Take Home Exam #2

Vishwa Koppisetti

DA6823

Time Series Analysis

1. Select a scientific, biomedical, business or other issue that appeals to you and go looking online for relevant time series data sets. The good news here is that there are tons of free and interesting time series data sets online. If you have problems locating them let me know and I will help. Tell me what issues you are investigating and cite the data sets so that I know where they came from. I want you to select at **least two different data sets** and perform the following analyses on each data set. For the first data set find one **without seasonality** in it. You will use this data set in steps 1 through 8. **The second of these data sets has to have seasonality in it and will only be used in step 9.** Suck the data set into your platform(s) of choice.

In order to perform time series analysis and forecast we choose two following datasets:

Non-Seasonal time series dataset: Total number of water consumers per month from Jan 1983 – April 1994.

## <https://datamarket.com/data/set/22wb/total-number-of-water-consumers-jan-1983-april-1994-missing-value-for-june-1988-66th-obs-estimated-by-intervention-analysis-london-united-kingdom#!ds=22wb&display=line>

## Seasonal Time Series data: Monthly gasoline demand in Ontario millions of gallons from 1960 to

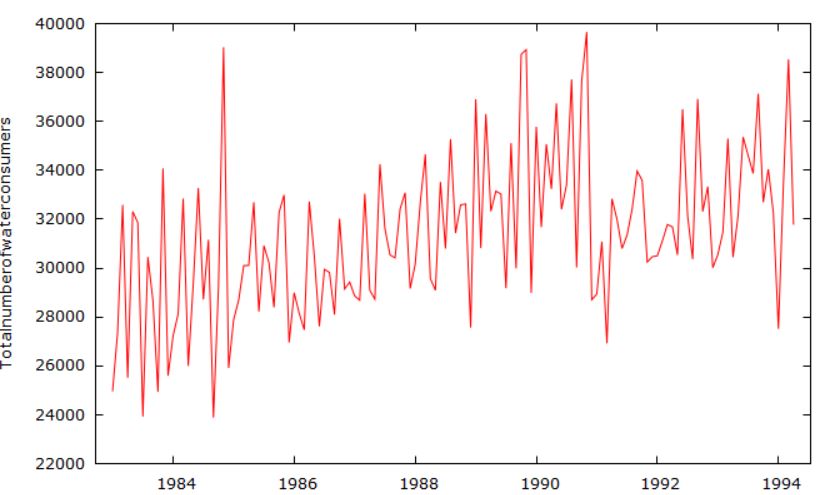
## 1975

<https://datamarket.com/data/set/22of/monthly-gasoline-demand-ontario-gallon-millions-1960-1975#!ds=22of&display=line>

**The following steps 2 through 8 are done only on data set #1 (the one without seasonality).**

1. Plot out your time series variable. Tell me using your Mark I eyeball whether or not you think the time series data set is stationary in terms of **constant mean** and also **constant variance**.

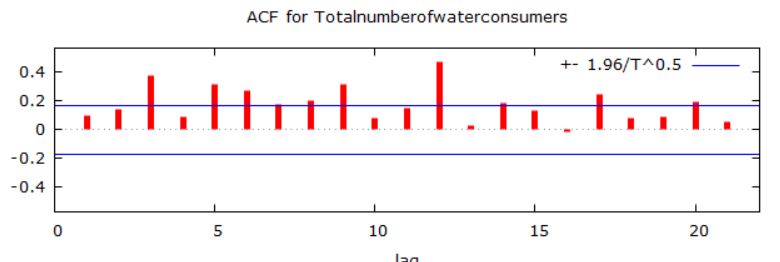
Just from eyeball analysis, the series does not seem to have constant mean and variance. Therefore, the series seems to be non-stationary.



1. Plot the ACF for the time series data set. Looking at ACF, does it look like there may be a trend or non-constant mean for each time series?

In the Autocorrelation plot we can observe that for most of the part, the autocorrelations are positive and don’t alternate into positive and negative lags as we would expect for a stationary series.

