QMM 1002 Case Study 2

# Himani Dave

# A00178443

# Due: April 26th, 2019

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 marks | 1 mark | 2 marks | | 3 marks | | Gr. |
| **Data Analysis** | No moving average model is included | A moving average model is included with more than 1 error | | A moving average model is included with 1 error | | A correct moving average model is included |  |
| No exponential smoothing model is included | An exponential smoothing model is included with more than 1 error | | An exponential smoothing model is included with 1 error | | A correct exponential smoothing model is included |  |
| No random sample is taken | A random sample is taken with at least one error | | A correct random sample is taken from the population | | N/A |  |
| No chi-square test is included | A chi-square test is included with more than 1 error | | A chi-square test is included with 1 error | | A correct chi-square test is included |  |
| **Report** | The key questions are not answered | The key questions are answered with some missing detail | | The key questions are clearly answered in full detail | | N/A | **X3** |
| No correct hypotheses are included for the test conducted | Hypotheses are included for the test conducted with at least 1 error | | Correct hypotheses are included for the test conducted | | N/A |  |
| No results of the statistical test and an interpretation are included with error | The results of the statistical test and an interpretation are included with error | | The results of the statistical test and a correct interpretation are included | | N/A |  |
| No assumptions and conditions are checked | Conditions and assumptions are checked with 1 error | | All conditions and assumptions are checked correctly and in detail | | N/A |  |
| No plot included | Plots are included and interpreted with at least 1 error for:  time series data, MA model, ES model, forecasts of model, quality of model, differences in statistical test | | Plots are included and correctly interpreted for: time series data, MA model, ES model, forecasts of model, quality of model, differences in statistical test | | N/A | **X6** |
| Incorrect or no time series models are chosen with | The best time series model is chosen with more than 1 error and/or explanation missing | | The best time series model is chosen with 1 error and/or explanation missing | | The best time series model is chosen and an explanation is given for MA and ES model | **X2** |
| The model forecasts are not included | The model is used to forecast 5 periods into the future and an interpretation of the values is given with error | | The best model is used to forecast 5 periods into the future and an interpretation of the values is given | | N/A |  |
| **Format** | Numerous spelling errors, incorrect punctuation, and/or severe errors in grammar so that the report is hard to understand. | Frequent spelling errors, incorrect punctuation, and grammar problems that sometimes interfere with understanding. | | Occasional lapses in spelling, punctuation, grammar, but not enough to seriously distract the reader. | | Very few spelling errors, correct punctuation, grammatically correct, complete sentences. |  |
| Incorrect notation and terminology used throughout | More than 2 errors in notation and terminology | | 1-2 minor errors in notation and terminology | | No errors in notation or terminology |  |
| There are not 3 sections of the report | There are 3 sections of the report as given in the template | | N/A | | N/A |  |
| **Bonus** | No additional test(s) are conducted |  | | N/A | | Additional time series model is included and interpreted correctly | **0 or 3 Bonus** |

|  |
| --- |
|  |
| /50 | |

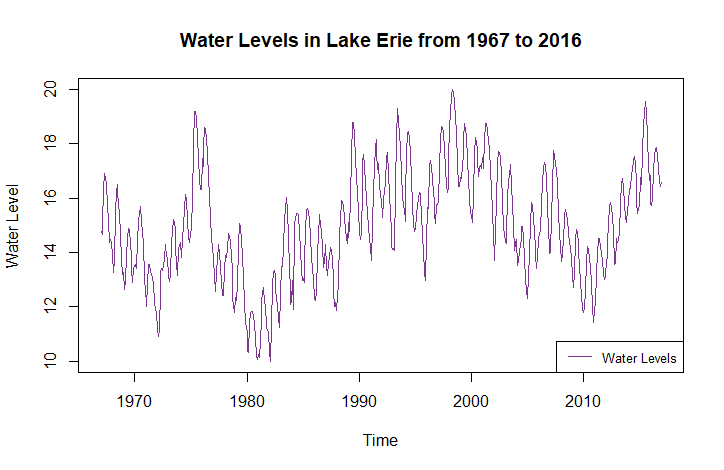
# Introduction

An environmental monitoring agency wants to monitor the water levels of Lake Erie to ensure that they are within the normal range. The given data has the water levels recorded monthly from January 1967 to December 2016. The levels have been on the higher side lately. If the water levels go too high, there is a risk of floods. On the other hand, if the water levels go too low, there is a risk of receding shorelines. Therefore, the agency wants the analysis of the given data in order to answer the following questions:

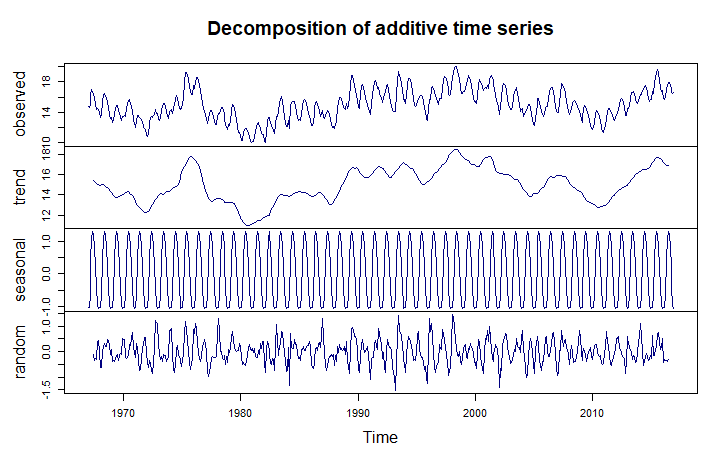
1. What is the average water level of Lake Erie?
2. What is the predicted water level of Lake Erie for the next five months? Are water levels increasing or decreasing?
3. A high water level is defined as 15 tens of meters or greater above the sea level and a low water level is defined as less than 15 tens of meters above the sea level. If a random sample of 100 water level readings is taken, is the count of high and low water levels independent of season (winter, spring, fall, and summer)?

# Methods and Analysis

In order to start the analysis, we will first find the average water level of Lake Erie. The average comes out to be 14.99305. Next, we would have a look at the data by converting it into a time series object and plotting it. The plot is as given below:



It can be seen from the plot that the water levels have been very erratic over the years. There seems to be a slightly upwards trend with a few dips in-between. There also seems to be a seasonal component. In order to examine these components more closely, the data can be decomposed and plotted. The plot is as shown below:



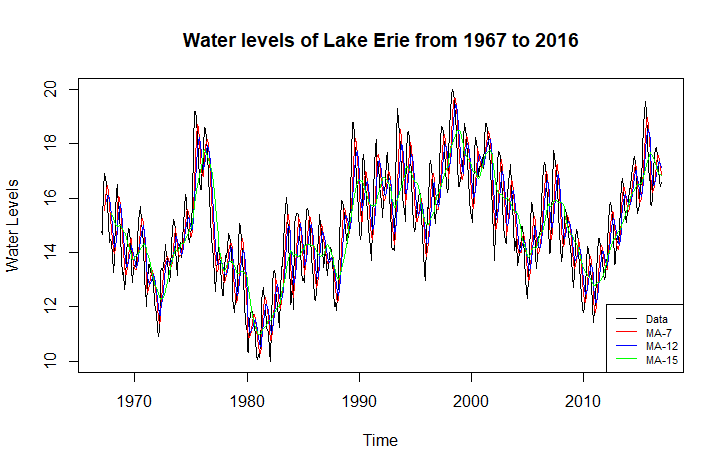
Trend: There was an upward trend near 1976 which dropped back down in about 1981. There was also a slight dip near 2010. Overall, the trend is an increasing one even though the increase is less.

Seasonality: There is a definite seasonal component in the water levels. The water levels seem to rise in summer and recede during winters.

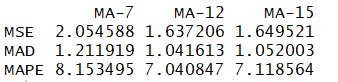
Cyclical: There does not seem to be any cyclical component in the given data.

Irregularities: There are definite irregularities in the data with a few outliers.

In order to start the initial analysis, the moving averages method can be applied to analyse the data. For this purpose, the lengths 7, 12, and 15 have been taken. We can plot all the three models in the same graph to compare it with the data.

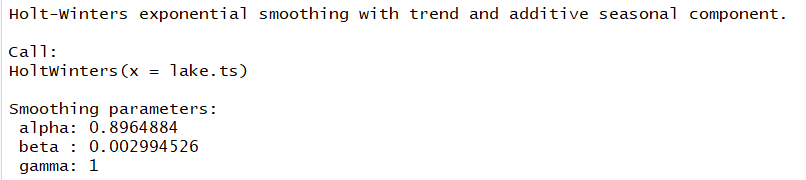


The plot shows the smoothing curves for MA-7, MA-12, and MA-15. MA-15 is the smoothest curve, while MA-12 and MA-7 capture the fluctuations in a more detailed manner which is why their curves are close to those of the original data. In order to assess the best model for the data, the errors for all the three models should be examined and the one with the least errors would be the most appropriate model.



