FINAL PROJECT:
SPECIALIZED MODELS: TIME SERIES AND SURVIVAL ANALYSIS

Antonio Jimenez

IBM ON COURSERA . April, 2023

Table of Content

Introduction

The quality of the water that we drink is vital for our health. For that reason, Bolivia's government has strict norms regarding with this subject. It requires constant measurements of key properties of the water at key points of the distribution network. One of those key points are treatment plants. In our case, the facility is the Achichicala treatment plant. It is located in the city of La Paz, Bolivia.

One key property is the PH level of the water. So, daily measurements are performed to make sure it is under safe levels. The objective of the current work is to build a model to predict future levels of PH in this plant. The resulting model will be used to improve resource management in the company that is in charge of it.

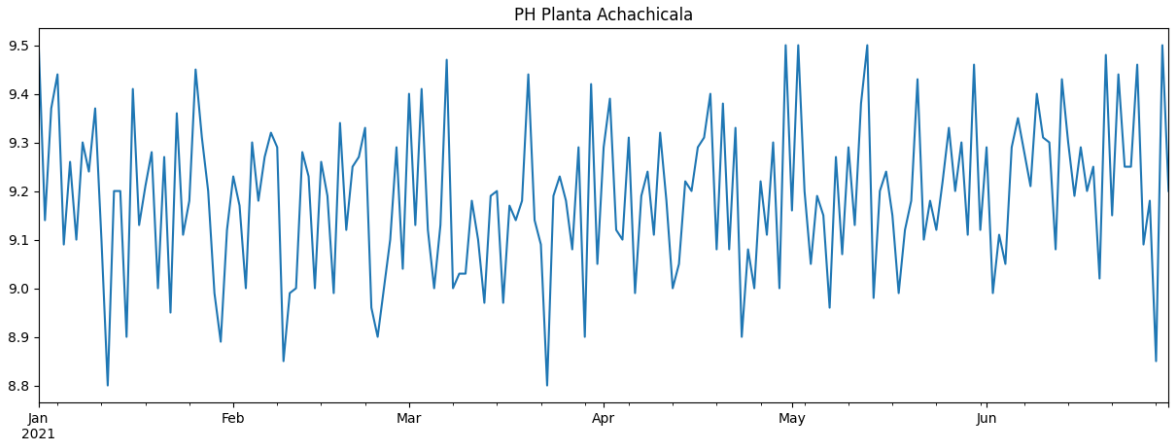
Description of the Data

The data was collected as a part of continuous testing program, required by regulation. The data comes from the government entity that is in charge of regulating water supply companies:" AUTORIDAD DE FISCALIZACIÓN Y CONTROL SOCIAL DE AGUA POTABLE Y SANEAMIENTO BÁSICO".

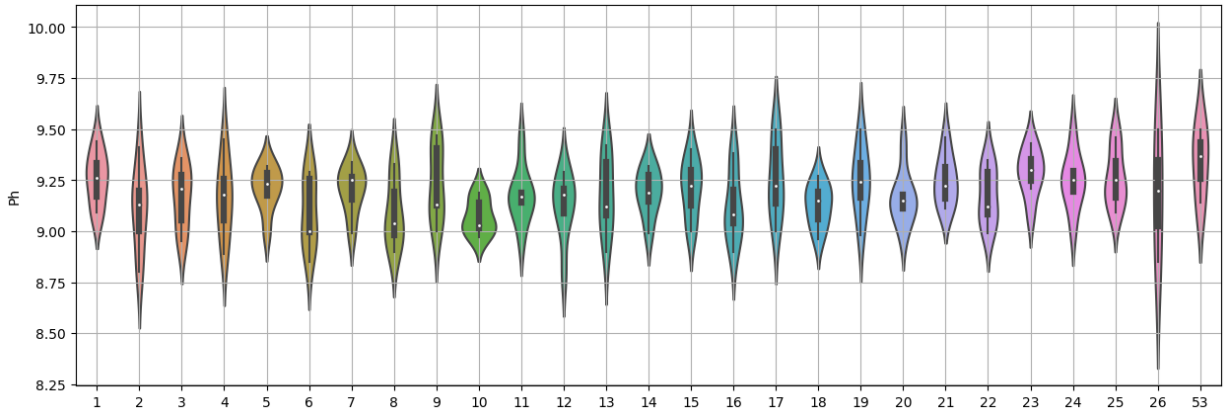
It contains PH measurements of the Achichicala treatment plant, covering the period from January 1st to June 30th, 2021.

