homework4

April 29, 2024

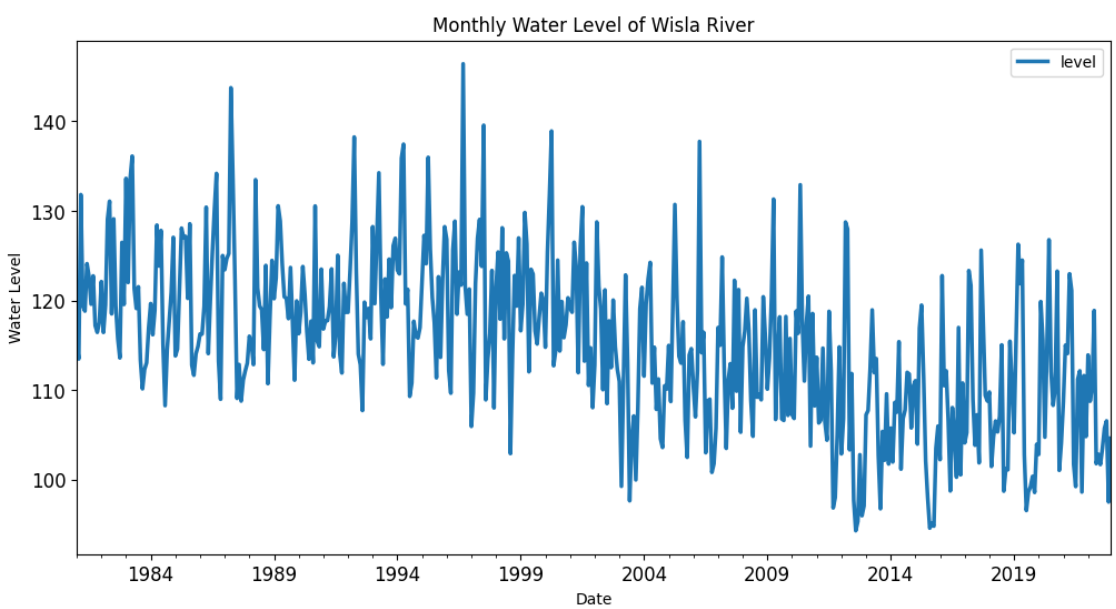
1 Homework 4 - time series forecasting

2 Dataset preparation

	stationid	name	water	hyear	hmonth	day	level	month
0	149180140	WISŁA	Wisła (2)	1981	1	1	114	11
1	149180140	WISŁA	Wisła (2)	1981	1	2	114	11
2	149180140	WISŁA	Wisła (2)	1981	1	3	114	11
3	149180140	WISŁA	Wisła (2)	1981	1	4	114	11
4	149180140	WISŁA	Wisła (2)	1981	1	5	113	11

I combine the year, month and day of the measurment into one feature `date`, which will simplify the futher modelling.

I decided to investigate the monthly time-series forecasting, therefore for each month I extract the mean value of the water level in Vistula river.