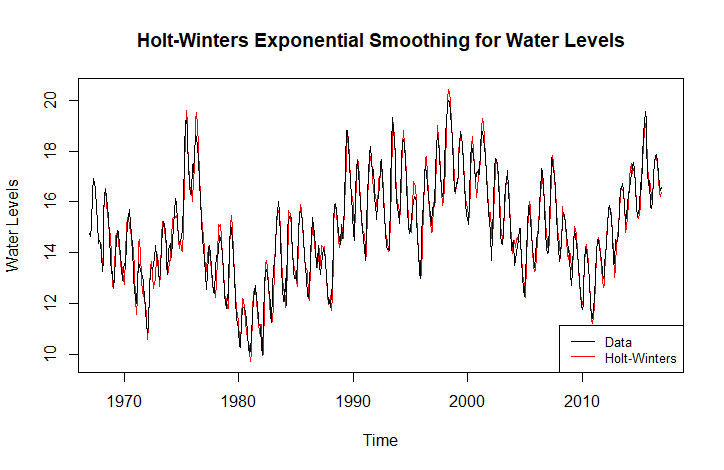
It can be seen that the MA-12 model has the least errors with a percentage error of nearly 7.04%. Thus, the best moving averages model for the given data would be MA-12.

However, since the data has both trend and seasonality components, the most appropriate model would be Holt-Winter’s Exponential Smoothing model. In order to find the Holt-Winter’s model with the best fit, we would not specify the alpha, beta, and gamma components. The results are as follows:



The optimal Holt-Winter’s model for the given data would have alpha = 0.8964884, beta = 0.002994526, and gamma = 1. Thus, for seasonality, all the weight would be placed on the most recent data whereas for the trend, maximum weight would be placed on the previous data.

The plot of the model is as shown below:



It can be seen from the plot that the model closely follows the given data and captures the trend and the seasonality.

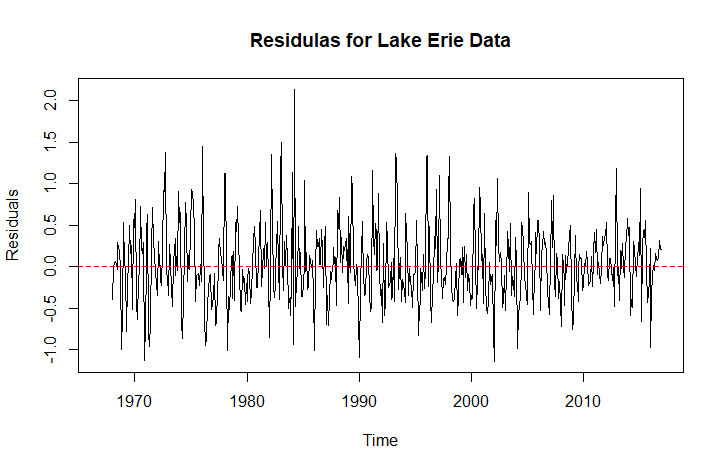
In order to check the accuracy of the model, we can look at the errors.



The errors are at a minimum with the percentage error being just 2.3%. Thus, it can be said that the model is accurate.

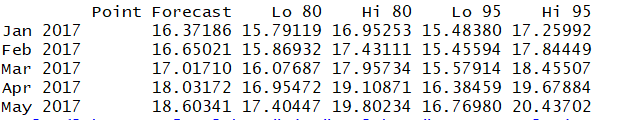
As specified above, since the data has both seasonal and trend components, the best model for forecast would be the Holt-Winter’s Exponential Smoothing Model. This can also be seen from the fact that the error for Holt-Winter’s is lesser than that for the Moving Averages model.

A plot can also be created to assess the quality of the model. This plot would be of the residuals of the given data.

