Progress Report
Big Bang to Big Data

Kamalpreet Kaur, Subarna Chaki

Supervisor: Dr. Ralf Timmerman, Prof. Dr. Frank Bertoldi

Contents

1 Aim

The aim is to develop a time series analysis machine learning algorithms for forecasting precipitable water waver content (PWV from here onwards) for cutting-edge Cerro Chajnantor Atacama Telescope.

2 Motivation

This project is a part of the "Big Bang to Big Data" cluster which combines radio astronomical research with data science expertise. It belongs to the work package (WP from here onwards) 4. WP 4 sets itself the task of ensuring connectivity to data streams and repositories relevant for astrophysical research.

The scientific motivation behind the project lies in CCAT, a telescope designed to operate within the submillimeter to millimeter wavelengths. Located at an elevation of 5600 meters on Cerro Chajnantor in the Atacama desert of northern Chile [1], the project faces a significant challenge due to the presence of water vapor in the atmosphere during observations. Our development of a machine learning algorithm for forecasting aims to address this challenge by optimizing observation scheduling. By doing so, we not only enhance our ability to achieve scientific objectives but also mitigate operational costs associated with telescope usage.

3 About the data

As the CCAT is not yet operational, we are currently utilizing data from the Atacama Pathfinder Experiment (APEX). APEX, boasting a 12-meter diameter telescope, operates within the millimeter and submillimeter wavelength ranges - falling between

infrared light and radio waves. Situated in the Chajnantor region, this location is optimal for such observations due to its status as one of the driest places on Earth and its elevation exceeding that of observatories on Mauna Kea by over 750 meters, and the Very Large Telescope (VLT) on Cerro Paranal by 2400 meters [5]. Historical weather data from APEX can be accessed here. Users can download daily data from this source. Employing the script available here, we aggregated data for the entire year, concatenating daily records for comprehensive yearly weather data retrieval.

The data consists of radiometer readings, which include the PWV content, temperature, dew point, humidity, pressure, wind speed, and wind direction. there is no fixed frequency in which data is recorded. Our primary focus for prediction is PWV, as it is a crucial parameter that determines atmospheric transparency for sub-millimeter observations [2]. Once we are able to forecast PWV, we can similarly forecast other weather parameters by better understanding their respective trends.

The primary issue with the dataset is the occurrence of faulty radiometers or inoperable telescope periods, resulting in numerous missing values. Consequently, we need to devise a method to fill these gaps while preserving the daily trends and minimizing disruption to the dataset.

4 Gap filling: exponential smoothing

4.1 Introduction

Exponential smoothing techniques yield forecasts that are weighted averages of historical data, with the weights decreasing exponentially with increasing observational time. Stated otherwise, the higher the related weight, the more recent the observation [4].

There are two key components in the time series: seasonality and trend and how we

use the smoothing method (additive, damped or multiplicative) help us to select the method of exponential smoothing [4]. We use seasonal additive method to fill our gaps and if there are not sufficient points available then we switch to trend additive method.

The Holt-Winters seasonal additive method with an additive trend involves the following components [4]:

1. Level Equation
$$l_t = \alpha(X_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1})$$
2. Trend Equation
$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$
3. Seasonal Equation
$$s_t = \gamma(X_t - l_t) + (1-\gamma)s_{t-m}$$
4. Forecast Equation
$$F_{t+k} = l_t + kb_t + s_{t+k-m}$$

4.2 Parameters

- l_t: Level component at time t. It represents the smoothed value of the series after removing the seasonal component.
- b_t: Trend component at time t. It captures the rate of change in the level.
- s_t: Seasonal component at time t. It captures the seasonal fluctuations.
- X_t: Actual observed value at time t.
- α: Smoothing parameter for the level component ($0 < \alpha < 1$). It determines the weight of the current observation in the level equation.

- β: Smoothing parameter for the trend component ($0 < \beta < 1$). It determines the weight of the current trend estimate.
- γ: Smoothing parameter for the seasonal component ($0 < \gamma < 1$). It determines the weight of the current seasonal estimate.
- m: Length of the seasonal cycle. For example, if the data has a yearly seasonality with monthly data points, $m = 12$.
- k: Number of periods ahead to forecast.