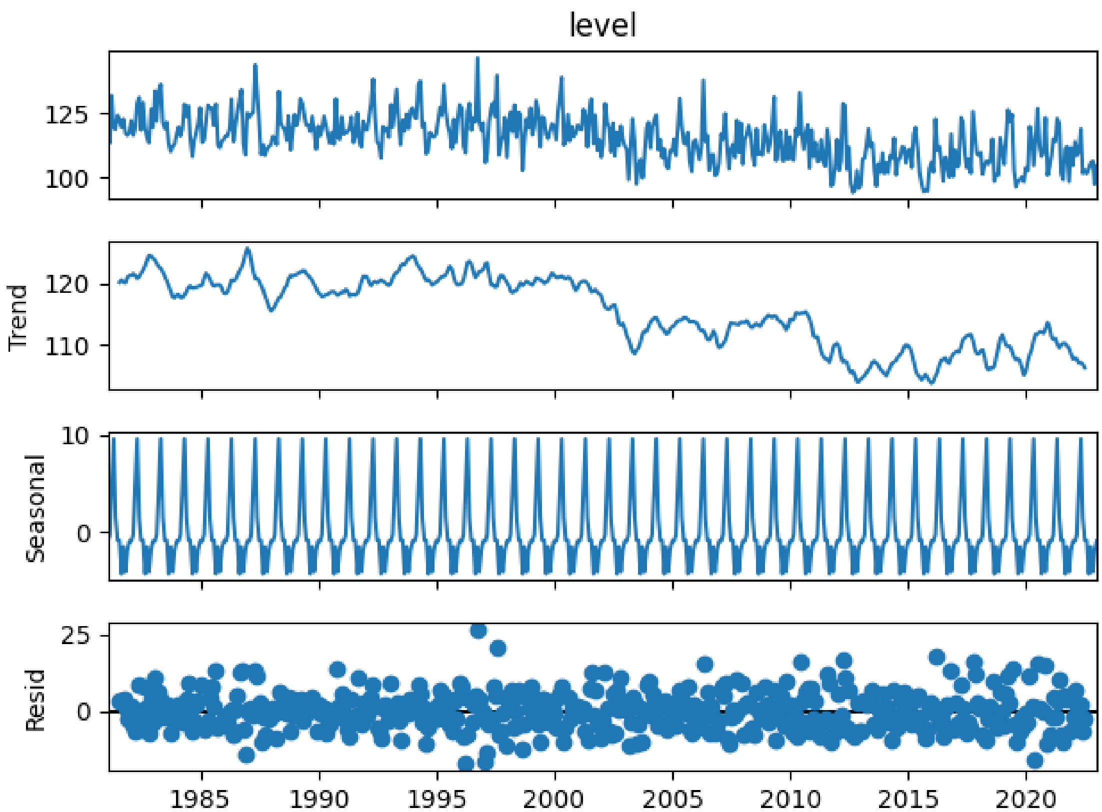


3 Trend and seasonality

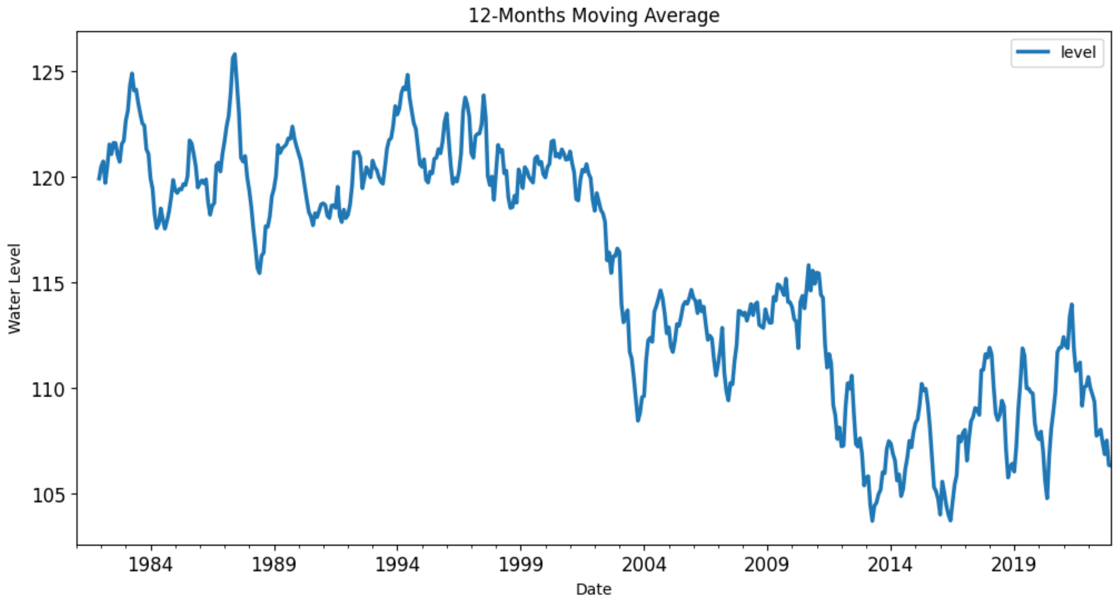
3.1 Additive model

In the additive model, the time series (Y) is assumed to be the sum of the three components: $Y = T + S + R$ where:

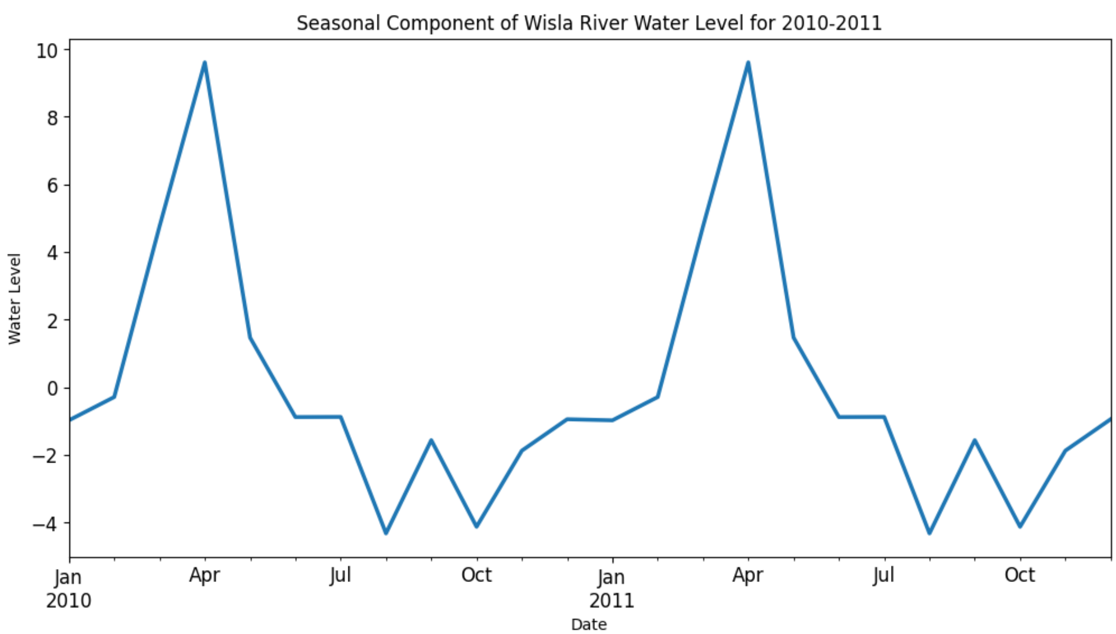
1. Y is the observed time series value
2. T is the trend component and it represents the long-term pattern or direction of the time series.
3. S is the seasonality component, which captures the regular, periodic fluctuations in the time series
4. R is the residual component and represents the remaining part of the time series after removing the trend and seasonality, capturing random fluctuations