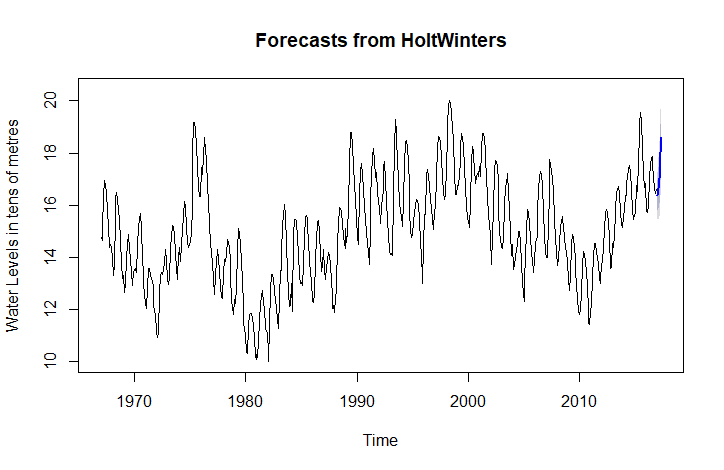


The plot shows that there are no major outliers and the plot is mostly balanced around the 0 mark. Thus, it can be said that the model is accurate as well as optimal.

Using the Holt-Winter’s model, the predictions for the next five months would be as follows:



The forecast also shows the 80% and the 95% confidence intervals as can be seen above. The plot of the forecast is shown below:



The blue line at the end of the plot shows the forecasted values.

In order to find whether the water levels (High >= 15, Low < 15) are independent of the seasons (Fall, Winter, Spring, and Summer), a chi-square test can be used. For this, we will take a random sample of 100 water levels out of the given 600 and count the number of High and Low water levels in all the seasons. Firstly, we will check the conditions for the chi-square test.

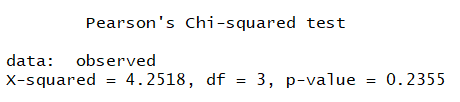
1. Counted Data Condition: Since we count the number of High and Low water levels in each season, this condition is satisfied.
2. Independence Condition: The sample is taken randomly so the readings are independent of each other.
3. Randomization: Since the sample is random, this condition is met.
4. Sample size condition: Sample size is 100 > 5. Thus, this condition is met.

Next, we will formulate the null and the alternative hypothesis.

Null Hypotheses, H0: Levels (high or low) are independent of seasons.

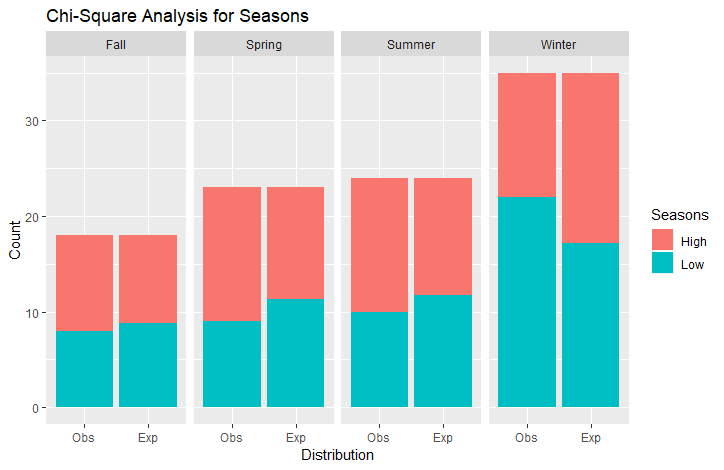
Alternative Hypotheses, HA: Levels and seasons are not independent.

Performing the chi-squared test, we get the following results:



As the p-value of 0.2355 is greater than alpha which is 0.05, we fail to reject the null hypotheses. Thus, it can be said that the seasons and the levels are not independent of each other, or, in other words, seasons and levels are dependent on each other.

The number of observed and the expected values for seasons can be visualised as in the plot below:



It can be seen from the plot that there are quite a few differences in the number of observed values for high and low water levels for every season. For fall, spring, and summer, the number of observed values for high water levels are more than the number of expected values, whereas the observed number for low values are less than expected. The trend is the exact opposite for Winter.

Overall, it can be said that the water levels are increasing.

# Results and Discussion

From the analysis, we find the answers to the questions as follows:

1. The average water level in Lake Erie is 14.99305.
2. The predicted water levels for the next five months are:

|  |  |
| --- | --- |
| Month | Predicted Water Levels |
| January 2017 | 16.37186 |
| February 2017 | 16.65021 |
| March 2017 | 17.01710 |
| April 2017 | 18.03172 |
| May 2017 | 18.60341 |

Also, the water levels are increasing.

1. Water levels and seasons are not independent of each other.