Also, the positive correlations reach beyond the blue line of p-value 0.05, which means they are non-zero. This suggests that the series has trend or non-constant mean.



1. Now let’s examine the time series data set using unit root tests. First use the KPSS test for the time series data set and tell me if the test suggests if there is a constant mean or not. Then see if you can confirm your KPSS evaluation using the Augmented Dickey Fuller (ADF) or the ADF-GLS test for each time series. What is your decision concerning constant mean?

KPSS Test:

H0: no evidence data has more than one mean (fits stationary criteria)

Halt: data has more than one mean (e.g. non-stationary)

The p-value from KPSS test is < 0.01 hence we have enough evidence to reject the null and conclude that the series is non-stationary. The test suggests that data has more than one mean or non-constant mean.

Augmented Dickey Fuller test:

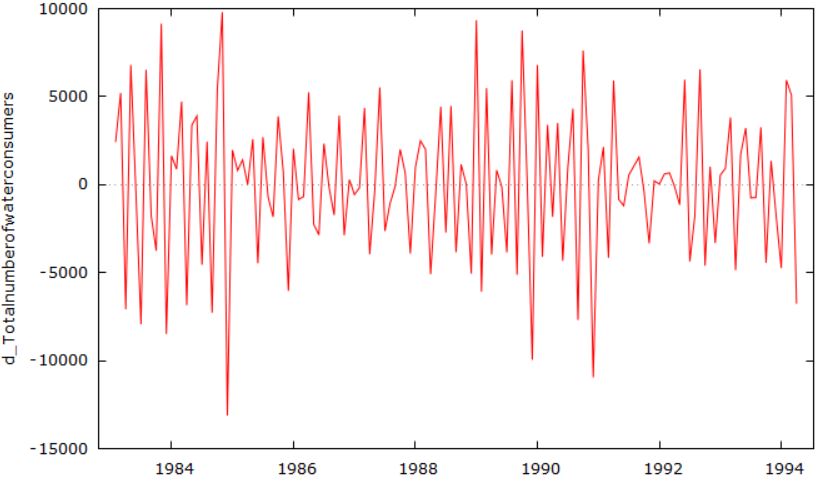
H0: no evidence data has single mean (non-stationary)

Halt: data have single mean (fits stationary criteria)

The p-value from ADF test is 0.6156. Because p-value >0.05 we do not reject the null and conclude that series is non-stationary. Hence the ADF test confirms the KPSS evaluation. The series does not have constant mean.

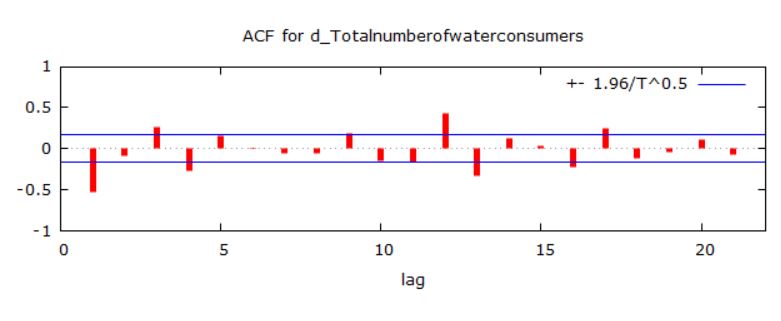
1. Review the decisions in step #4. If the test suggests that there is a non-constant mean then use differencing to create a new differenced variable for the time series data set.
   1. Plot out the data for the new differenced data set. Tell me if it looks like the differencing got rid of the trend or non-constant mean.

Since the series was found to have non-constant mean from KPSS and ADF test as well as the ACF plot, we decide to difference the series to the first order.



The differencing got rid of the trend and the series seems to have somewhat constant mean after differencing it.

* 1. Plot the ACF for the differenced time series. Tell me if this new ACF plot looks like there now is no trend.



