UNIVERSIDADE D
COIMBRA

Weather Time Series Analysis and Forecast

Time Series Analysis and Pediction

Master in Computer Science Engineering
2024/2025

October 25, 2024

Authors

João Moreira

joaomoreira@student.dei.uc.pt

PL1

2020230563

Tomás Pinto

tomaspinto@student.dei.uc.pt

PL1

2020224069

Contents

1. Introduction

1.1. Problem description

The analysis and forecasting of weather patterns play a significant role in a variety of fields, from agriculture to energy management and infrastructure planning. By studying historical weather data, we can identify trends and seasonal variations that help us better understand local weather conditions and predict future patterns.

This project focuses on the time series analysis of mean temperature data for a selected location. The goal is to model and forecast temperature trends using different time series analysis techniques, including exponential smoothing and other machine-learning methods, to assess their predictive accuracy.

In addition to forecasting mean temperature, we extend the analysis by incorporating multivariate models that consider other weather variables. This approach allows for a more comprehensive understanding of the relationships between different weather factors, ultimately leading to more reliable forecasts.

1.2. Dataset description

The dataset used for this project was obtained from the Open-Meteo Historical Weather API, which provides daily weather observations for various locations worldwide. For this analysis, we selected data from *Tokyo* for the period spanning from January 1st, 1940, to the present. The primary variable of interest is the mean daily temperature, supplemented with additional weather variables for multivariate analysis.

The dataset contains daily observations structured with a date column and corresponding values for each weather variable. An initial inspection of the dataset revealed some missing entries, in the first day. We opted to remove the first day of the dataset because do not impact the time dependency of the time series.

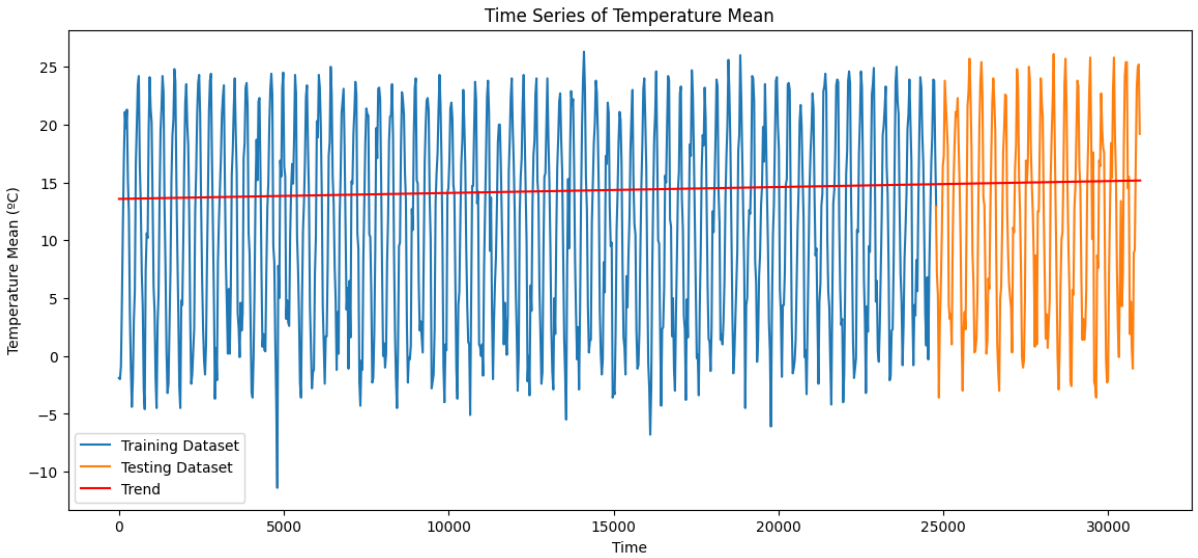


Figure 1: Visualization of mean temperature TS

In order to generate the figure 1 and facilitate the visualization, we plotted the mean temperature value of the first day of each month. First things first, there are some aspects to notice in our mean temperature time series:

- There is a very small trend
- The time series is remarkable because of the constant seasonal pattern
- The seasonal pattern does not change with the trend

Based on these observations, the additive model is appropriate. In this model, the three components are independent and it is usually used when the magnitude of seasonality and noise do not change with the trend, which aligns with the characteristics observed in the time series:

$$x(n) = tr(n) + sn(n) + e(n)$$

As follows, this time series is not stationary at all. In chapters **TO REF** we will see more about this topic.