4.3 Explanation

1. Level Equation: The level equation updates the smoothed level of the series by combining the current deseasonalized observation ($X_t - s_{t-m}$) and the previous level adjusted by the trend ($l_{t-1} + b_{t-1}$).
2. Trend Equation: The trend equation updates the estimate of the trend by comparing the current and previous levels ($l_t - l_{t-1}$) and smoothing this difference.
3. Seasonal Equation: The seasonal equation updates the estimate of the seasonal component by comparing the current observation and the current level ($X_t - l_t$) and smoothing this difference.
4. Forecast Equation: The forecast equation combines the current level, the trend, and the seasonal component to generate forecasts for future periods.

   By using these equations, the Holt-Winters method can adapt to changes in the level, trend, and seasonality of the data, providing accurate forecasts even in the presence of complex patterns.

4.4 Algorithm

The algorithm systematically processes a time series dataset by identifying intervals with missing values or reaching the dataset's end, segmenting these into distinct time series [3]. It then applies exponential smoothing to fill the gaps between adjacent segments, ensuring continuity. The filled data is plotted alongside the original time series, providing a clear visualization of the imputation process. This iterative approach ensures comprehensive handling of data gaps, enabling the algorithm to efficiently prepare the time series for further analysis or visualization tasks.

1. **Start:**
   - Begin the execution of the gap-filling algorithm.
2. **Initialize variables:**
   - Create an empty list to store segmented time series.
   - Determine the start time of the time series.
   - Set a flag to indicate the presence or absence of a gap.
3. **Loop through data:**
   - Iterate through each index-value pair in the resampled dataframe.
   - **Check for NaN or end of series:**
     - If the current value is NaN or if it's the last index in the dataframe:
       - Determine the end time of the current time series segment.
       - Append the segment to the list of time series.
       - Set the gap flag to indicate the presence of a gap if necessary.
   - **Check for gap:**
     - If a gap is detected:

- Update the start time for the next time series segment.

4. **Generate prediction and fill gap between adjacent segments:**
   - Apply a specified prediction method to fill the gap between adjacent time series segments.
5. **Plot prediction:**
   - Visualize the filled gap alongside the original time series data.
6. **Repeat steps 3-5 until all gaps are filled:**
   - Iterate through the remaining segments of the time series, filling gaps and plotting predictions as needed.
7. **Stop:**
   - End the execution of the gap filling algorithm.

4.5 Application of the code

The following plots will help to visualize what the code is doing: As an example, for one year i.e. 2012,

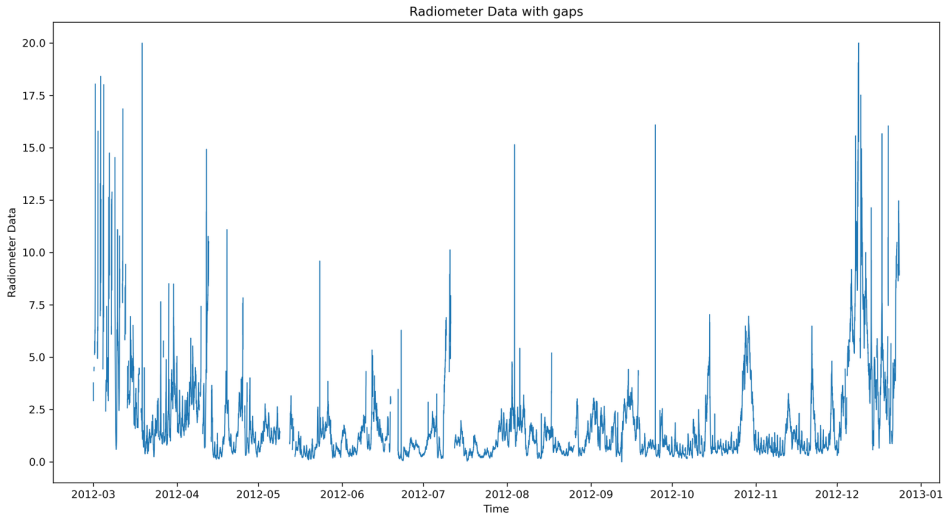

Figure 1: Radiometer Data for the year 2012 with gaps.

