Shuo Chen
2020012894
chenshuo20@mails.tsinghua.edu.cn

Homework Assignment 3
Environmental Simulation Practice

2023-05-09

# 1 Introduction

This experiment provides a dataset of sewage pipe network length for each month from 1978 to 2010, with a total of 396 records. It is required to use appropriate time series tools to forecast the length of the urban sewage pipe network for a period of time in the future.

# 2 Smoothing of Time Series and Seasonal Decomposition

The raw data is shown below.

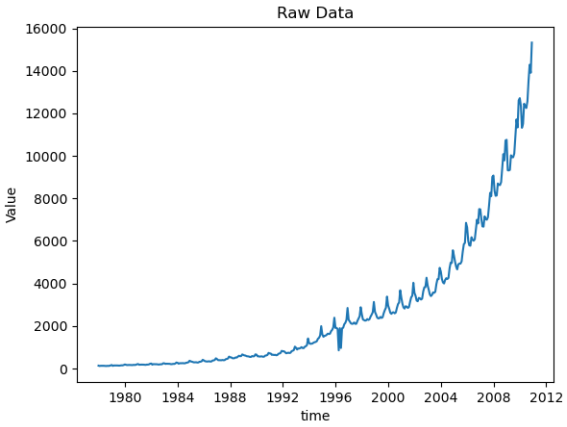


Abbildung 1: Raw Data

## 2.1 Smoothing of Time Series

We use lowess method to make the curve smooth. The details can be found in the code. And the result is

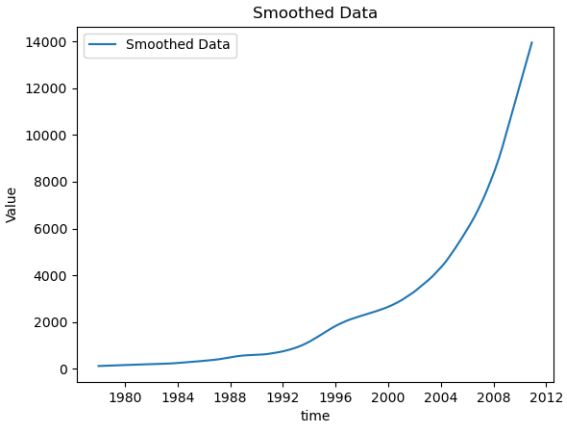


Abbildung 2: Caption

We find that the raw data almost obedient to exponential growth. So we take a logarithm for all values. And we use the ADF test. The original sequence is differenced once, that is, a first-order lag. ADF test is performed on the differenced sequence, and the test result shows that the sequence is stationary at this point.

| | ADF | p-value |
|---|---|---|
| Raw Data | -3.612 | 0.006 |
| Smoothed Data | -2.661 | 0.081 |

In this case, it means that we have conducted the Augmented Dickey-Fuller (ADF) test on a time series data set and the test result indicates that the null hypothesis of non-stationarity of the time series can be rejected at a significance level of 0.05, which means the time series is likely stationary.

## 2.2 Seasonal Decomposition

There are several reasons why we should do seasonal decomposition. Firstly, by decomposing the time series into its individual components, we can gain a better understanding of the underlying patterns and trends in the data, which can help to inform forecasting and decision-making. Secondly, the decomposition can help to identify the sources of variation in the time series, which can be useful for detecting anomalies or outliers in the data. Lastly, the decomposition can provide a basis for modeling and forecasting the time series using different techniques.
And we get the result:

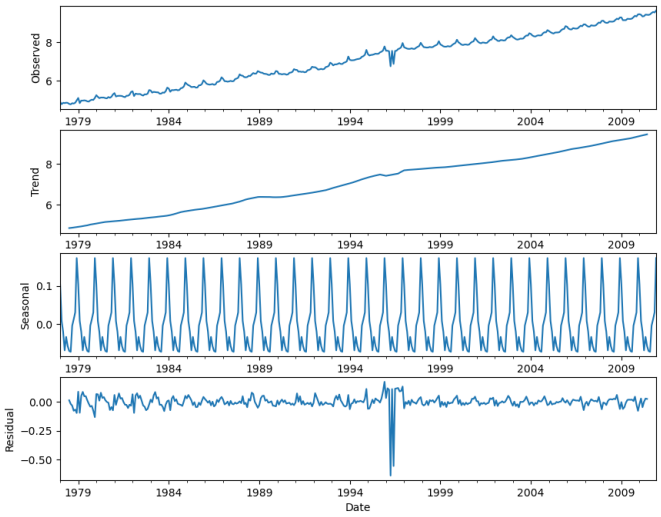


Abbildung 3: Seasonal Decompose

# 3 Forcast by ARIMA Model

We can use ACF and PACF plots to choose candidate models, and the plots are shown below:

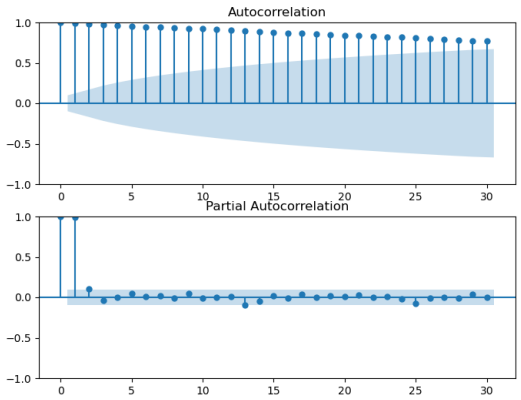


Abbildung 4: ACF and PACF

And we choose the $(p, d, q) = (1, 1, 1)$ to forecast.
And we get the result:

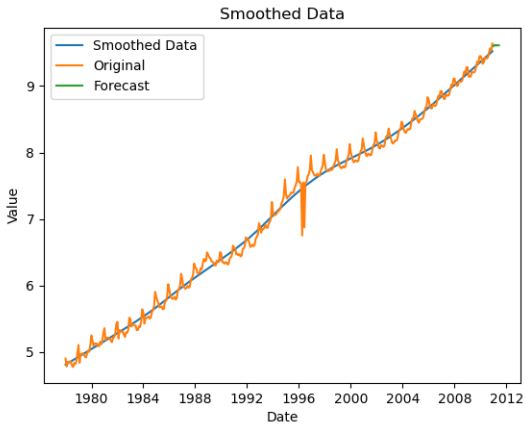


Abbildung 5: Forecast

| Time | Value |
|---|---|
| 2011-01-01 | 14762.192644 |
| 2011-02-01 | 15026.554904 |
| 2011-03-01 | 14901.537499 |
| 2011-04-01 | 14960.251894 |
| 2011-05-01 | 14932.586828 |
| 2011-06-01 | 14945.602131 |

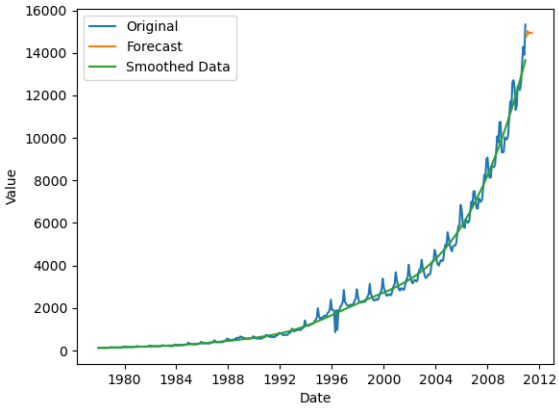


Abbildung 6: True Forecast Result

# 4 Discuss analysis

The ARIMA model can fit the data well, despite the presence of some outliers in the data, such as those around 1996 (as can be seen in the residual plot). However, the overall trend can still be well captured by this model, and it can be used to make predictions within a certain range. The limitation is that the predictive performance may not be ideal, and more data may be needed to train a better model (such as neural networks).
Also, by using this model, we can find that ARIMA models have both advantages and disadvantages. One of the benefits of ARIMA models is their interpretability and flexibility in capturing various temporal patterns, such as trend, seasonality, and autocorrelation(you can know this by the whole report pipeline). Additionally, they can handle missing data and irregular time intervals, and there are established techniques and tools available for their implementation.
However, ARIMA models also have several limitations. They assume linearity and stationarity, which may not hold for some time series data, and can be sensitive to outliers and unusual observations, which can affect the

model's accuracy and stability. The 1996 years' data make the model performance worse. Furthermore, they may require a large amount of data to estimate the model parameters accurately, and they may not be able to capture complex nonlinear patterns or interactions among variables, which could require more advanced modeling techniques. In this homework the data are not that abundance, if we have more data, the model may have better performance.