse	NA	31	31	-31.307	.0487

#	MODEL COMPONENT	LAG (BOP)	COEFF	STANDARD ERROR	P VALUE	T VALUE
1	CONSTANT		59.8	6.01	.0000	9.95
	INPUT SERIES X1 I-T00001		TIME	2		
2	Omega (input) -Factor # 1	0	-1.48	.755	.0582	-1.96
	INPUT SERIES X2 I-T00011		TIME	12		
3	Omega (input) -Factor # 2	0	2.32	.856	.0105	2.70
	INPUT SERIES X3 I-P00021		PULSE	22		
4	Omega (input) -Factor # 3	0	-38.3	9.87	.0004	-3.88
	INPUT SERIES X4 I-P00017		PULSE	18		
5	Omega (input) -Factor # 4	0	-37.9	9.94	.0005	-3.81
	INPUT SERIES X5 I-P00031		PULSE	32		
6	Omega (input) -Factor # 5	0	-29.6	9.89	.0050	-3.00
	INPUT SERIES X6 I-P00027		PULSE	28		
7	Omega (input) -Factor # 6	0	26.7	9.85	.0103	2.71

If you consider deterministic variables like outliers, level shifts, time [trends]





Tags: Tagged in: