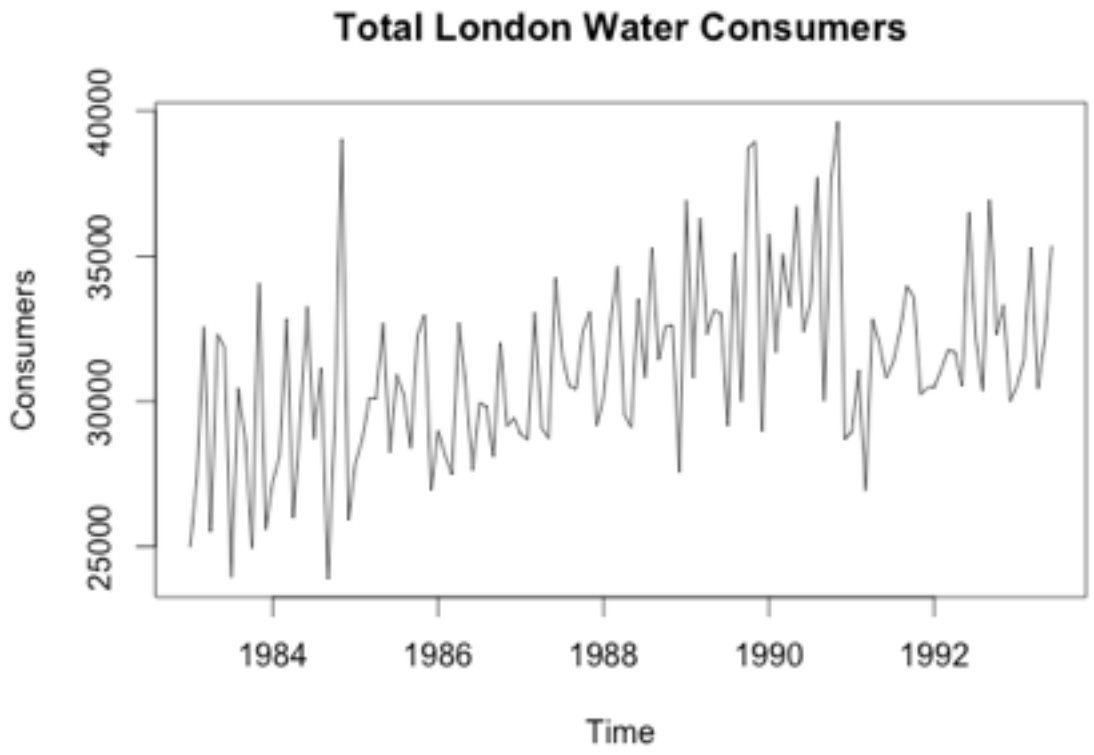
James Henderson
06/2/2016
Terdik

Time Series Analysis on London Water Consumers
Jan 1983- Jun 1993

For this project I am analyzing the data for the total number of water consumers in London from January 1983- Jun 1993. There is a missing value for June 1988, which had been filled in by intervention analysis. I changed this value to the average of its neighboring values as requested by my TA. The purpose of this analysis is to forecast 10 future values of data, as well as precast the last 10 values of the given data. The 10 future values in this case actually exist, however I am not using them for my model. I am treating them as if they do not exist until I forecast them, at which point I will use them to compare with my predictions. The time series plot of the original data is shown in figure 1.

Figure 1.1



Stationarity and Transformations:

By only looking at the graph, there appears to be an upward trend, as well as some seasonality. This means that the data is not stationary. The models that will be used cannot work with non stationary data, so I must apply transformations in order to make it stationary. This is done by removing the trend and seasonality.

My suspicions are confirmed in Figure 1.2 where we can see an obvious trend, as well as seasonality which looks to be about every 12 months. I decide to take the 1st difference of the data as well as the log to get rid of any trend as well as a non uniform variance. I also take the 12th difference of to get rid of the seasonality. In figure 1.2 it seems that my transformations have paid off as the trend has disappeared as well as the seasonality. The residuals are now much more normal as well. This is a good sign that our data is now stationary which allows us to move on in our analysis.

Figure 1.2

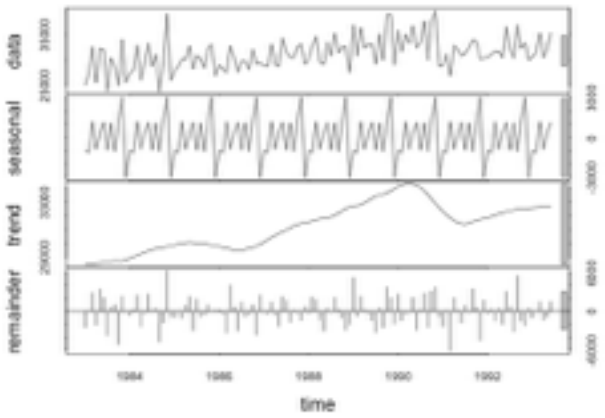
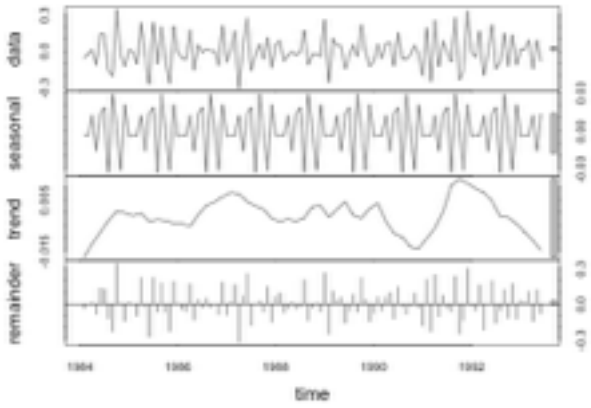


Figure 1.3



In figures 2.1, and 2.2 we can also see how the ACF and PACF reflect a positive result from the transformation. The ACF is the Autocorrelation function, and it shows how values are correlated with other values t points away. The PACF is the Partial autocorrelation function, and it is similar to the Autocorrelation function except it accounts for all points in between the two that you are looking at. The plots of these two functions can give us a good idea of how many parameters we will need in our model. There are two main sets of parameters that we will use in our model. They are AR and MA. These stand for autoregressive and moving average respectively. In addition there are seasonal versions of these parameters that we will also include.

The original data's ACF doesn't cut off, or exponentially decrease very nicely so this shows non-stationarity. The ACF of the transformed data fixes this, and shows that there is possibly an MA(2) component to the model, because it cuts off after 2 lag. The PACF shows there could be a AR(2) component as well because it also pretty much cuts off after 2 lag. There is also a repeating small spike in the ACF which may be a seasonal MA component. The AR and MA stand for autoregressive and moving average respectively, and they are 2 major components in the model we will use. In addition there are seasonal versions of these parameters that we will also include.

Figure 2.1

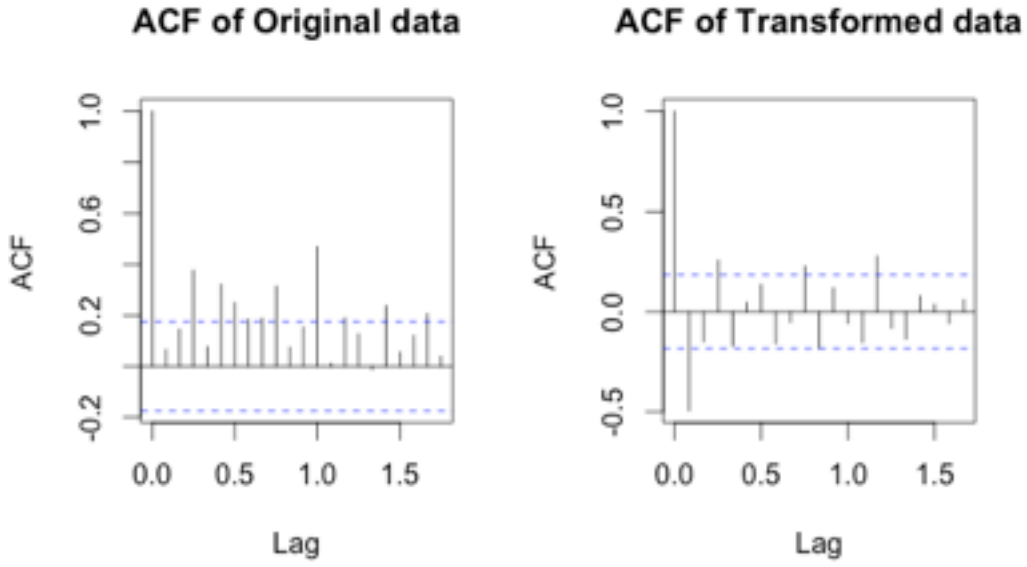
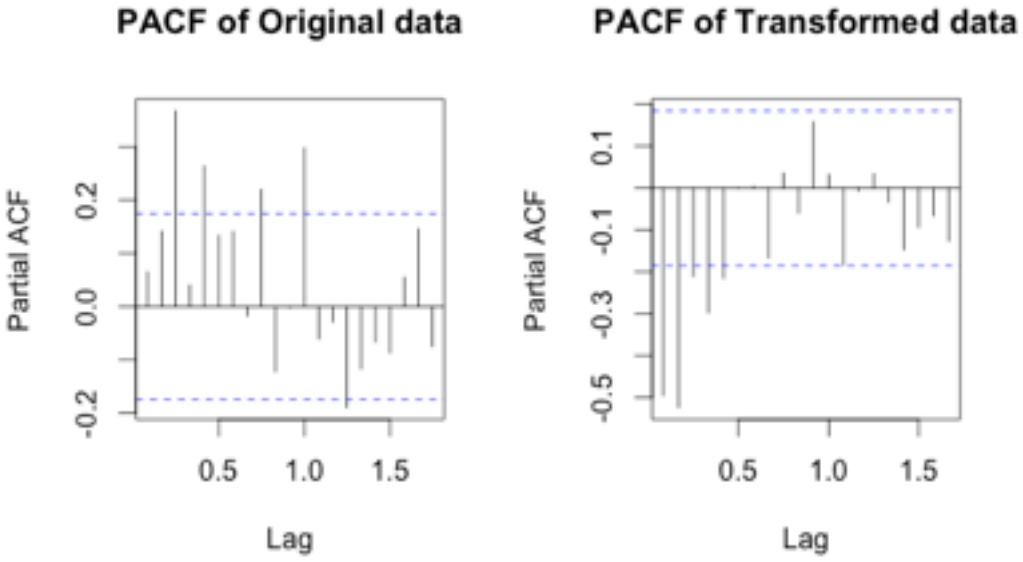


Figure 2.2



Model Selection:

Now it is time to start finding a model that will be satisfactory for our data. I am using a SARIMA model for this. This stands for Seasonal Autoregressive Integrated Moving Average. What this means is that our model will include a combination of AR, MA, Seasonal AR, Seasonal MA, as well as any differences we want to apply to the data.

Based on my inferences from the ACF and PACF, as well as the seasonality of the data, I had an idea that the model would be something similar to SARIMA(2,1,2,1,0,0,12). I tinkered around with a lot of the combinations of parameters, and from this I was convinced that my first 3 nonseasonal parameter assumptions were correct. This was because most models I tried without these parameters did not pass the Ljung-box statistic. I also am convinced that the 12th difference is correct. It was more tricky trying to hunt down a good model however, when it comes to the seasonal parameters. After playing around I came up with 3 good models to choose from. They are shown in table 3.1. The diagnostics for models 1, 2, and 3 are shown in the appendix as figures A.1, A.2, and A.3 respectively.

Each of these three models has normal standardized residuals, ACF of residuals which are 0 for the most part, satisfactory Q-Q plots, and p values for Ljung-box statistics which all pass. The Ljung-box statistic is a hypothesis test that tests whether or not all of the ACF residuals are 0, vs just one at a time as in the plot of the ACF residuals.

Table 3.1

Model	SARIMA	AIC	BIC
Model 1	SARIMA(2,1,2,1,0,1,12)	16.68557	15.84314
Model 2	SARIMA(2,1,2,1,0,0,12)	16.75394	15.88900
Model 3	SARIMA(2,1,2,0,0,1,12)	16.70645	15.84151

From looking at the AIC and BIC criterion, the correct model is now narrowed down to either Model 1 or Model 3. In Model 1 the AIC is lower, however in Model 3 the BIC is lower. I am choosing to go with model 3 because

the BIC penalizes for more parameters in the model, so I believe it is a more accurate representation.

Table 3.2

Parameter	AR(1)	AR(2)	MA(1)	MA(2)	SMA(1)
Estimate	-0.881	-0.291	-0.138	-0.632	0.522
Standard Error	0.176	0.110	0.172	0.183	0.109
Confidence Interval 95%	[-1.23,-.536]	[-.507,-.075]	[-.475, .199]	[-.991, -.273]	[.308, .736]

In table 3.2 we can see the estimations for the parameters as well as the standard error and confidence intervals. From this we can get an equation for our model:

$$(1+.881B+.291B^2)(1-B)X_t = (1-.138B-.632B^2)(1-.522B^{12})W_t$$

Where X_t is the value of our data at lag t, and W_t is a white noise random variable with distribution

$$W_t \sim N(0,2454.3^2)$$

The B^h is a backshift operator, which when multiplied to a time series variable transforms it into the value of the data at t-h. For instance $B(X_t)= X_{t-1}$.

Forecasting:

Using the model from this equation we can finally forecast our values! Figure 4.1 shows the forecasting predictions, plotted with their actual values. The blue circles show the predicted last 10 values, with the green dashed line showing the prediction intervals for them. The red circles show the predicted 10 future values, with the purple dashed lines being the prediction interval. The black solid line shows the actual past 10 values, and the orange line shows the actual 10 future values, which were not included while creating the model.

The model does a pretty good job forecasting the 10 future values, and an even better job predicting the 10 past values. Figure 4.2 shows the previous 10

values and 4.3 shows the 10 forecasted future values. Overall I am happy with my model. It may seem like there is a high variance but in the context of the scale of the data I believe it makes sense.

Figure 4.1

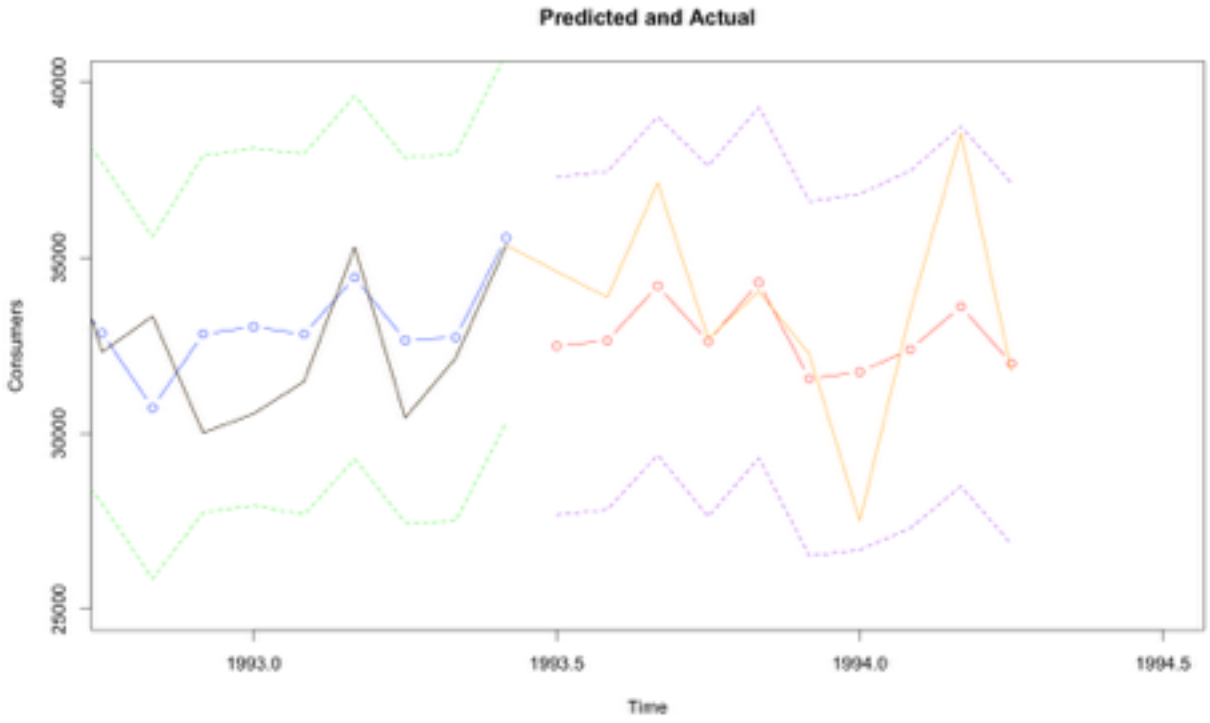


Figure 4.2

```
     Jan      Feb      Mar      Apr      May      Jun Jul Aug      Sep      Oct      Nov      Dec
1992                                                          34718.68 32858.13 30728.48 32832.67
1993 33037.76 32826.25 34440.88 32651.73 32735.24 35566.81
```

Figure 4.3

```
     Jan      Feb      Mar      Apr May Jun      Jul      Aug      Sep      Oct      Nov      Dec
1993                                      32489.82 32640.49 34199.21 32619.27 34296.31 31560.58
1994 31753.56 32391.45 33614.04 31987.28
```