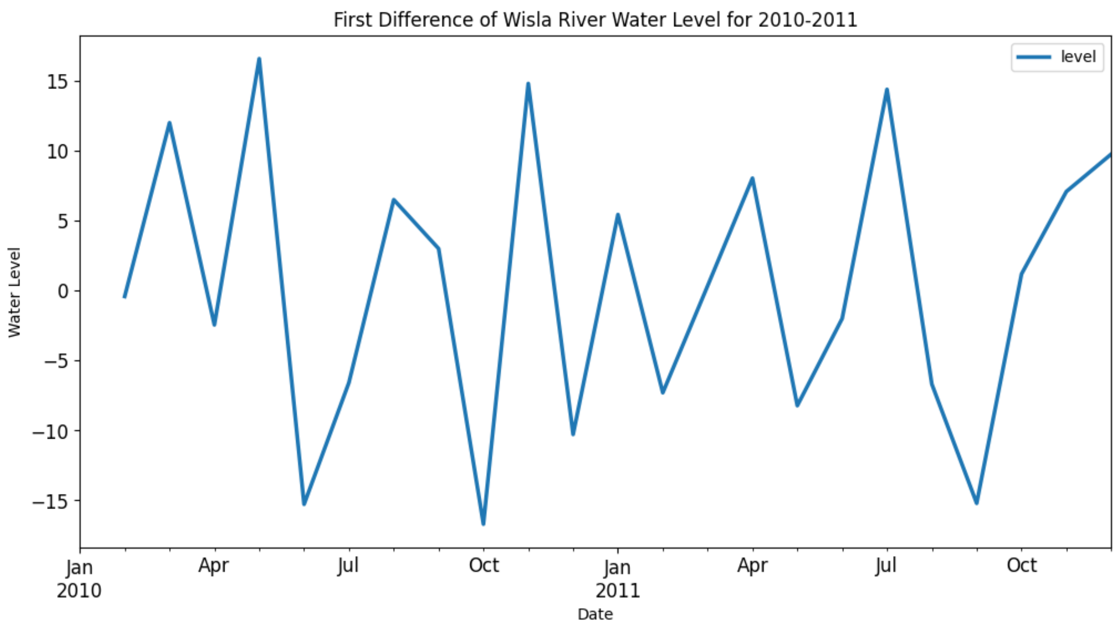


1. Trend: Over the period from 1981 to 2024, there is a definite decreasing trend observed in the water levels. Theobservation is validated by examining the 12-month rolling mean of the water level data, which shows a clear downward trajectory when calculated over the moving window of 12 months.



2. Seasonality: We can notice that the additive model identified some regularities within 12-month time intervals, which overall makes a lot of sense, as they fit into the natural annual cycle. I further investigate the seasonality factor, looking closer at years 2010-2011. As we can see, the biggest spike appears in April, when the winter snows melt and the potential of high water level in river is





4 Stationarity

4.1 Augmented Dickey-Fuller test

```
ADF Statistic: -1.338098
p-value: 0.611567
Critical Values:
        1%: -3.444
        5%: -2.868
        10%: -2.570
```

Based on the p-value of the ADF test we should not reject the null hypothesis. Therefore it is likely that this time series is non-stationary. I validate this finding by looking at complementary KPSS test.

4.2 Kwiatkowski-Phillips-Schmidt-Shin test

```
KPSS Statistic: 3.619952
p-value: 0.010000
Critical Values:
        10%: 0.347
        5%: 0.463
        2.5%: 0.574
        1%: 0.739
```

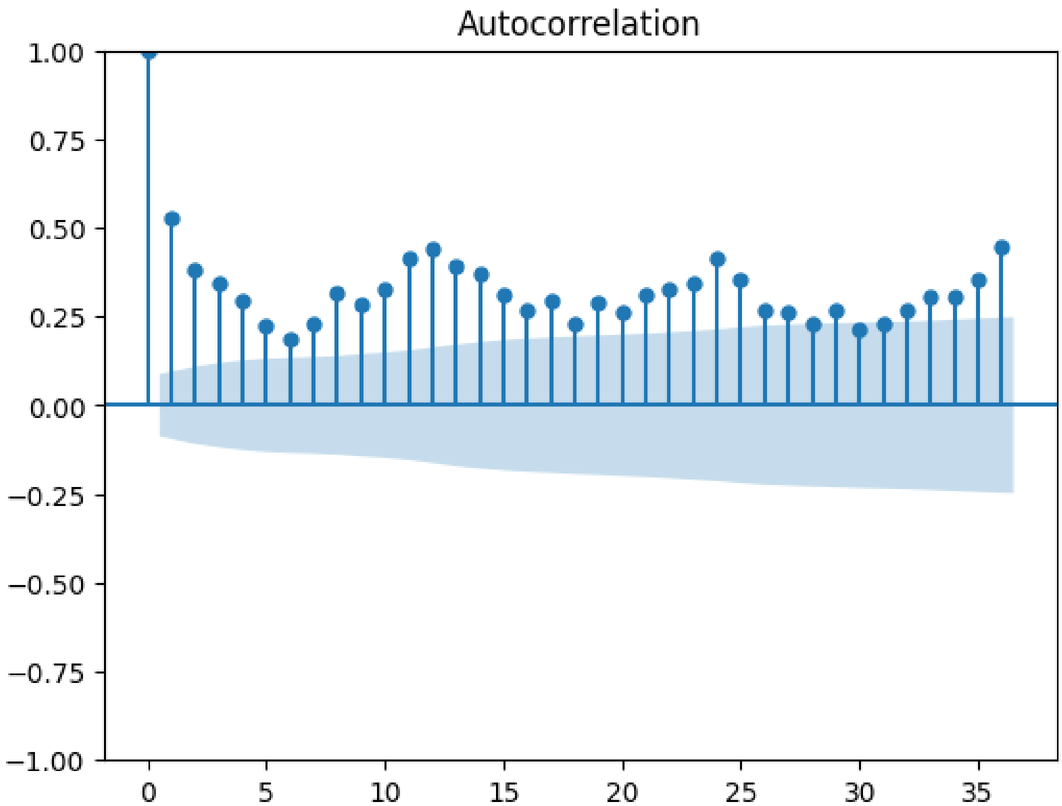
```
C:\Users\jan20\AppData\Local\Temp\ipykernel_16572\3663233560.py:1:
InterpolationWarning: The test statistic is outside of the range of p-values
available in the
```

```
look-up table. The actual p-value is smaller than the p-value returned.
```

```
result = kpss(monthly_df['level'].dropna(), regression='c', nlags='auto')
```

