Team Control Number

12015

Problem Chosen

B

2022
HiMCM/MidMCM
Summary Sheet

Drought Model Based on Mechanism Analysis, K-Means and ARMA Time Series

With the increasing demand of human beings, more and more natural resources are being consumed, resulting in more and more serious consequences. The sudden drought in Lake Mead is one of many consequences. Therefore, to solve the problem, we construct our model to evaluate factors that could influence the volume of Lake Mead and predict the future trend of drought of Lake Mead, then we come up with a plan to recycle the wastewater as to solve this problem.

We first determine the factors that will affect the water volume of Lake Mead: inflow, outflow and loss. Then we divide the above three main factors into several smaller factors. We obtain the relationship between these factors through **Mechanism Analysis**. In order to verify the relationship between the water volume, water level and altitude, and area of Lake Mead, we used **Binary Numerical Integration** and Linear Regression methods, and the obtained R^2 is very close to 1.

Secondly, we focused on changes in elevation by visualizing the data we obtained, and summarized the overall pattern. We use the **K-Means** method to do cluster analysis to define the criteria for drought. For the prediction problem of Lake Mead, we first use **ADF test** to test the stability of the data. For the recent data model, we directly use **ARMA model** to predict the altitude change of Lake Mead in the future. For the long-term data model, we first use **Linear Regression** to extract the trend term, and then use ARMA model to make predictions. The results we get show that if we do not take any action, the water level of Lake Mead will continue to fall, which means that the government must respond to this.

Come to recycling of wastewater, we investigate some factors that might influence the amount of wastewater through **Reviewing Literature**, the factors are usually different depending on the individual or group that use water. Based on these factors, we come up with regulations and plans that the government could use in order to increase the efficiency of wastewater recycling. We also obtained a **Differential Equation** Model as to measure the impact of these regulations which shows wastewater recycling can indeed alleviate the trend of gradual reduction of available water, and the more reclaimed water recycled, the slower the attenuation of available water.

To ensure the validity and stability of the model, we conduct sensitivity analysis using **Elbow method** and **Contour Coefficient** to analyze the stability of the model. Finally we summarized the strengths and weaknesses of the model and wrote a one page news article to report the key points of this research and provide some suggestions for the government to apply.

Keywords: Lake Mead; Drought; Wastewater; K-Means; ARMA model; Contour coefficient

Contents

1 Introduction **3**

1.1 Background . . . 3

1.2 Problem Restatement . . . 4

1.3 Our Work . . . 4

2 Assumptions and Justifications **5**

3 Variables **5**

4 Modeling **6**

4.1 Lake Mead Volume Model . . . 6

4.1.1 Lake Mead Volume Factors . . . 6

4.1.2 Lake Mead Volume Calculation . . . 9

4.2 Lake Mead Water Level Analysis . . . 10

4.2.1 Lake Mead Overall Patterns . . . 10

4.2.2 Defining Drought Criteria . . . 11

4.3 Lake Mead Water Level Forecast . . . 13

4.3.1 Model Based on Recent Data . . . 13

4.3.2 Model Based on Long Term Data . . . 15

4.4 Wastewater Recycle Model . . . 18

4.4.1 Factors Influencing Wastewater . . . 18

4.4.2 Actions and Plans of Local Leaders . . . 19

4.4.3 Measure the Impact . . . 20

5 Sensitivity Analysis **22**

5.1 Elbow method . . . 22

5.2 Contour Coefficient . . . 22

6 Strengths and Weaknesses **23**

6.1 Strengths . . . 23

6.2 Weaknesses . . . 23

Daily News

News Hour

2021.11.16 Tuseday

Morning News

Lake Mead is a large artificial freshwater lake in the center of a desert. It is completed around 1935 and it provides water resources for the residents nearby. Thanks to Lake Mead, the local economy is prosperous, but the beauty seems that it will never come back.
As there are more and more residents, the need for water is becoming greater and thus the volume of Lake Mead is dropping continuously. Climate change is also responsible for the shortage of water.

To discover what factors are affecting the volume of Lake Mead, we make an analysis and identify three main factors, which are inflow, outflow and loss through evaporation. The inflow is the sum of tributaries and precipitation. Tributaries can be calculated with the intersecting surface and the speed of inflow when tributaries interact with Lake Mead. We take precipitation as a constant as in less than ten years, rainfall does not change a lot. We make deeper analyses and can obtain the actual volume as long as measurements take place.

After that, we visualize and rearrange the water level elevation of Lake Mead from 1935 to 2021, and based on the data, we use make a standard value to distinguish the drought period and the other through rating the volume of different years from high to low. We discover that the dry periods are becoming longer and more severe recently. In more detail, data shows that since 2010, Lake Mead is constantly lack of water and the volume is way smaller than before. Due to the result, we make a model to predict if Lake Mead can meet human needs in the future. We choose the time series analysis to make a reasonable prediction. In this approach, we include the trend and take the possible sudden changes into account. We find that the future volume of Lake Mead is not enough for water usage in the future.

Therefore, action needs to take place. Wastewater recycling is a solution, but the efficiency needs improvement. Wastewater recycling is a wide topic, so we turn it into smaller sections. The wastewater can be made up of six parts: industrial, farm, commercial, green infrastructure, domestic, and storm wastewater. for example, industrial wastewater is determined by the usage of electricity and water. We make judgments on factors of water recycling and give some advice to local leaders on how to make specific reactions.

In short, we conducted a comprehensive and detailed investigation and analysis of all aspects of Lake Mead. We hope that our report can contribute to mankind's tackling the drought.

1 Introduction

1.1 Background

As people's consumption of natural resources increases, more and more crises are emerging, and drought is one of them. Drought is a condition in which there is not enough freshwater to sustain human survival. It is usually caused by a decrease in the amount of precipitation and an increase in the number of people around it. Lake Mead, the largest water reservoir in the United States, is at its lowest level since it was first filled in the 1930s mainly due to the reduced amount of precipitation and increased demand from 25 million people. As a result, Lake Mead has reduced to approximately 36 percent of its full capacity. On August 16, 2021, the Bureau of Reclamation announced the first-ever water-shortage declaration on the Colorado River[1].

Water is the source of life and the most important resource for all living things to sustain life. Although water occupies 71 percent of the earth's surface, only 2.5 percent of it is freshwater that humans can drink directly. The remaining 97.5 percent is saltwater which humans can't drink directly [10] . Therefore, in order to save water, we decide to recycle wastewater that flows out of our sinks, toilets, and showers. Of course, the wastewater goes through a series of treatments before being reused. The plan for recycling water is roughly shown in Figure 1.1.

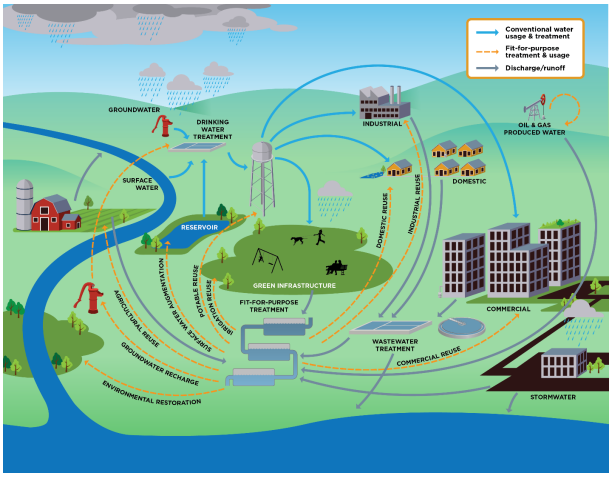

Figure 1.1: Examples of water sources and use application for recycled water (Environmental Protection Agency[2])

For data, We include the relationship of elevation, area and volume at different point which is shown in Table 1.1 below[3].

Elevation (feet)	Area of Lake (acres)	Volume of Lake (acre-feet)
1229.0	159,866	29,686,054
1219.6	152,828	28,229,730
1050.0	73,615	10,217,399
895.0	30,084	2,576,395

Table 1.1: Elevation, area and volume for Lake Mead in different position

As we can see, there are four lines in total in which different line stands for the elevation, area, and volume of the lake at a different position. Line one stands for maximum designed

water-surface elevation; line two stands for the crest of drum gates on spillway (raised); line three stands for Intake tower, upper gates; line four stands for intake tower, lower cylinder gate entrance liners.

Besides this, we also include elevation of the water at Lake Mead (in feet above sea level) at the end of each month by year and the highest and lowest water elevation at Lake Mead (in feet above sea level) by year, these data will be adopted in the following problems.

1.2 Problem Restatement

1. As the volume of Lake Mead will either gain or lose over time, we are going to decide on the factors which influence inflow, outflow, and loss in Lake Mead, then find out the relationships between these factors and find the relative impact of these factors on the volume and water level of Lake Mead.

2. Because there is always water flowing in Lake Mead, it was inevitable that the bottom of the lake would be pockmarked in the course of its natural formation, causing the shape of the lake to be irregular. Therefore we have to consider the data and information needed for finding the relationship of elevation, area, and volume.

3. A drought is a period when there is not enough freshwater to sustain local people. How can we define the drought? In this study, we will set drought criteria based on recent water levels and compare recent dry periods with earlier ones.

4. Based on the conclusion of question 3, we will build two models to predict the change of water level in the future in 2025, 2030 and 2050. The conclusions of the first model are based on recent drought data; the second model is based on data from 2005 to 2020. So we can predict the water level in the future.

5. As water levels continue to fall, recycling wastewater is also a viable solution. Then what factors should we consider for this solution? We also need to consider how the policies that local leaders will pose might affect the program.

6. We will also describe the plan and consider the method to measure the impact of implementing this plan.

7. A one-page non-technical news report will also be included in this article and we will make suggestions based on the result we obtain.

1.3 Our Work

In this research, we first regard the water volume of Lake Mead as a function of inflow, outflow and loss, and consider the factors affecting the inflow, outflow and loss of Lake Mead in detail. In addition, we established a calculation formula for the water volume of Lake Mead through mechanism analysis and linear regression, which can verify whether the previous data is correct. For the elevation of the water level of Lake Mead, we used K-Means cluster analysis to define drought and summarized the overall pattern of the water level of Lake Mead. We use the ARMA model to establish two models to predict future drought trends. Our results indicate that if the government does not take any action, the water level will continue to fall. Therefore, we evaluated the source of wastewater and formulated some regulations and plans for the government to achieve more effective wastewater recycling. We also establish differential equations to quantify the impact of our planned implementation. Finally, we wrote a news article to report the key points and recommendations of this issue.

2 Assumptions and Justifications

- **Assumption 1:** While measuring, we assume that the water surface is flat without fluctuation for the ease of calculation. The water surface is unpredictable and therefore brings great difficulties to our calculation. In addition, the difference in height has little change on the result due to its scale is low.

- **Assumption 2:** We assume that the volume of Lake Mead is only effected by the amount of inflow, outflow and loss for the ease of calculation. Based on our research, the volume is mainly influenced by inflow, outflow and loss. There are some other factors such as water accidentally flowing into Lake Mead, which is negligible.

- **Assumption 3:** We assume that the amount of inflow is only affected by the water flow in from the four tributaries and precipitation, inflow from other sources are negligible. Although there may be other factors that affect the water volume of Lake Mead, these three are the most important factors, so other factors can be ignored.

- **Assumption 4:** We assume that the climate will hardly change during a short term so as to show the length of t will affect the amount of precipitation. In fact, climate change in the short term is a very rare event, so annual precipitation can be considered a fixed figure.

- **Assumption 5:** We assume that during time t, the water in Lake Mead will not seep into the ground. In fact, the water in the lake does seep into the ground, but once the water content of the soil reaches saturation, the water seeping into the ground can be ignored.

3 Variables

Table 3.1: Variables Table

Variables	Description
t	time of the investigation period
V	volume of Lake Mead
S	the surface area of cross section of tributary
k	the number of clusters
x, y	the coordinate of the centroid
h_0	distance from sea level to the lowest point of the lake
h_i	distance from sea level to the bottom of the lake
h_s	distance measured by sonar from the surface of the lake
E	elevation of the lake
A	surface area of Lake Mead
v	speed of the flow
e	evaporation rate
I	inflow
O	outflow
N	number of sluice gate
W	wastewater
W_A	available water
W_R	wastewater recovery

4 Modeling

In this section, we contain the way to measure Lake Mead's volume and the model that predicts the future trend of the drought. The plan of recycling the wastewater is also listed and the impact of it is also shown in this section.

4.1 Lake Mead Volume Model

4.1.1 Lake Mead Volume Factors

From what we have assumed, the volume of the lake is only influenced by the amount of water inflow, outflow, and loss while the initial volume depends on the choice of researchers. Therefore the factors affecting are shown in Figure 4.1.

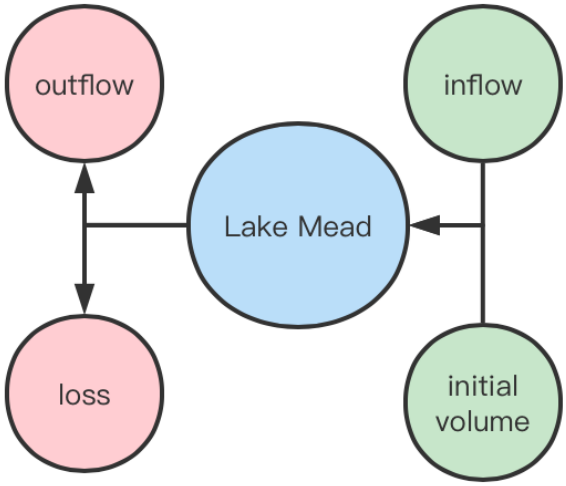

Figure 4.1: Overview of factors affecting volume

Therefore equation (4.1) is the relationship of Lake Mead's volume and its factors.

$$V = V_{init} + Inflow - Outflow - Loss \tag{4.1}$$

Among these factors, inflow is determined by the water brought by the four tributaries and the amount of precipitation. (We have explained it in our assumptions) To calculate the amount of water from four tributaries, we multiply the cross section of a tributary where it meets Lake Mead by the flow rate and the time period of measurement, then plus the amount from each tributary together, the equation is shown below.

$$I_{tri} = t \sum_{j=1}^{4} S_{tri}^{j} v_{tri}^{j} \tag{4.2}$$

The cross-section must locate at the point where the tributary meets Lake Mead since the bottom of the tributary is not smooth and can be in different shapes at a different point. The joining process is shown in Figure 4.2.

To calculate the area of cross section, since it is irregular in shape, we need to integral the cross section to cut them into small pieces just like shown in Figure 4.2 , and therefore by supposing the surface of the tributary is flat, the area of cross section is shown in equation 4.3, where x and y can be measured by sonar and $f(x)$ is the final result of data put together.

$$S_{tri} = x_1 y_1 - \int_0^{x_1} f(x)\, dx \tag{4.3}$$

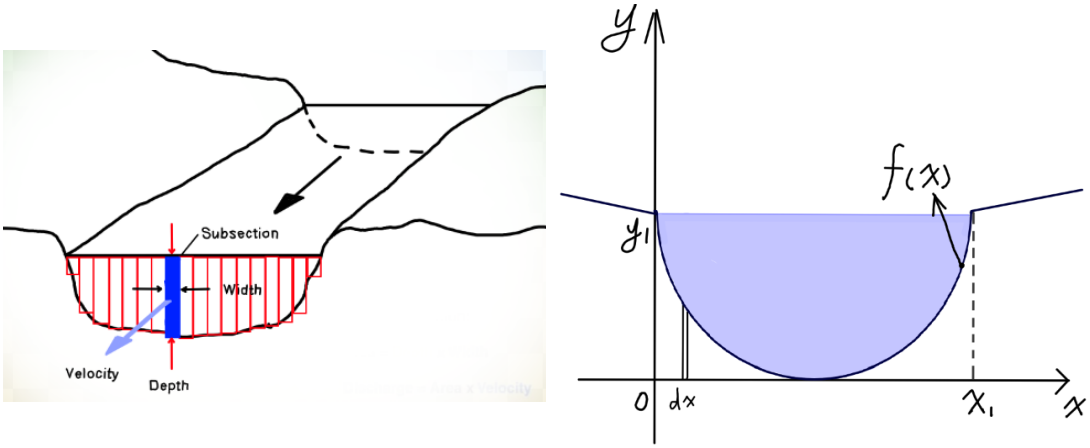

Figure 4.2: The figure on the left shows the cross-section of the river confluence and the flow velocity. The figure on the right shows the calculation of the cross section.