From the ACF plot above we can observe that the pattern of autocorrelations varies from positive to negative peaks which indicates that there is no trend now after differencing. Though autocorrelations for lag 1 exceed the p-value of 0.05 signified by the blue line so the lag 1 autocorrelation is non-zero

* 1. Apply the KPSS test and the ADF or ADF-GLS test to the differenced data – does the trend disappear?

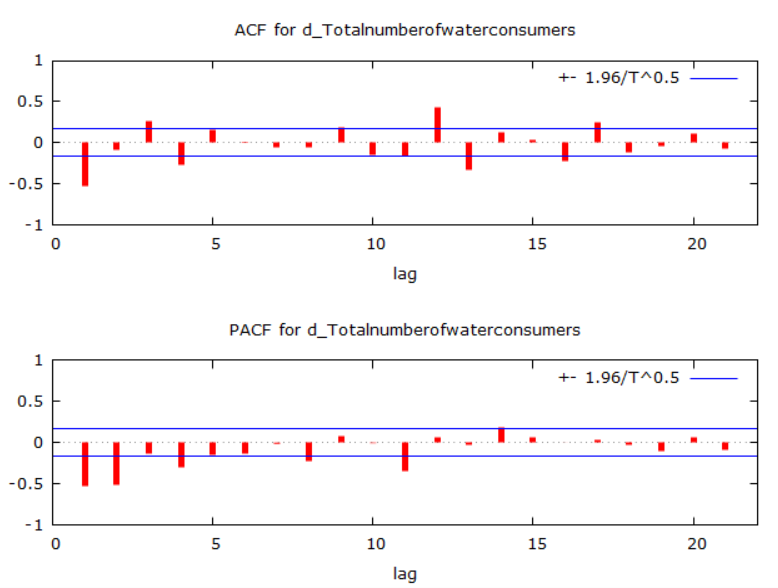
KPSS test gives a p-value >.10, so there is no evidence that data has more than one mean, so the trend disappears after differencing. Similarly, for ADF test, the p-value is < 0.05 (4.134e-011) which concludes that data has single mean, the trend disappears.

**Note: From this point onward through step 8, if the time series was differenced, use the differenced time series data set for all the rest of the questions. Otherwise you can use the undifferenced data set.**

1. Test each of the time series data sets for constant variance using the ARCH test (GRETL does this nicely). Tell me which ones might have issues with constant variance and so not be so nicely stationary. Note that we will not do anything about this issue for the moment, but it’s good to know.

The ARCH test gives a p-value 0.579 which is insignificant at p< 0.05 therefore we can conclude that the series has constant variance((homoscedasticity) and is stationary.

1. Plot the PACF for the time series data sets. Using the combined information from the ACF you plotted earlier along with the information in the PACF, tell me if you see autoregressive and/or moving average processes in the data set. To help with interpretation you may want to refer to online resources – here is a decent resource from Duke University [**https://people.duke.edu/~rnau/411arim3.htm**](https://people.duke.edu/~rnau/411arim3.htm) or Penn State <https://onlinecourses.science.psu.edu/stat510/node/64>



There is a single negative spike at the ACF plot which indicates MA (1) model to be used. There also seems to be negative spikes at lag1 and lag2 in PACF plot which indicate AR (2) but we can first do an AR to the first order and check if the remaining autocorrelation is corrected in the differenced series.