Exploratory Data Analysis

The first step is to plot the series. In this case we have daily measurements collected over a period of six months.

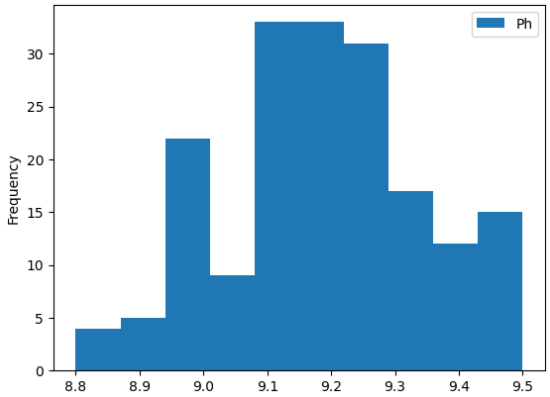


From the plot, no tendency is observed, and also variance seems to be constant. To have a different perspective, violin plot is also shown.



Again, mean and covariance seems to be constant in the range of analysis.

In order to be certain of the stationarity of the data, a frequency plot is shown and the Dickey-Fuller is performed.



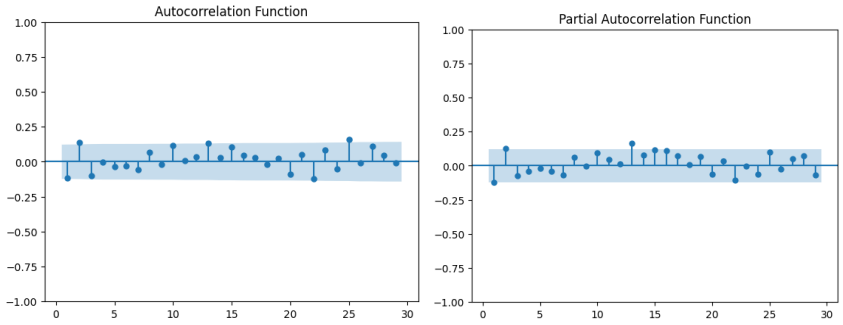
ADF: -7.912719371390771

p-value: 3.908958003929843e-12

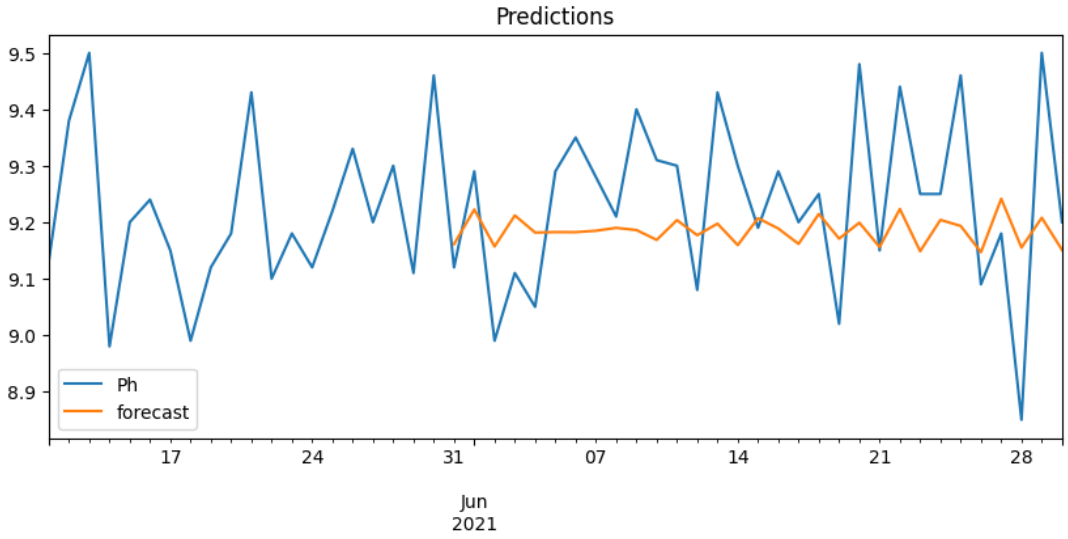
From the results, it is clear that the data is stationary. So we can proceed with the model building process.