For the ease of calculation, here we suppose that the bottom of the lake forms a second order function like shown in Figure 4.2.

As for a second order function, we have $f(x) = a(x-b)^2 + c$ to be the basic function, after integration if we put it into equation 4.3, the result is shown below.

$$S_{tri} = x_1 y_1 - \frac{a(x_1 - b)^3 + ab^3}{3} - cx_1 \tag{4.4}$$

Note that here v stands for the average speed of flow, the average here either means the flow speed per second, it also means the average velocity per square unit of cross sectional area. Therefore the speed of the tributary can be calculated by the equation below.

$$v_{tri} = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_1'} v_{ij}}{n_1 n_1'} \tag{4.5}$$

Therefore the steps for calculating the amount of inflow from tributaries are shown above, for precipitation, the only thing that can be controlled is the measured time t. t could be any period from one week to even a year. As the length of t will determine the amount of precipitation. If we have t equals to one month, the precipitation can't be determined as a constant since there are rainy and dry seasons. In contrast, if we have t which stands for one year, then it can be a constant, in this essay we usually see it as one year, therefore we have it as a constant.

Now come to outflow, as we have assumed that the amount of outflow is only determined by the water released and the water consumed directly from the lake, we will explain the way to

calculate it.

$$Outflow = O_{rel} + O_{con} \tag{4.6}$$

The release of the water is mainly through the dam, as the dam takes on the largest percentage of released water, the other ways of releasing will be considered as others.

$$O_{rel} = O_{dam} + O_{oth} \tag{4.7}$$

Where $O_{release}$ is the outflow through releasing, O_{dam} is the outflow through the dam and O_{oth} represents the outflow in other ways.

The shape of the dam is shown in Figure 4.3 with sluice gates on it, the dam release water by opening the sluice gates. Therefore we will use the same calculation steps as the inflow amount from the tributaries. Here we suppose the dam has n sluice gates and get the product of the speed of the flow when it passes through the gates and the area of the gate, the steps are the same for others. Here t_1 means the period for the dam to release wate and t_2 stands for the time period for others.

$$O_{rel} = NS_{dam}v_{dam}t_1 + S_{oth}v_{oth}t_2 \tag{4.8}$$

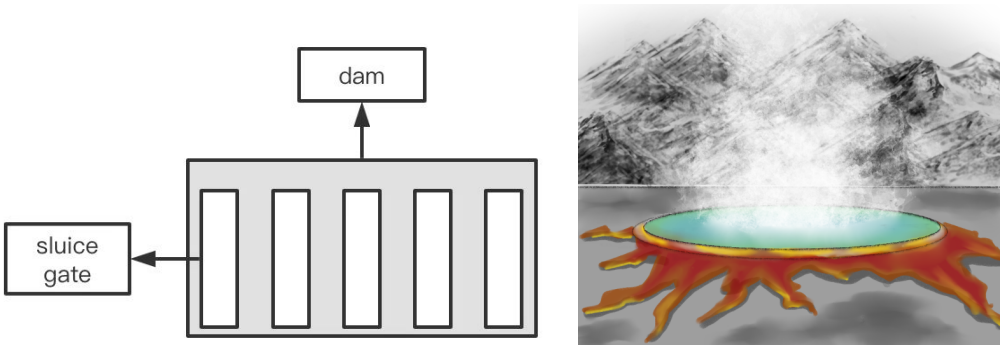

Figure 4.3: The left figure shows Lake Mead releases water through the dam. The right figure shows lake evaporation

The loss of the water, as we assumed, will only be affected by the amount of evaporation.

As evaporation will be mainly affected by the surface area of the lake, the temperature and the time, we could know that,

$$Loss = eAt \tag{4.9}$$

where e stands for the evaporation rate and we assume during t, the surface area of the lake will have a negligible change. There is a more accurate equation to calculate the loss which is the Penman formula, a formula used specifically to calculate the water evaporation. Due to the complexity of this formula, if we would like to calculate the loss accurately, we can use the simplified version instead which is shown below[4].

$$e = \frac{\frac{700T_m}{100 - La} + 15(T - T_d)}{80 - T} \tag{4.10}$$

In this formula, $T_m = T + 0.006h$, h is the altitude (meters), T stands for the average temperature and La stands for the latitude, which means e has nothing to do with longitude. One more thing to notice is that $T - T_d = 0.0023h + 0.37T + 0.53R + 0.35Rann - 10.9\,°\text{C}$. Where R is the average daily temperature, R_{ann} is the difference between the average temperature of the hottest and coldest months.

Therefore above shows the relationship between factors and the volume of Lake Mead.

4.1.2 Lake Mead Volume Calculation

As we can infer from Table 1.1, each set was measured at different location, so by combining these four data sets, we can get a rough model of Lake Mead as follows.

From Table 1.1, as elevation stands for the distance from sea level to the surface of the lake, therefore if the lake locates below the sea level, the product of elevation times the area of the lake shall be smaller than the volume of the lake. However, if we calculate the data in the table, we will find that the result appears to be larger than the actual volume of the lake, which we can infer that the location of the lake is higher than the sea level.

Therefore we will first consider the lake Mead model as a cylinder, if we cut down the cylinder of height h_0, distance from sea level to the bottom of the lake, the remaining cylinder's height will be $E - h_0$. Therefore by removing the volume of Lake Mead, we get a dual integral in terms of A and h shown below, where A stands for the bottom area of small cuboid.

$$Volume = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f(x, y)\, dxdy = Ah - \sum_{i=1}^{N} A_i h_i \tag{4.11}$$

By breaking up main variable h, we can get the processed equation (4.12) below,

$$Volume = A(E - h_0) - \sum_{i=1}^{N} A_i(E - h_0 - h_s) \tag{4.12}$$

For this equation, we first calculate the volume of the entire cylinder using $V_{cylinder} = Ah$,

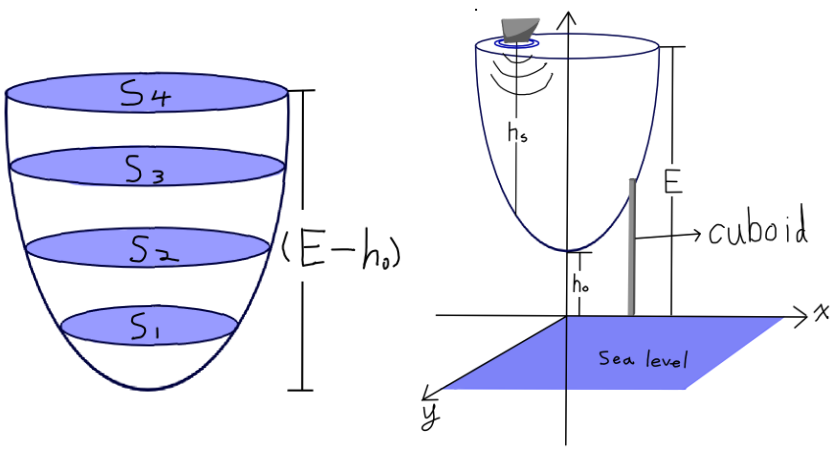

Figure 4.4: The figure on the left is a plane corresponding to different altitudes in the lake, corresponding to the altitude of the Table 1.1. The figure on the right is a schematic diagram of the numerical integration of the differential element method to obtain the lake volume.

where the height of the cylinder is E elevation minus h_0.The second term uses E minus h_0 minus h_s which represents the height from the bottom of the cylinder after removal to the bottom of the lake. As the equation contains all the factors we need, this will be the relationship between the three factors.

Now we look for the relationship between the volume, water level elevation, and area of Lake Mead from a data-driven perspective.

First, we used the volume of Lake Mead as the dependent variable and the water level elevation and area as the independent variables, and performed two linear regressions. The results of linear regression are shown in the first two graphs in the figure below. We can see that the data points appear approximately on a straight line, which means that the linear relationship between them is very strong. This inspired us to use the water level elevation multiplied by the area $E \times A$ as a common variable to do linear regression. The result of linear regression has become better, R^2 is close to 1.