The last thing to point out is that the first 80% of the dataset will be used for characterisation, modelling, and training and the last 20% will be used to assess the forecasting performance of our models.

2. Decomposition Methods

In this section we present the TS decomposing methods in its main components:

- Trend: Captures the long-term progression in the series. It represents the overall direction in which the data is moving (upward, downward, or stationary) over time.
- Seasonality: Refers to patterns that repeat at regular intervals, often driven by external factors such as seasons of the year.
- Erratic Components: Random variation in the data that cannot be explained by the trend or seasonality. It represents unpredictable fluctuations that might arise from various short-term factors.

2.1. Model Fitting Approaches

Model fitting is a corrector trend method, which involves selecting the right mathematical function or statistical model to capture the trend and seasonality. One common method is *Polynomial Fitting*, where the theory says that a TS can be described as a polynomial as follows:

$$T(t) = a_0 + a_1 t + a_2 t^2 + ... + a_n t^n$$

where $a_0, a_1, ..., a_n$ are the polynomial coefficients, and t is the time variable. For instance, a higher-degree polynomial might capture more complex trend patterns, but it also risks overfitting the data.

2.2. Local Smoothing Approaches

Local smoothing techniques are methods used to extract the trend by averaging over a small, sliding window of the data. Those involve averaging or weighting nearby data points to smooth the series.

2.2.1. Moving Average (MA)

The moving average is a simple and effective technique for smoothing. It calculates the average of a fixed number of neighbouring points (a window) and replaces each point with its average:

$$T(n) = \frac{1}{\sum \omega_k} \sum_{k=-\frac{M-1}{2}}^{\frac{M-1}{2}} \omega_k x(n+k)$$

beign $\omega = [-\frac{M-1}{2}, ..., \frac{M-1}{2}]$ the weights of the filter.

2.2.2. LOWESS/LOESS (Locally Weighted Scatterplot Smoothing)

LOWESS is a more flexible, non-parametric method for smoothing. It fits multiple regressions to local data segments, weighting each point based on its distance from the current data point. This approach is ideal for capturing non-linear trends without assuming a specific form for the trend component.

Unlike moving average, which uses a fixed window, LOWESS adjusts its fit based on the local structure of the data, making it more adaptable in the presence of complex, non-linear trends.

2.3. Seasonality assessment

Identifying and measuring seasonality is crucial for understanding the recurring behaviours in the data, which helps in forecasting and making data-driven decisions.

Two common techniques for assessing seasonality are Filtering and Epoch Averaging.

2.3.1. Filtering

Filtering is a technique used to isolate the seasonal component of a time series by removing trend and noise. Common methods include the already presented moving average filter (low-pass), with a window size often chosen based on the seasonal period, and Fourier transform filtering, which identifies and removes specific frequency components corresponding to the seasonal cycle.

2.3.2. Epoch Averaging

Epoch averaging, also known as seasonal averaging, involves breaking the time series into repetitive cycles (epochs) and averaging the corresponding points from each cycle. For instance, averaging the values for the same month across different years can reveal the seasonal pattern. This method is useful for highlighting consistent seasonal trends by smoothing out random fluctuations.

2.4. Trend and Seasonality removal by differencing

Differencing is a key technique used to remove trend and seasonality from a time series, transforming a non-stationary series into a stationary one.

2.4.1. First-order differencing

This method works by subtracting the value of the current observation from the previous one.

$$\Delta x(t) = x(t) - x(t-1)$$

This operation removes trends or long-term fluctuations in the data, highlighting short-term changes and making the series more stable in terms of its mean and variance. First-order differencing is particularly useful when the series exhibits a linear trend.