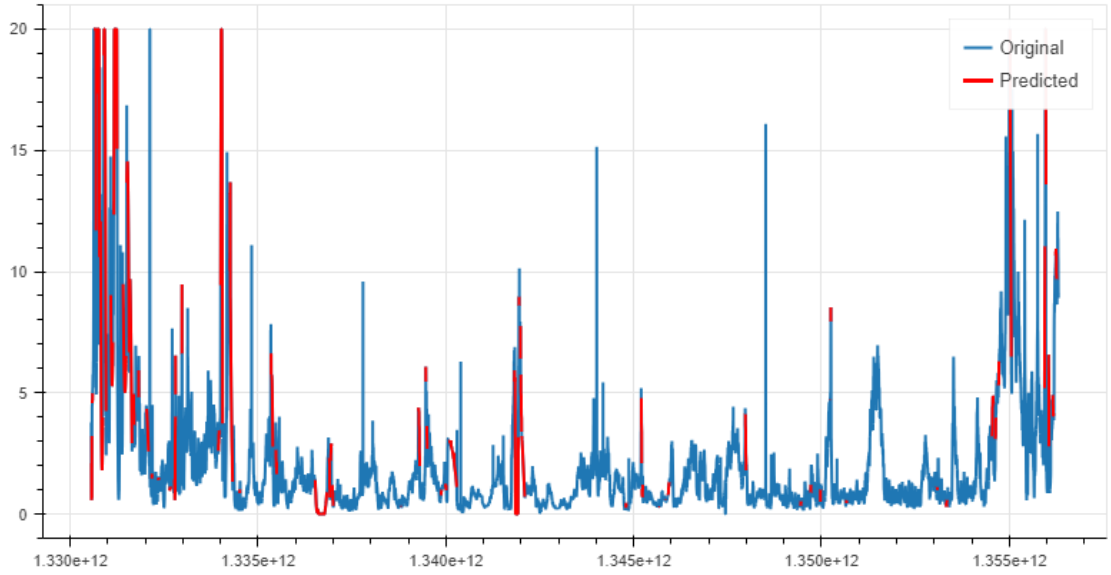

Figure 2: The gaps are filled with exponential smoothing for the year 2012.

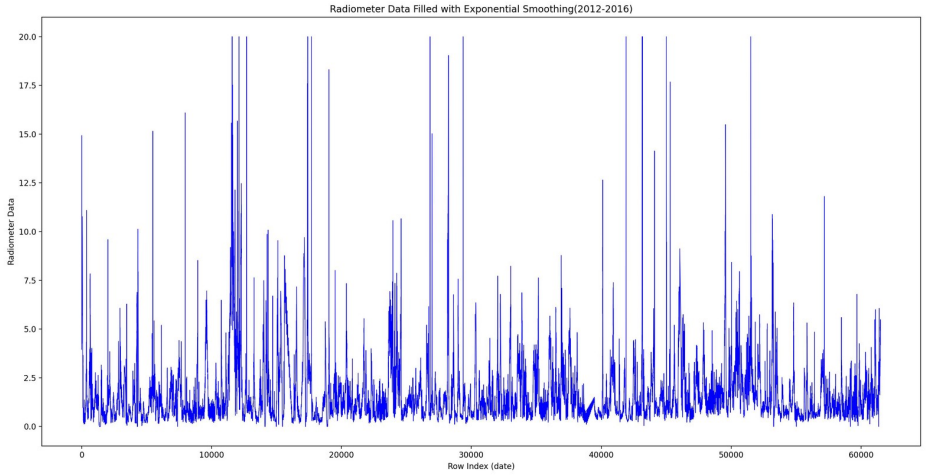

Figure 3: The plot shows the radiometer data from 2012 to 2016 where the gaps are filled using exponential smoothing.

References

[1] CCAT prime. `http://www.ccatobservatory.org/index.cfm/page/about-ccat.htm`, 2024.

[2] APEX. APEX - weather. `https://www.apex-telescope.org/ns/weather/`.

[3] Carlo Alberto Carrucciu. Filling large gaps in time series using forecasting. `https://medium.com/targetreply/filling-large-gaps-in-time-series-using-forecasting-2f6db5f5286b`, April 2021.

[4] Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*.

[5] European Southern Observatory. APEX. https://www.eso.org/public/teles-instr/apex/.