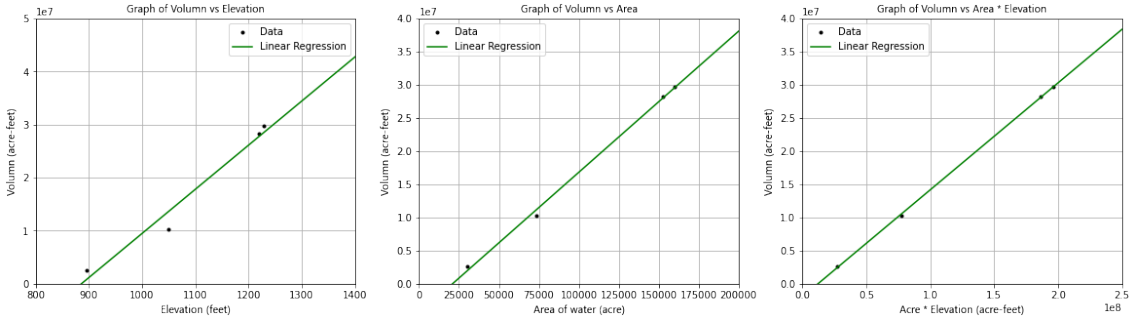

Figure 4.5: The result of linear regression.

4.2 Lake Mead Water Level Analysis

4.2.1 Lake Mead Overall Patterns

Lake mead's water level changes over time. However, it has some overall trend and it also follows a lot of patterns.

We visualize two data files to get the overall pattern of the water level of Lake Mead. It looks like below:

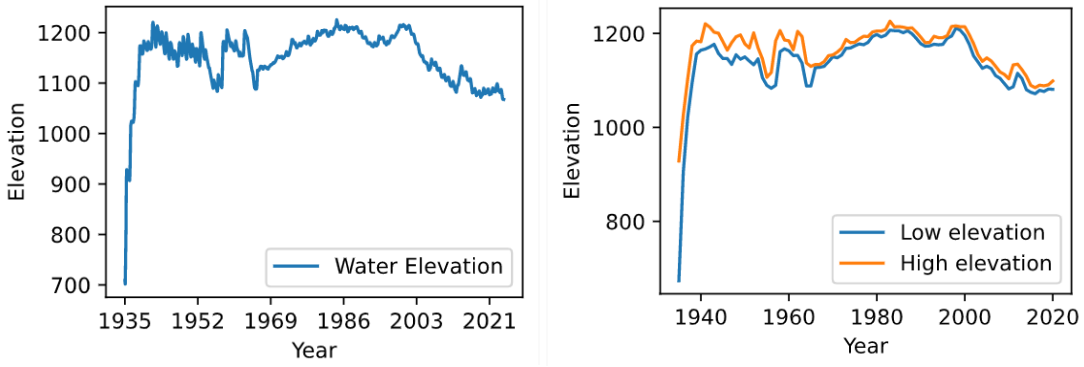

Figure 4.6: The figure on the left is the water level information of Lake Mead from 1935 to 2021 at the end of each month. The figure on the right is the highest and lowest water level information from 1935 to 2020.

Given this figure, we could easily summarize the following patterns:

1. From the plot, excluding some earlier data points, it is very clear that the water level varies a lot over time in the earlier period and varies less in the later time.

2. Also, starting from 2010, lake Mead constantly shows a lack of water. The water level was way lower than before.

3. From 1950 to 1970, the water level of Lake Mead fluctuated greatly, and there was a pattern of drought for a period of time.

4. The changes in the highest and lowest water levels of Lake Mead show similar trends. During the period from 1940 to 1960, the highest water level was significantly higher than the lowest water level, while at other times, the highest water level was similar to the lowest water level.

4.2.2 Defining Drought Criteria

This section contains the method of the cluster analysis to decide whether the water level should be considered as drought. Usually, the data will be distributed into two clusters if it has a clear drought period. Unfortunately, this data is pretty cohesive. However, we still choose to use k-means because it is one very effective way to do cluster analysis, especially in this case where we need to divide up the data into two various cases and define the drought manually.

We utilize the method of the K-means and decide to divide the data into k clusters. In this case, we suppose k equals 2, which are droughts and non-droughts. Our model firstly chooses k random data points as centroid (x, y). After that, the model chooses the center position of every point and makes it the new centroids through the following equation.

$$a_j = \frac{1}{|c_i|} \sum_{d \in c_i} d$$

However, some other points might find themselves closer to the other centroids after the above steps, which divided themselves into the other cluster. This process keeps repeating itself and finally reaches a relatively optimized answer. The iteration won't stop until it repeats 300 times.

Provided with the data in the attachment, we plot a scatter plot to show the distribution of the water elevation.

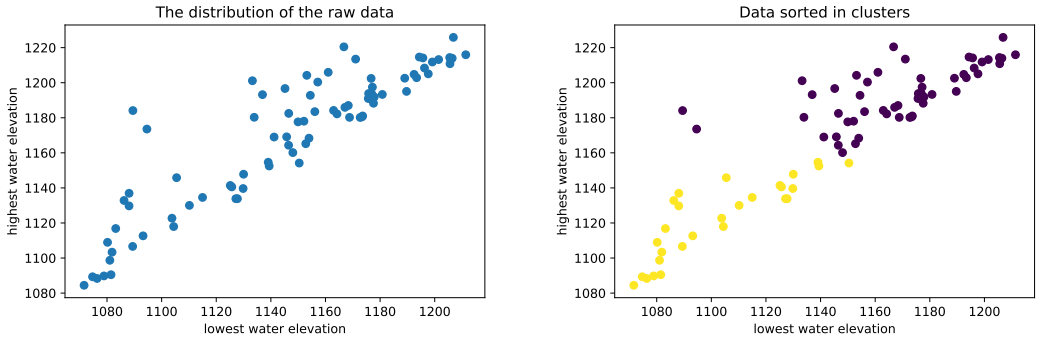

Figure 4.7: The result of the original data scatter plot and K-means clustering.

After that, we use the method described above to justify whether the given data shows a pattern of drought or not. We define the drought as water elevation lower than the mean value of the high elevation and low elevation of the lower centroid, which is the center point of the yellow cluster. The center point (x, y) for the clusters is $(1094.61, 1193.73)$, so the water elevation lower than 1144.17 is considered as drought, which is where the drought line in Figure 4.8. And we also draw the flow chart of the K-means algorithm.

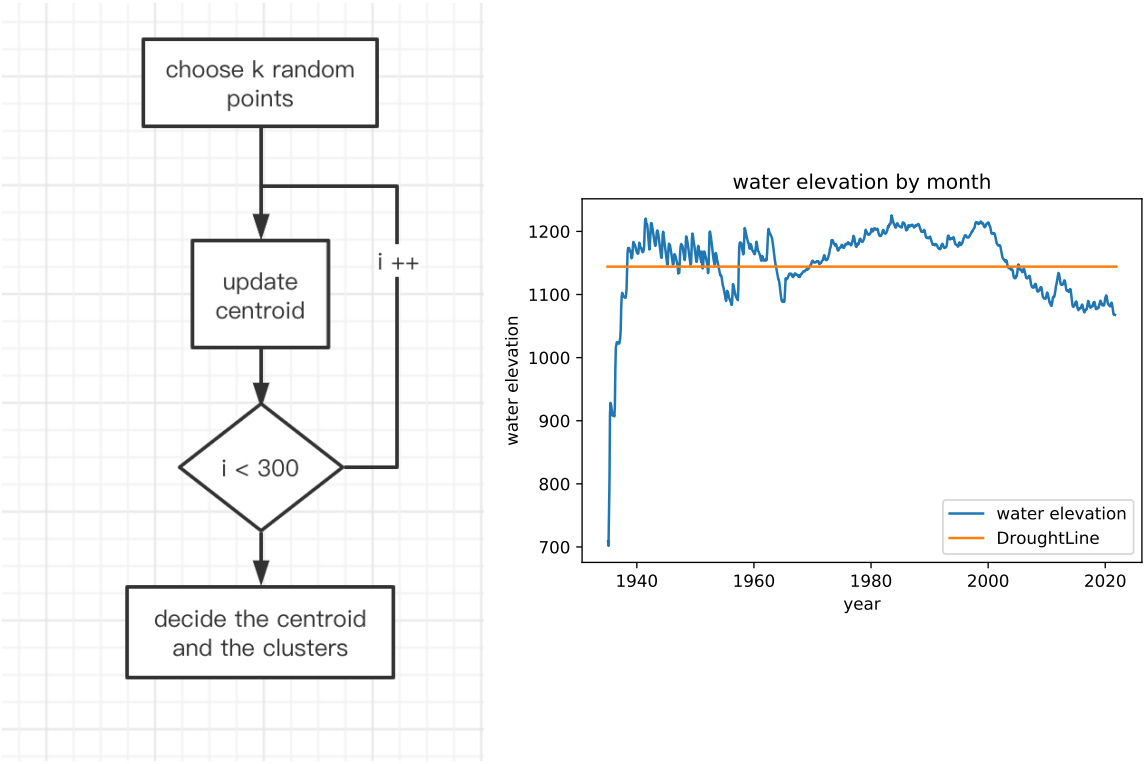

Figure 4.8: The algorithm flow chart of K-means clustering and the display of drought standard line.