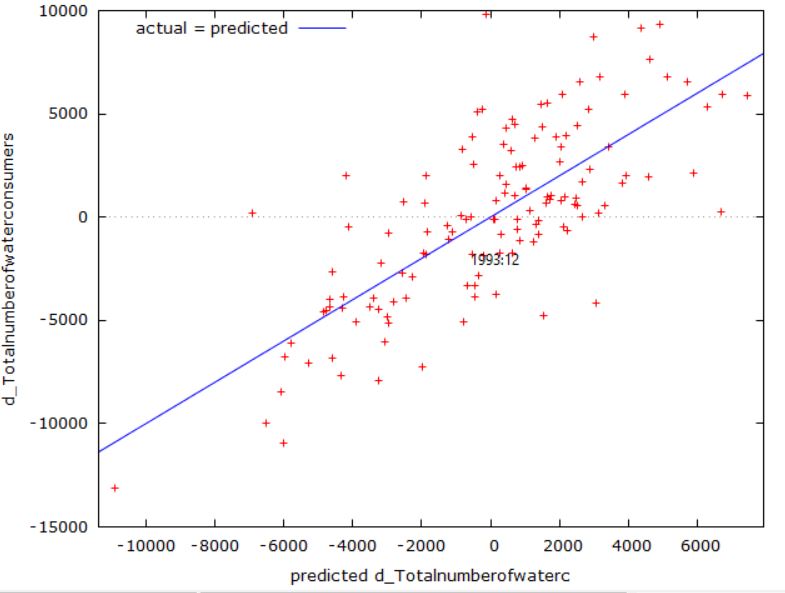
1. For your time series data set, experiment with different ARIMA models for them. As you try them, list out the results of the various models and
   1. Comment on how each one is working and compare it to the previous model using various metrics such as SBC, BIC, Box Leung, etc. Most students end up creating a small able with these statistics across the models tried so it is easy to compare them.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **AIC** | **BIC** | **HQ** |
| **(1,1,1)** | **2541.2** | **2552.8** | **2545.9** |
| **(0,1,1)** | **2543.4** | **2552.2** | **2547.04** |
| **(1,1,0)** | **2603** | **2611.8** | **2606.6** |
| **(0,1,0)** | **2646** | **2651.8** | **2648.4** |

Out of all the ARIMA models, ARIMA (1,1,1) is a better fit to the time series data since the AIC, BIC and HQ are lowest. The others lacking either MA or AR component have higher AIC, BIC and HQ values.

* 1. Plot the observed versus fitted data for the time series data set and comment on how well the model seems to be working

The ARIMA (1,1,1) model is able to moderately fit the actual data.

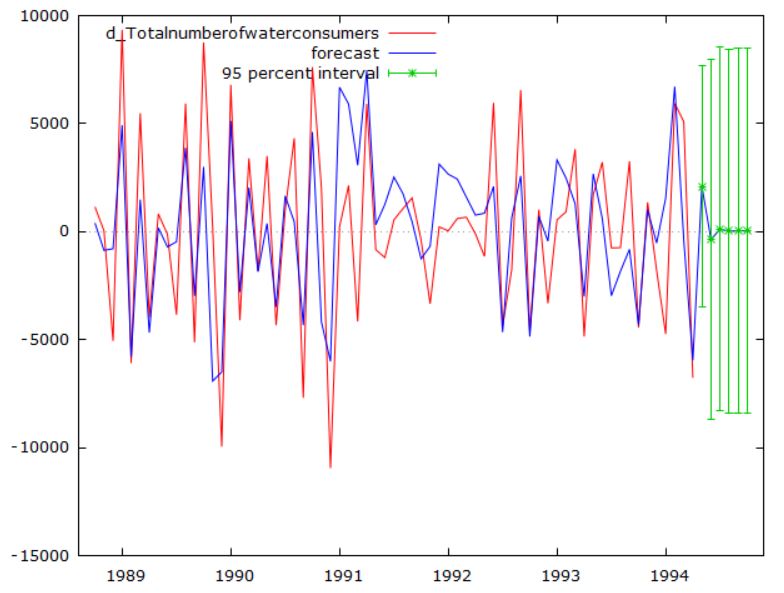
****

* 1. Pick one of the models as your favorite and tell me why you like that one the best.

I choose ARIMA (1,1,1) since it has lowest AIC, BIC and HQ values out of all the different ARIMA models.