Unlike the ADF test, which has a null hypothesis of non-stationarity, the KPSS test has a null hypothesis of stationarity. P-value is lesser than 0.05 indicating that indeed, the time-series related to water level in vistula river is not stationary.

5 Autocorrelation



We observe the highest correlation among observations when the lag is a multiple of 12.

6 ARIMA model

In previous section I discussed the non-stationarity of the Vistula river water level time-series. Therefore, I decided to use the ARIMA model instead of ARMA model, as it can handle non-stationary time series data.

I decided to split the dataset into 3 parts: 1. Training: 70% 2. Validation: 15% 3. Testing: 15%

I also define the grid of values for `p`, `d` and `q` - parameters of ARIMA model. 1. `p` - autoregressive order (AR), which specifies number of lags used as predictors 2. `d` - the integrated order (I), representing the number of times needed to difference time series to get stationary process 3. `q` - the moving average order (MA), specifing number of lagged errorrs used as predictors in the model.

In order to find the optimal model, I perform the grid search. In advance I suspect that the `p` might be a multiple of 12, while `d` should be at least equal to 1 (as $d = 0$ would imply stationarity of process, which we assessed as not true).

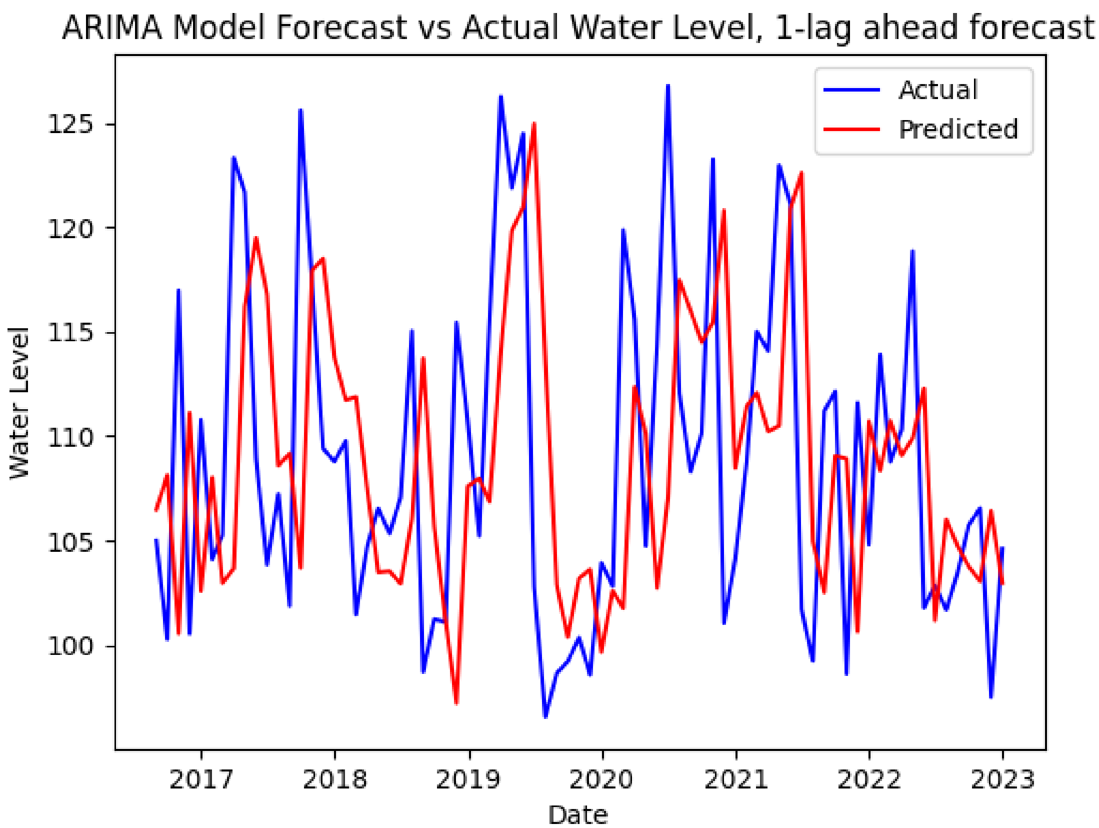
```
Best Parameters: (12, 2, 0)
Best RMSE on Validation Set: 8.252670692211167
```