Given the criteria of the drought, it is easy to find out all the drought period. It is showed in the following table:

Start period	End period
1935.2	1938.4
1946.11	1947.3
1951.1	1951.3
1951.12	1952.3
1953.11	1957.4
1963.9	1969.6
2003.4	2005.1
2005.3	2021.9

We are ignoring the first drought period in 1935 because this is considered as the period when the dam is first built and the water elevation is still raising. The mean and median water level over all the drought period has a distribution like this:

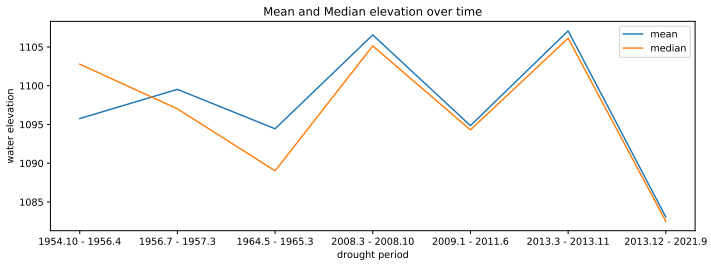

Figure 4.9: Mean and Median water level during the drought periods

From this figure, we can see that the drought levels of the several drought periods in Lake Mead's history are different, especially the average and median of the recent drought periods are lower, which means that the recent drought has been more serious.

4.3 Lake Mead Water Level Forecast

One of the very effective ways to predict future events is through the time series model. The whole idea of the time series model is to recognize the pattern and predict the phenomenons or events that would happen in the time series. In this case, the pattern we want to find is the relationship of drought in terms of time and predict the future trend of the drought. We decided to use the ARMA model in this particular context. ARMA model is made up with the AR model and the MA model, which stands for auto regressive model and the moving average model.

The first thing we are doing here is to check whether the data is "stable" or not. The data seems "stable" through simple observation, but we choose to use the ADF test to make sure.

ADF test is a test of unit root test to check the stationarity. The unit root is an important factor in the time series model. The reason we are doing this is a lot of models, including the ARMA model we are using, require the data to be stationary.

ADF test is about examine if the data set contains a unit root. Consider a If the data set is stationary, it would not contain a unit root. The p value in the ADF test the important factor in deciding whether the data is stationary enough. The p value in our model is 0.00464, which is under 0.01. This p value represent a stationary enough series and could use ARMA to predict the future trend.

4.3.1 Model Based on Recent Data

The water level data from 2015 to 2021 is shown in the figure below:

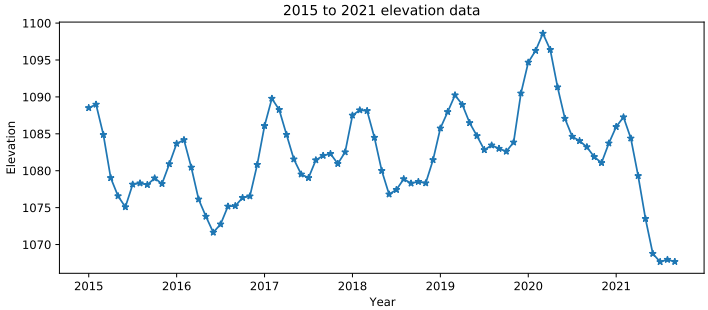

Figure 4.10: 2015 to 2021 elevation data

The ARMA model can be expressed in the following expression:

$$X_t = c + \epsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{4.13}$$

The third term in the equation is known as the Auto Regressive model and the forth term is called the Moving Average model. Those equations combines and make up the ARMA model.

Before establishing the actual model, we need to decide the order of the ARMA. The method of getting those two values is through the Auto correlation function(ACF) and partial correlation function(PACF). Those two function calculates the correlation of the past data to current data and apply those data to future context. To be more specific, ACF is the lag l between the series value that is l intervals apart. PACF is almost the same except that it also account for the value between the lag l. We plotted those two functions and it looks like this:

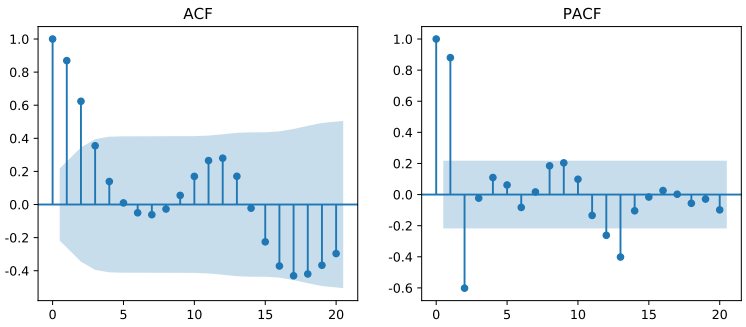

Figure 4.11: PACF and ACF

From the Figure 4.11, we could tell that for $ARMA(p, q)$ the order $p = 3, q = 3$. Thus, we choose to build $ARMA(3, 3)$.

The solution results of the model are as follows:

ARMA Model Results

Dep. Variable:	y	No. Observations:	81
Model:	ARMA(3, 3)	Log Likelihood	-160.988
Method:	css-mle	S.D. of innovations	1.701
Date:	Sat, 13 Nov 2021	AIC	337.976
Time:	15:32:27	BIC	357.132
Sample:	0	HQIC	345.661

	coef	std err	z	P>|z|	[0.025	0.975]
const	1081.1387	2.376	455.005	0.000	1076.482	1085.796
ar.L1.y	1.6964	0.167	10.149	0.000	1.369	2.024
ar.L2.y	-1.4980	0.241	-6.209	0.000	-1.971	-1.025
ar.L3.y	0.6425	0.143	4.506	0.000	0.363	0.922
ma.L1.y	-0.0324	0.179	-0.181	0.856	-0.383	0.318
ma.L2.y	0.6479	0.092	7.038	0.000	0.467	0.828
ma.L3.y	0.4825	0.142	3.389	0.001	0.203	0.762

Figure 4.12: ARMA(3, 3) model results.

The prediction effect of the model on the known data points is as follows:

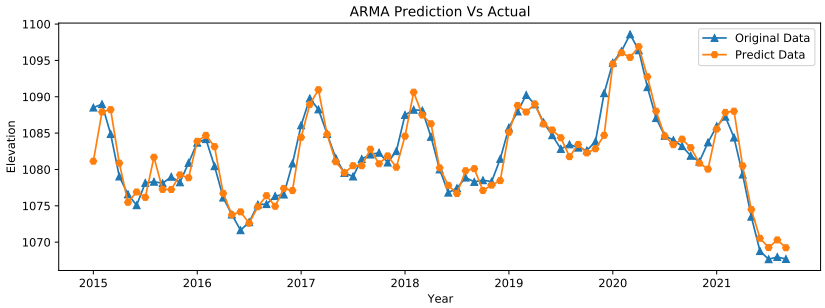

Figure 4.13: ARMA Prediction vs Actual

From the figure, we can see that the data that was predicted fits well into the actual data in the real life case, which means that it would possibly work well in the future cases as well if the trend remains.