Modeling

The first model to be build is ARMA. This is ideal for the no-trend, no-seasonality data we have. For this, it is necessary to choose the right values for p and q. These values can be determined from the autocorrelation and partial autocorrelation functions.



We choose p=2 and q=2. With these parameters, the model is fitted. The Prediction results is shown below



The parameters for this model are:

```
const     9.186835
ar.L1    -1.150626
ar.L2    -0.701474
ma.L1     1.068954
ma.L2     0.734240
sigma2    0.023326
```

The second model to be build is a simple RNN model. For this model, the following parameters were used.

```
series_days = 56
input_hours = 12
test_hours = 24
```

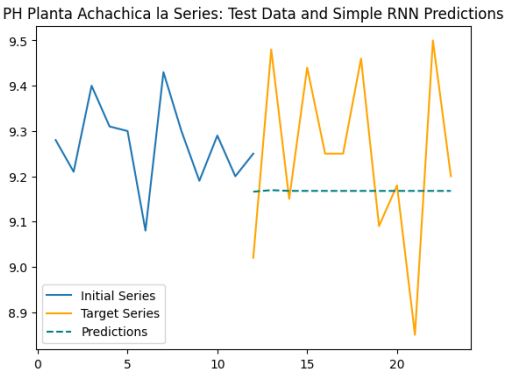
```
Model: "sequential_4"
_________________________________________________________________
```

```
 Layer (type)                 Output Shape              Param #
=================================================================
 simple_rnn_3 (SimpleRNN)    (None, 30)                960

 dense_4 (Dense)             (None, 1)                 31

=================================================================
Total params: 991
Trainable params: 991
Non-trainable params: 0
_________________________________________________________________
```

The results were the following



Finally, an LSTM model was fitted. The parameters for this algorithm were the following.

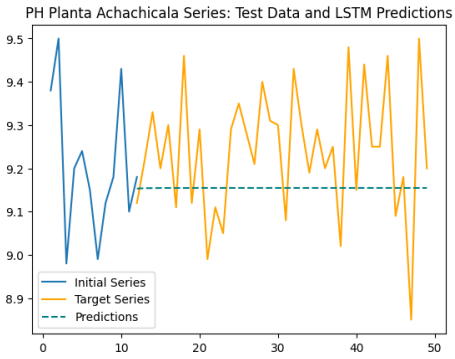
```
series_days = 100
input_hours = 12
test_hours = 50
```

```
Model: "sequential_5"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 lstm_1 (LSTM)               (None, 70)                20160

 dense_5 (Dense)             (None, 1)                 71

=================================================================
Total params: 20,231
Trainable params: 20,231
Non-trainable params:
_________________________________________________________________
```

With the results:



Final model

Even though the high complexity of the machine learning models, the increase in prediction power in not significant compared to the simple ARMA model. Therefore, ARMA is the model chosen for future use. Its simplicity and explainability makes it very portable and easy to understand for the future users of the model.

Key findings and insights

From the initial plot of the series and the frequency plot, it can be seen that the variability in the data is random and normally distributed. Furthermore, the data is stationary which means that their statistical properties remain constant.

The prediction from the ARMA model shows some variance; But, the Deep Learning models output a constant prediction. This shows that it is possible that the ARMA model has some overfitting.

Even though the complexity of the LSTM algorithm is significantly higher than the RNN model, the improve in prediction is not significant. So, RNN is preferable over LSTM for its relative simplicity.

Suggestions and next steps

It is necessary to acquire more data, year round and for several years. This could help to build a more powerful model, one that can capture long term seasonality and trends. These long term patterns were not present in the small range of our data.

Also, with more data and with long term patterns, it is possible to use more complex models, such as ARIMA or SARIMA. In addition, power prediction of deep learning models is going to increase with additional data.