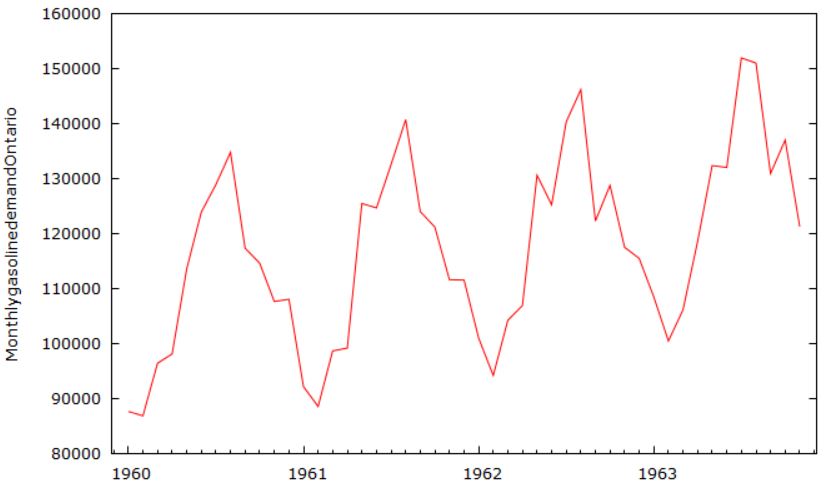
* 1. Forecast out your favorite model for the next 6 time periods and plot your time series plus the forecasted data. Does it look good or funky?

The model was able to predict the actual data and capture the variation in actual data to certain extent, but it could not forecast the data properly, after the first 2 years in to the future, rest of the years that followed them ended up being the mean itself which makes the model ineffective.

****

**Now switch to data set #2 – the one with seasonality in it. Perform the following steps.**

1. **For the time series data set with seasonality** **– start with the raw data before differencing here please!!!**
   1. Plot out the time series and suggest whether a type 1, type 2 or type 3 Holt Winter model should be applied and why.

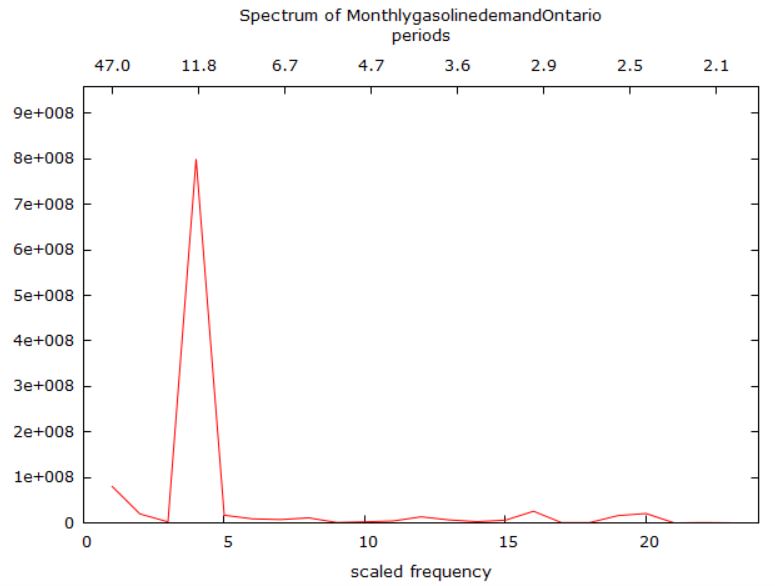
****

Since there is trend with unchanging seasonality, type 2 (Additive) Holt Winter model should be applied.

* 1. Eyeball the size of the period. What do you think it might be? Why is that?

The size of the period might be 12 months since we see peaks or declines in gasoline demand repeating for every 12-month time period

* 1. Use GRETL to do a periodogram for the data. What does the periodogram suggest might be the period length for the data?

****

Looking at the periodogram which has a single peak at period 11.8 we can conclude the period length as 12 approximately. From the x-axis in the time series plot, the series has 4 cycles each of period length 12.

* 1. If a type 2 or type 3 model, then apply a KPSS or ADF test to test for trend.