Based on this model, we could calculate the water elevation in 2025, 2030 and 2050 to be $1081.13, 1081.14$ and 1081.14 respectively. Those three values are very close to the recent values that was around 1080. This is because we only use the recent data that is stationary.

4.3.2 Model Based on Long Term Data

Having done the Model with the recent data, we plan to use the data from a longer term to predict a future trend.

As usual, the first step is to visualize the data from the year 2005 to 2021, and the figure is showed below.

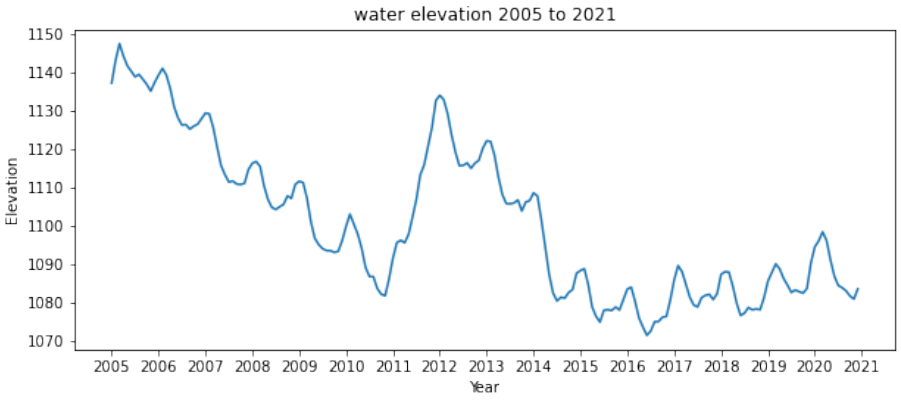

Figure 4.14: water elevation from 2005 to 2021

As mentioned above, the prerequisite for using ARMA is that the time series data is stationary. However, it is obvious that the data shows a pattern of decreasing. Thus, we decided to extract the trend of decrease and use the ARMA model.

We use the linear regression to extract the trend. Regressions likes polynomial and exponential might fits more into the data set. However, for long terms predictions like 30 years later,

those method definitely works poorly and the linear regression should have been better option. This is why we are using the linear regression.

$$E_{trend} = a_1 Y_M + b_1 \tag{4.14}$$

We could get a result of:

$$a_1 = -0.2798, \quad b_1 = 1127.66 \tag{4.15}$$

And the result can be visualized as below:

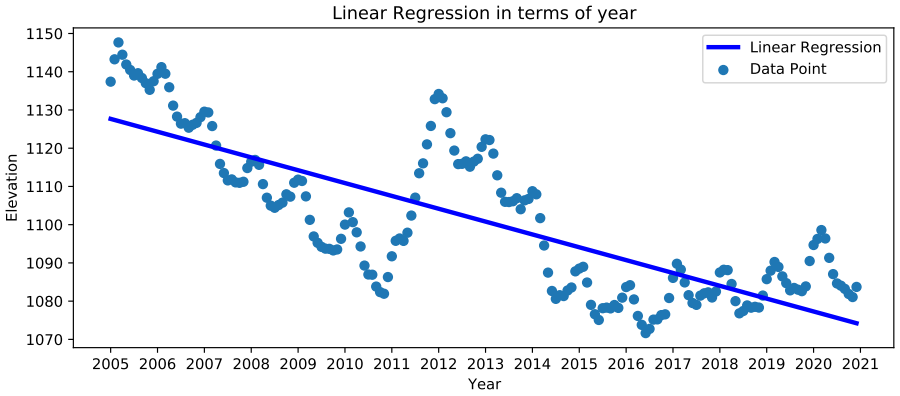

Figure 4.15: Result of Linear Regression

We define E_R as the interference items and it could be calculated by

$$E_R = E - E_{trend} \tag{4.16}$$

For the new E_R, we could use them to plot the figure of E_R, which is like below:

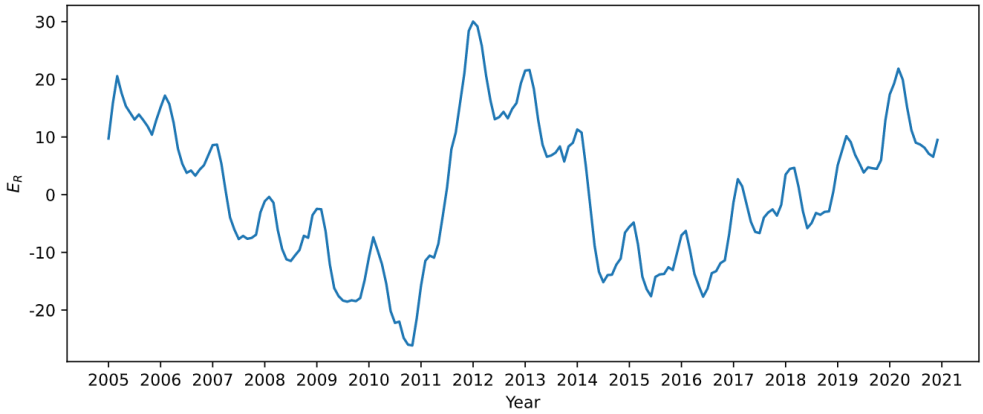

Figure 4.16: E_R as year goes

It can be seen that after subtracting the trend item, E_R has no downward trend, showing a relatively stable characteristic. Next, we follow the previous approach and use the ARMA model to predict E_R.

Just like what we did in the model with a short term data, we are going to find the ACF and the PACF function. Those two functions in this model looks like below:

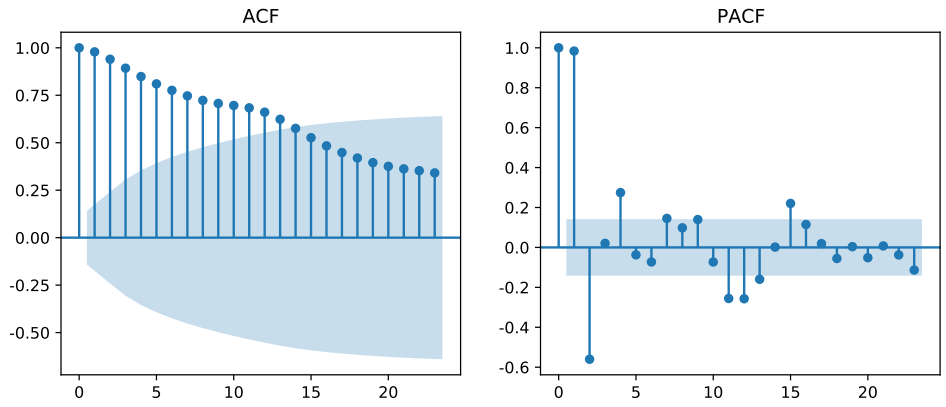

Figure 4.17: Trend as year goes

We could make sure that this model is an ARMA(3,0). Since $q = 0$, it is also the AR(3) model. From Figure 4.19 we can see that the current result provided by the ARMA model is also very close to the actual value.

The final result is the sum of the value from the ARMA modeland the value from the linear regression model.

$$E_{pred} = E_{trend}^{pred} + E_{R}^{pred} \tag{4.17}$$

which means the water level for 2025, 2030 and 2050 are $1059.7279, 1042.9088$ and 975.7493. It is clear that the water level is decreasing and it would be less than 1000 feet by the year 2050, which is why we need new policies to save the lake from drought.

ARMA Model Results

Dep. Variable:	y	No. Observations:	192
Model:	ARMA(3, 0)	Log Likelihood	-411.258
Method:	css-mle	S.D. of innovations	2.038
Date:	Sat, 13 Nov 2021	AIC	832.516
Time:	16:55:56	BIC	848.803
Sample:	0	HQIC	839.112

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.7286	3.760	0.194	0.846	-6.641	8.099
ar.L1.y	1.7881	0.071	25.200	0.000	1.649	1.927
ar.L2.y	-1.0339	0.128	-8.071	0.000	-1.285	-0.783
ar.L3.y	0.2083	0.072	2.902	0.004	0.068	0.349

Figure 4.18: ARMA(3, 0) model results.