Appendix

Figure A.1

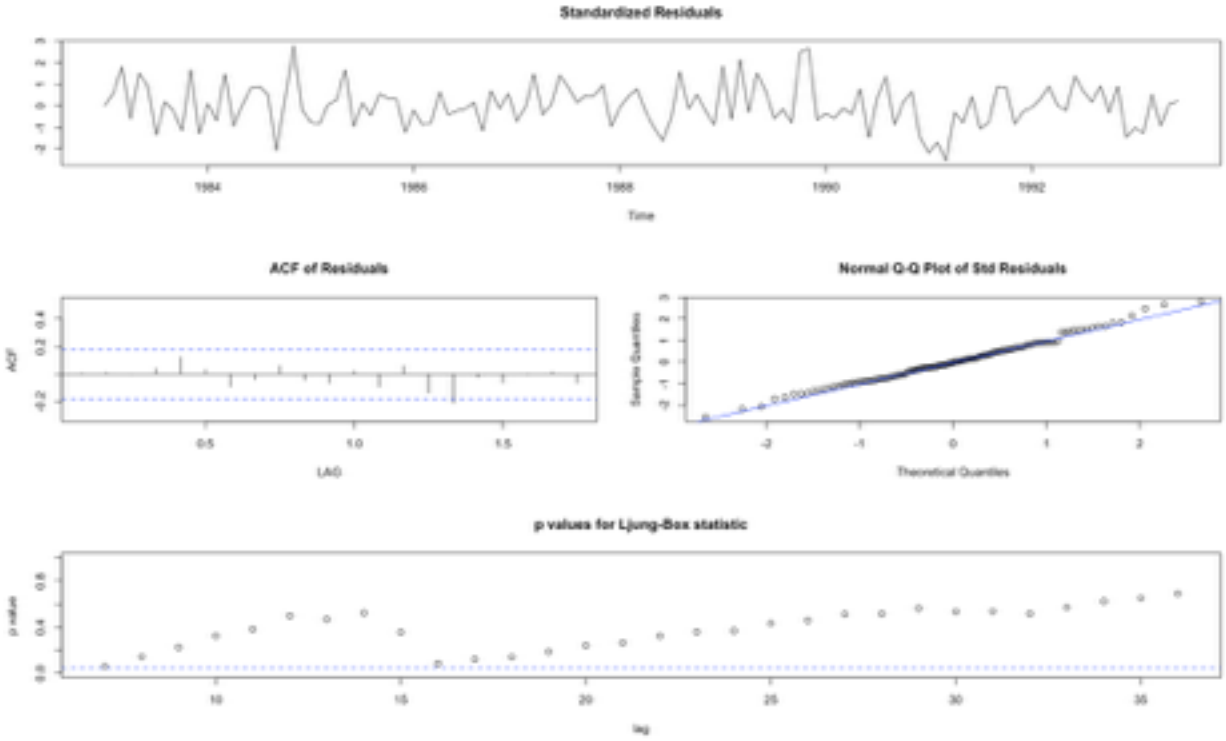


Figure A.2

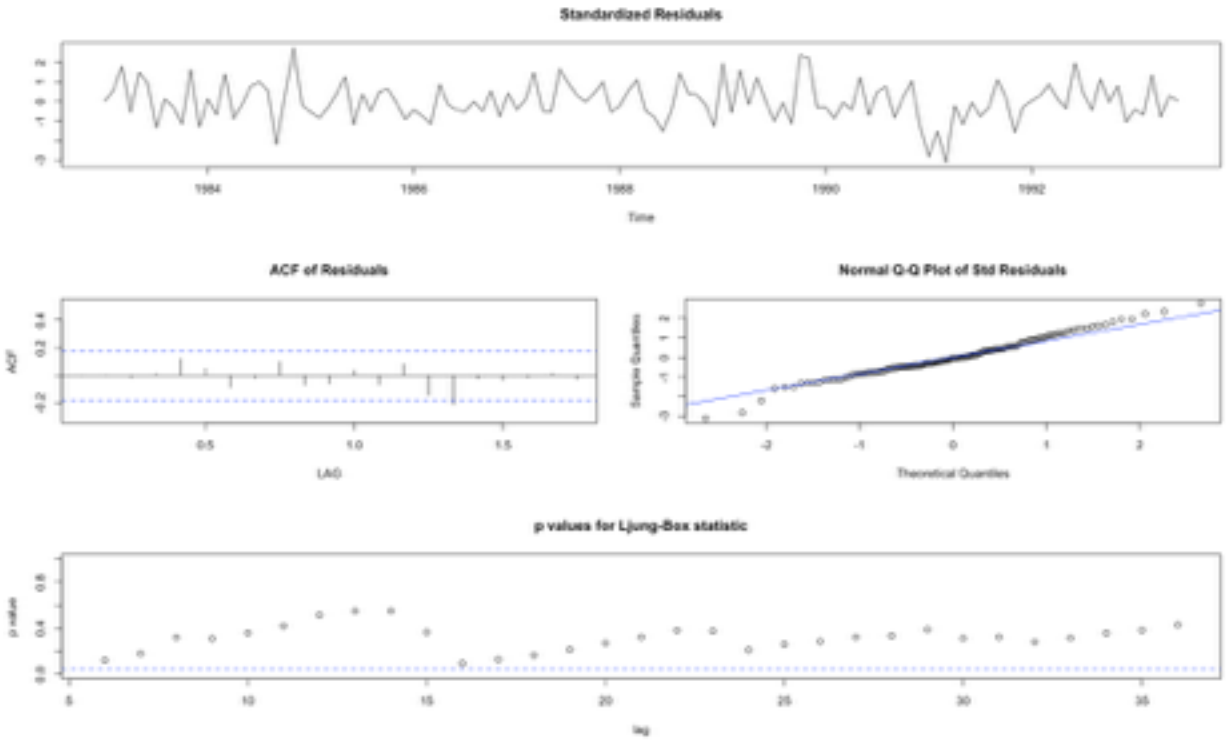
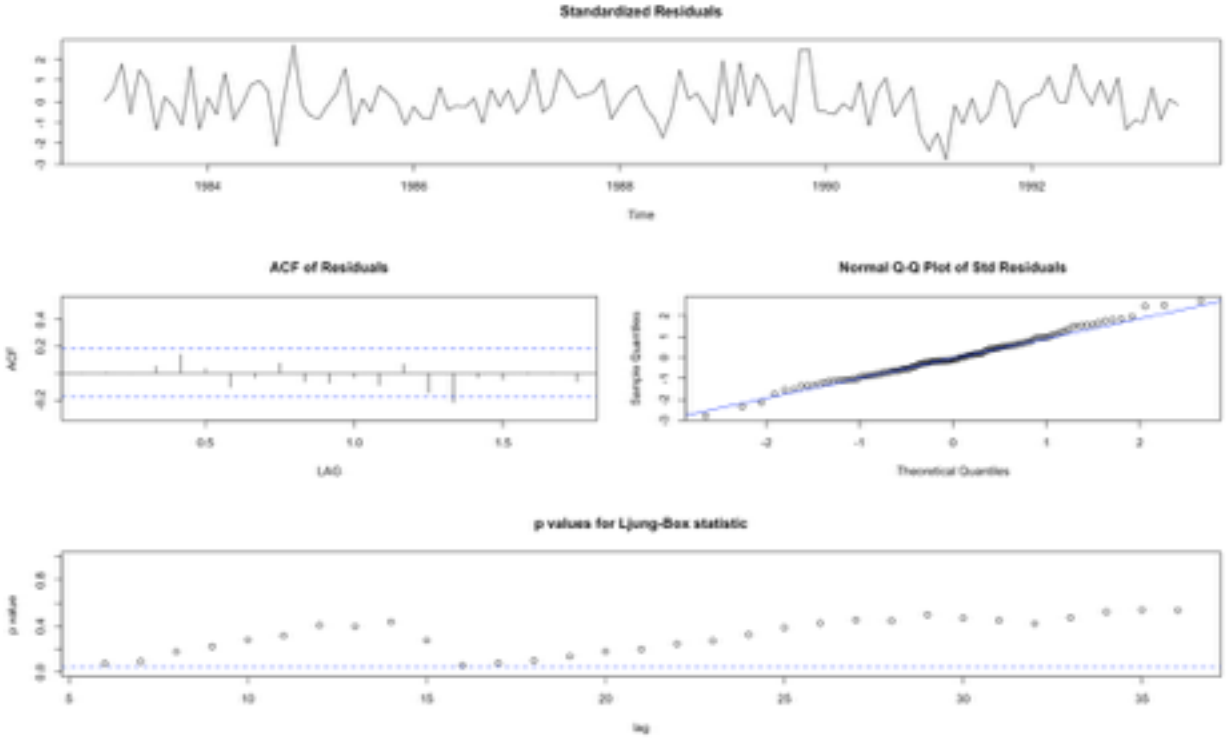


Figure A.3



* Data was obtained from **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**
* Source: Time Series Data Library (citing: Hipel and McLeod (1994))