Since we have chosen type 2 model, we use KPSS or ADF to test for trend

From KPSS test, we get p-value as 0.076 since p-value > 0.05, we should be rejecting the null and conclude that the series fits the stationary criteria, hence no trend. But since the p-value is only slightly greater than 0.05 and nearly significant we can expect a slight trend if not a steep one and confirm of this by observing the time series plot that there is a slightly increasing trend.

From ADF test, however, suggests that series is stationary with highly significant p-value (3.072e-015), meaning there is no trend.

* 1. Decide the weights you will use for the three components of Winter Holt smoothing – constant, trend and seasonality – why these values? If you are using SAS then read about these weights in the proc docs, otherwise fish around in the R or Python docs.

The values for these three weights should be between 0 and 1. There are w1, w2 and w3.*w1* gives the weight for updating the constant component, *w2* gives the weight for updating the linear and quadratic trend components, and *w3* gives the weight for updating the seasonal component. They are certain default weights among which one set of values are 0.05, 0.001, 0.001.

* 1. Run the Holt Winter model and then using sgplot or your other favorite plotting poison plot the actual data and the fitted/forecast data on the same graph. How did the Holt-Winter model do in terms of forecasting?

Used ESM to run addwinters model and forecasted data.

**proc** **esm** data=seasonal\_add\_winters

lead=**12** /\* Forecast 12 months into the future \*/

print=(ESTIMATES STATISTICS SUMMARY)

plot=forecasts;

id Month interval=month;

forecast gasoline\_milgal / model=addwinters;

**run**;



On observing the graph that plots both the actual and fitted data, we can conclude that the holt additive winters model forecasts the actual data quite well, in terms of capturing the variation and the confidence bands width. Since the predicted graph is closely following the actual data that tells us that the variation in actual data is captured well by the fitted data. Since the width of the confidence band is narrow, we can conclude that it forecasts the data quite well.

* 1. Next run the same data set using the Unobserved Components Model time series analytic technique. Interpret the significance analysis of components table (based on final components in terms of trend, irregular (ARMA) and seasonality components in the data set – that is, which components are statistically significant?

**proc** **ucm** data=seasonal\_add\_winters;

id Month interval=month;

model gasoline\_milgal;

irregular plot=smooth;

level plot=smooth;

slope plot=smooth;

cycle plot=smooth;

season length=**12** type=dummy plot=smooth;

forecast lead=**12** back=**12** alpha=**0.05** plot=forecasts;

outlier;

**run**;

| **Significance Analysis of Components (Based on the Final State)** | | | |
| --- | --- | --- | --- |
| **Component** | **DF** | **Chi-Square** | **Pr > ChiSq** |
| **Irregular** | 1 | 0.00 | 0.9985 |
| **Level** | 1 | 155030 | <.0001 |
| **Slope** | 1 | 68.01 | <.0001 |
| **Cycle** | 2 | 29.97 | <.0001 |
| **Season** | 11 | 2760495 | <.0001 |

From the above table we can observe that the cycle, level, season and slope components are highly significant with p-values <0.0001 each but the irregular component is not significant with p-value 0.9751. By this we can conclude that the significant components of cycle, level, season and slope are responsible for the prediction behavior and irregular component is not contributing to the prediction in this scenario.

* 1. Have the UCM model produce fitted values for the existing data and forward 12 periods into the future and plot the original time series as well as the fitted/forecast data as well.

**proc** **ucm** data=seasonal\_add\_winters;

id Month interval=month;

model gasoline\_milgal;

irregular plot=smooth;

level plot=smooth;

slope plot=smooth;

cycle plot=smooth;

season length=**12** type=dummy plot=smooth;

forecast lead=**12** back=**0** alpha=**0.05** plot=forecasts;

outlier;

**run**;

With UCM we plot the actual and fitted values for existing data and forecast 12 months into the future. On Observing the plot, we can conclude that the fitted plot follows the actual data closely and the confidence bands are narrower too. Therefore, we can say that UCM model forecasts the data well