The prediction effect of ARMA(3, 0) on known data is as follows:

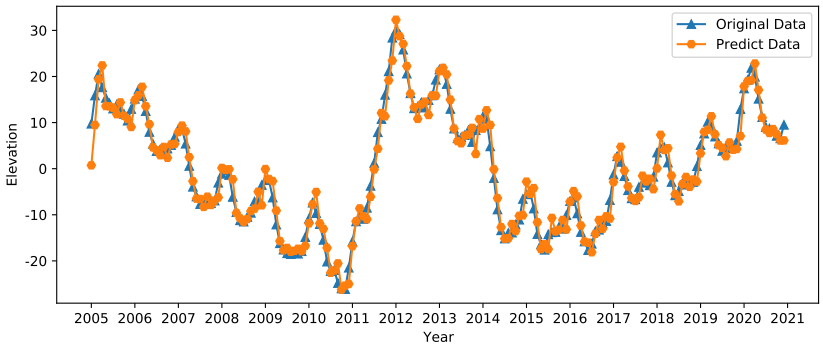

Figure 4.19: Predict vs actual

4.4 Wastewater Recycle Model

Based on the model before, Lake Mead is in the risk of water shortage. As a result, the recycling of wastewater is necessary to make up for the lack of water from Lake Mead. In this section, we will identify the factors that affect water recycling and give advice for local leaders on how to overcome the problem.

4.4.1 Factors Influencing Wastewater

The source of wastewater is mainly made up of six aspects, which is industrial, farm, commercial, green infrastructure, domestic, and storm wastewater. Therefore, the wastewater can be demonstrated using the equation below.

$$W = W_i + W_f + W_c + W_g + W_d + W_s \tag{4.18}$$

Where W stands for the wastewater, W_i means industrial wastewater, W_f stands for farm wastewater, W_c is commercial wastewater, W_g represents the wastewater from green infrastructure, W_d is a representation for domestic wastewater and W_s is the wastewater from storm water.

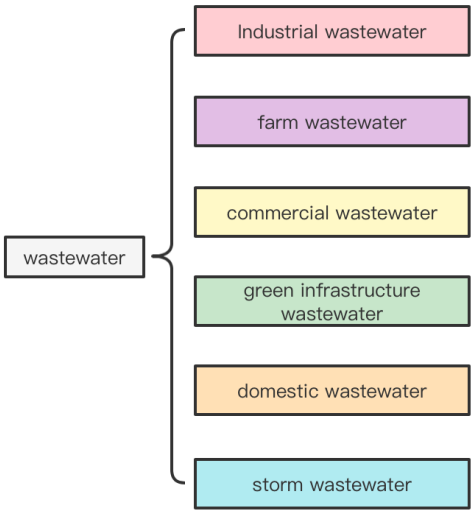

Figure 4.20: Factors of wastewater

These factors can be divided into smaller aspects as shown below.

- **Industrial wastewater** Industrial wastewater is the wastewater released by industries. In general, any manufacturing company that lets out industrial wastewater will be counted as a source of wastewater. Here we have to differentiate between sewage and wastewater, sewage means dirty water which can be cleaned by simple steps while wastewater mainly stands for water polluted by chemical reagents which have to be cleaned by more complex steps[5]. Here in this chapter, to simplify the problem, we will only distinguish between wastewater and sewage in the industrial part. As sewage can be counted as one kind of wastewater, we will use wastewater as a joint name for other sorts shown in Figure 4.20.

- **Farm wastewater** The wastewater from the farm consists of irrigation and feeding water. The wastewater produced by irrigation is formed when farmers spread chemical fertilizer into the field[6]. The excess fertilizer will mix with water, making the water unavailable. The feeding wastewater is water mixed with excrement from animals. As a result, the amount of fertilizer used will influence the volume of wastewater.

- **Commercial wastewater** Commercial wastewater is the wastewater released during commercial operations. It contains four aspects: catering, beautifying, construction wastewater, and the size of the company. When an industry grows larger, more bi-sections and needs for beautifying and construction are created which the wastewater will rise proportionally with it.

- **Green infrastructure wastewater** Green infrastructure wastewater can be divided into wastewater created by building facilities and the cultivation of green belts. It is the excess unavailable water after cultivating green belts in cities. It is decided by the size of the green belt and the type of plant planted. For public facilities, wastewater will be created by either cleaning or building.

- **Domestic wastewater** Domestic wastewater comes mainly from two sources, which are kitchen and bathroom wastewater[7]. The usage of water is directly proportional to the amount of wastewater. The residents' awareness of water recycling also plays a crucial role as if the residents are well educated about saving the water, they will realize the importance of water and will save as much water as they can.

- **Storm wastewater** Storm wastewater is water from storms that can not be utilized. Water from storms is unpredictable as it may flow through areas with pollutants and thus carry them towards clean sources of water. The only factor that can control the waste storm water is the seasons and the geographical location.

4.4.2 Actions and Plans of Local Leaders

For the factors above, the local leaders can take specific action to encourage water recycling.

1. Focusing on industrial wastewater, local leaders can raise the price for industrial water. A higher price will force the administrators of industries to cut down on water usage and thus release less wastewater. However such policy may cause a consequence that the industries will release wastewater containing lots of harmful substances that can not be recycled. Therefore the government shall also set regulations on the polluted level of the wastewater. For example, if the government sets three levels: A, B, and C, where A

stands for the most polluted wastewater and C stands for the least polluted wastewater, each level shall be charged with a different price as the way to deal with it will be at different complexity.

2. For farm wastewater, the recycling of wastewater can be improved by cutting down the use of chemical fertilizer and promote the development of dealing method inside the farm to promote the efficiency of water recycling.

3. For commercial wastewater, the wastewater is mainly influenced by the size of a company. Thus, local leaders can limit the size of companies. For example, companies can be required to take limited space. At the same time, more environment-friendly materials can be used during construction. Such practices can also cut down the release of wastewater.

4. For green infrastructure wastewater, plants need water to live and thus let out wastewater. We can introduce plants that need less water so that less wastewater can be released.

5. The key factors that influence domestic wastewater are water usage and the awareness of people. Therefore, policies can be posed to raise water prices and sponsor schools to publicize the importance of saving water resources.

6. The most important of all, the government should pay more attention on the researchers and give subsidy to the university in order to raise the average salary of academic staffs. Due to the living pressure from the society, more people will consider to work in companies instead of doing research in the university simply because the salary is higher. Therefore if government would like to solve the problem effectively, a support towards academic staff is needed and more advanced technology may be developed as to solve the problem of water.

7. As storm water is difficult to predict, the only thing government can do is to control the environment and set regulations on places that have large precipitation amount to make sure the safety of residents there.

4.4.3 Measure the Impact

As we have given our advice above, we will then predict the possible results if such policies are set.

- Industrial wastewater is a great issue, since it will probably be the most difficult one to recycle due to its harm. The classification of industrial wastewater will force the administrators to pre-process to pay less.

- Limiting the use of chemical fertilizer makes the wastewater from farm easier to process, thus making water recycling more efficient. In addition, the cut on the use of fertilizer will make food healthier.

- The introduction of plants which need less water is the solution to reduce green infrastructure wastewater. But we cannot predict what influence the introduced plants will have on the overall appearance of a human habitat.

- Attention on researchers and universities is a long-term process. In other words, the effects are not significant in a short period. It may take a few generations to find improvement in

technology. But the benefits are not limited only in water recycling. It can improve the local awareness of science.

Now we quantify the impact of wastewater recycling in a quantitative way.

The available water is recorded as W_A and $W_A = 1$ at the initial moment. Under the normal development of society, the available water will become less and less. As the available water becomes less and less, people's awareness of saving water will gradually increase, so the rate of attenuation of available water will also slow down accordingly. The above process can be described by an ordinary differential equation:

$$\begin{cases} \dfrac{dW_A}{dt} = -\alpha W_A \\ W_A(0) = 1 \end{cases} \tag{4.19}$$

The solution of this equation is in exponential form: $W_A(t) = e^{-\alpha t}$, the image of this function is shown in Figure 4.21.