We achieved a sensible RMSE on validation set for the best model. Now it is time to make a one-step ahead prediction for the test data.

6.1 Results

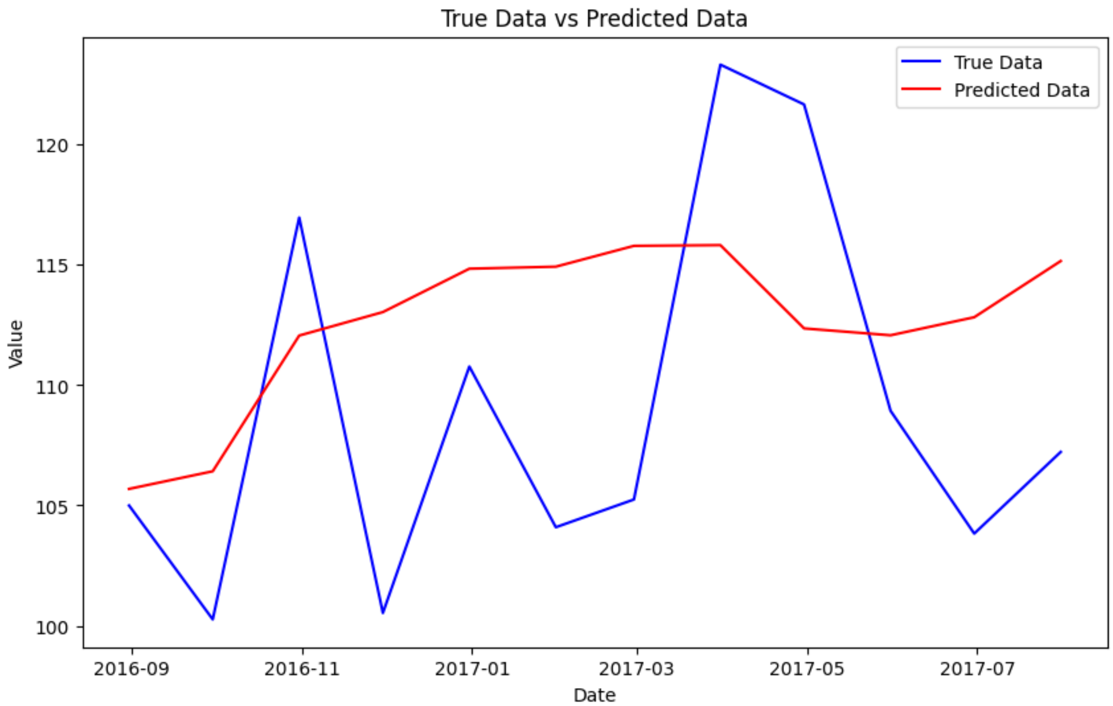
6.2 1-lag ahead forecast

```
Test RMSE: 9.266
```



As we can notice, the model was generally accurate. The biggest departures can be seen in places of big spikes, where the model struggled to be accurate.

6.3 1 year forecast



The long term forecast is much worse than 1-lag ahead forecast. There is close to none similarity between the prediciton and true value of water level.