Now we have introduced a waste water recycling plan. According to our plan, the used waste water will be treated and converted into usable water. The amount of waste water converted into usable water per unit time is W_R, then the above differential equation will be obtained A modified form:

$$\begin{cases} \dfrac{dW_A}{dt} = -\alpha W_A + W_R \\ W_A(0) = 1 \end{cases} \tag{4.20}$$

The solution of this equation is as follows:

$$W_A(t) = \frac{1}{a}(b + (a - b)e^{-\alpha t}) \tag{4.21}$$

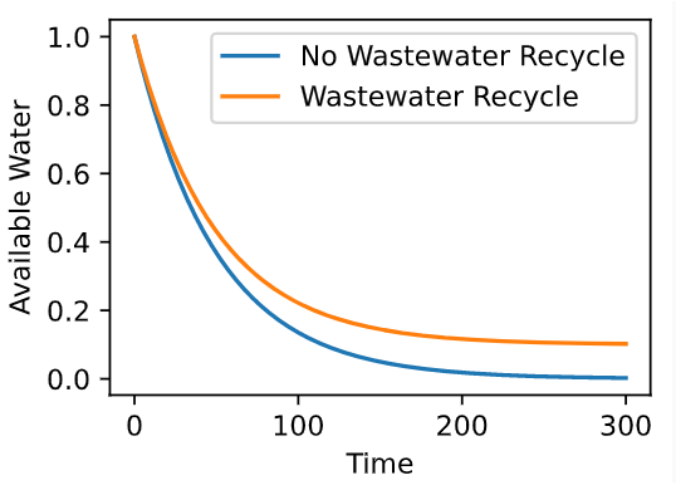

Figure 4.21: No wastewater recycle vs wastewater recycle

It can be seen that according to our above model, wastewater recycling can indeed alleviate the trend of gradual reduction of available water, and the more reclaimed water recycled, the slower the attenuation of available water.

5 Sensitivity Analysis

In k-means cluster analysis, we suppose the k value equals to 2. However, we don't know if it is actually the best k and how the model would change if a different k value is given. We apply the elbow method and the Contour Coefficient to decide the best k value. We are going to talk about the elbow method first.

5.1 Elbow method

Elbow method[11] was obvious. However, it needs human power to see where the turning point is. The way the elbow method work is to calculate the average distance of all the points to the centroids in different "k"s. It should have been a shape of an elbow as it would decrease very quickly at first and decrease less after, just look like a bent elbow.

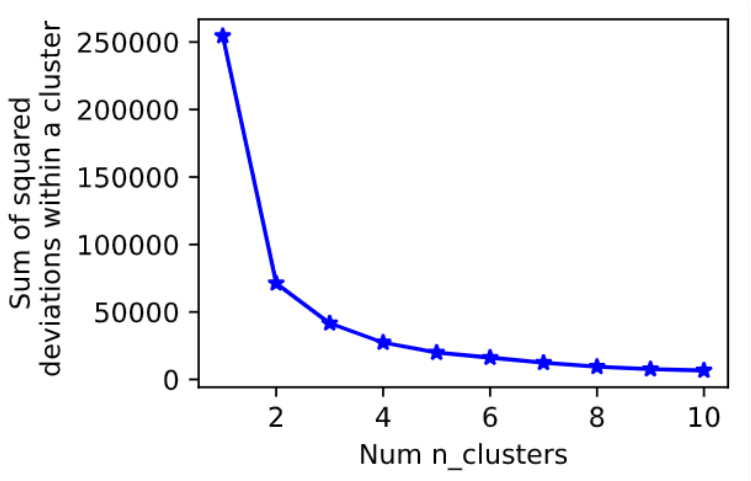

Figure 5.1: The sum of squared deviations within a cluster vs the number of clusters

It can be seen from this that when the number of clusters is 2, an obvious inflection point is formed. The changes in the sum of squared deviations within the clusters corresponding to the clusters after 2 are very small, and the reasonable value of k should be 2.

5.2 Contour Coefficient

This method[12] comprehensively considers the two information of cluster density and dispersion. If the data set is divided into ideal clusters, the samples in the corresponding clusters will be very dense, while the samples between clusters will be very scattered. Define the contour coefficient of the sample point i:

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{5.1}$$

When the S_i get close to 1, it means that the k value are sorted in very ideal clusters. When s_i get close to zero, this means that the points in the cluster are close to each other and clusters are far from each other.

From the Figure 5.2, we could see that the k value that has a corresponding contour factor closest to 1 is 2, which means k equals to 2 is the best option.

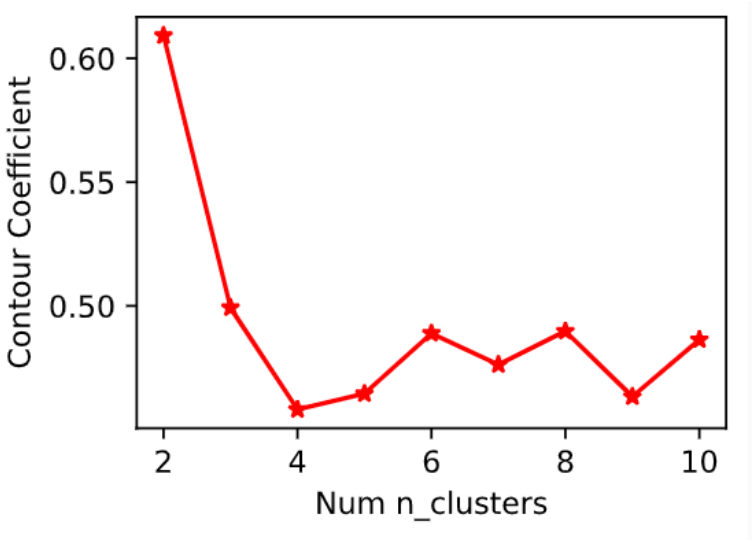

Figure 5.2: Contour Coefficient as k changes

6 Strengths and Weaknesses

In this section, we listed some strengths and weaknesses for our work.

6.1 Strengths

- **Visualization** In order to more vividly reflect our research, we have included many pictures in the paper. In order to express some modeling processes more clearly, we have drawn many sketches by hand, which can help to better understand the model.
- **Consider comprehensively** In the process of modeling, our team consulted more information and considered more comprehensive factors. Whether it is about the modeling of Lake Mead volume or wastewater recycling, we have completed the model based on the results of many documents.
- **Wide applicability** Our model is more general, not just for Lake Mead. For lakes in other parts of the world, our model can also be used to model and predict.
- **Mechanism analysis** Our model is mainly obtained through mechanism analysis, so it can clearly describe the relationship between variables. For example, in the part about quantifying wastewater recycling, we used a differential equation model.

6.2 Weaknesses

- **Not available in special conditions** In out model, we did not consider the situation that the lake may freeze. It is based on the fact that Lake Mead is in a desert. As a result, if the same model is used for lakes which will freeze, the result will not be precise.
- **Not effective in special conditions** Our model assumes that rainfall is constant and that the climate will not change significantly in the short term. In fact, under certain conditions, these assumptions will fail.

References

[1] Bureau of Reclamation. "Reclamation Announces 2022 Operating Conditions for Lake Powell and Lake Mead." August 16, 2021. Found at: https://www.usbr.gov/newsroom/#/news-release/3950.

[2] United States Environmental Protection Agency. "Basic Information about Water Reuse." Updated June 4, 2021. Found at https://www.epa.gov/waterreuse/basic-information-about-water-reuse.

[3] United States Environmental Protection Agency. "Basic Information about Water Reuse." Updated June 4, 2021. Found at https://www.epa.gov/waterreuse/basic-information-about-water-reuse.

[4] Linacre, Edward T.. "A simple formula for estimating evaporation rates in various climates, using temperature data alone." Agricultural Meteorology 18 (1977): 409-424.

[5] https://zhidao.baidu.com/question/585213573.html

[6] https://baike.baidu.com/item/

[7] https://baike.baidu.com/item/

[8] https://www.usbr.gov/lc/riverops.html

[9] https://www.usbr.gov/lc/riverops.html

[10] https://zhidao.baidu.com/question/1868014207188042507.html

[11] Wikipedia Elbow method. https://en.wikipedia.org/wiki/Elbow_method_(clustering)

[12] Zhang Y, Liu N, Wang S (2018) A differential privacy protecting K-means clustering algorithm based on contour coefficients. PLoS ONE 13(11): e0206832.