2.4.2. First-order seasonal differencing

This approach targets seasonal patterns, which occur at regular intervals. Instead of subtracting consecutive points as the previous technique, it subtracts values from the same season in previous cycles.

$$\Delta x(t) = x(t) - x(t-f)$$

3. Stationary Assessment Methods

Once both trend and seasonality have been removed, we assess the series's stationarity. Ensuring stationarity is essential, as non-stationary data can complicate forecasting models. We then evaluate the series' autocorrelation by constructing correlograms, which illustrate the time-dependent relationships within the data. Finally, we conduct a formal stationarity assessment using statistical tests.

3.1. Autocorrelation and Correlogram

Autocorrelation is a statistical measure that describes how the current value in a time series is related to its past values. In other words, it evaluates the degree of similarity between observations as a function of the time lag between them. The Autocorrelation Sequence (ACS) calculates these correlations for different time lags, allowing us to identify patterns, such as periodicity or seasonality, and understand the dependencies between observations over time.

The confidence bounds on the correlogram are critical for interpreting the results. These bounds indicate the range of values within which we expect the autocorrelations to lie if the data is random (i.e., if there's no meaningful autocorrelation at that lag). If the autocorrelation values fall outside the confidence bounds, it suggests that there is significant correlation at that specific lag, which could indicate a strong seasonal effect or trend.

3.2. Visual Approach

A visual inspection of the time series plot is an intuitive method to assess stationarity. By examining the graph, one can look for a constant mean and variance over time. If the series shows no noticeable trends or long-term shifts in behaviour, it might be considered stationary. However, trends or periodic patterns would indicate non-stationarity, requiring further transformation, such as differencing, to stabilize the series.

3.3. Statistical Approach (Unit Root Test)

For a formal stationarity check, statistical tests like the Augmented Dickey-Fuller (ADF) test are applied. The ADF test evaluates the null hypothesis that the series contains a unit root,

which implies non-stationarity. If the test rejects the null hypothesis, it indicates that the series is stationary. This is a crucial step before applying any time series modeling techniques, as many models assume the series is stationary for accurate forecasting.

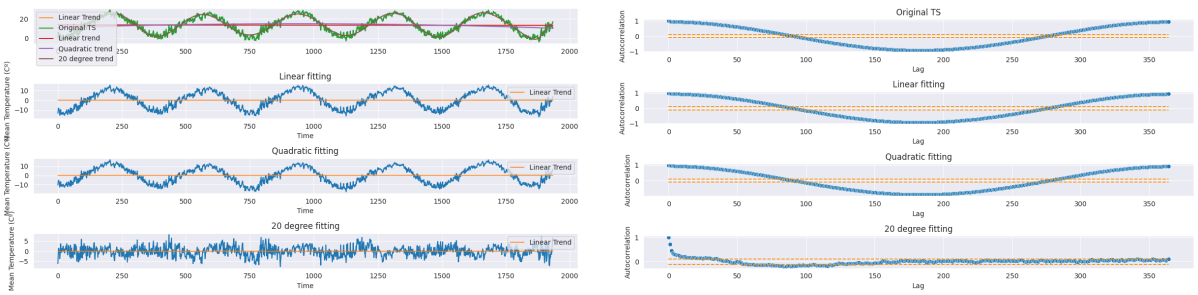
4. Results

This section presents the outcomes of various time series analysis techniques applied to model and understand the trend, seasonality, and noise within the dataset. The focus is on determining the most suitable techniques for achieving stationarity in the time series, ultimately supporting robust forecasting.

4.1. Polynomial Fitting on the series

In this method we tried to fit the trend with different polynomial degrees: 1ª order, 2ª order and 20ª order.

By the way the degree of polynomial increases, the trending tend to overfit the TS. As it is possible to see in figure 2, for a 20ª order we get a stationary TS. This happen because we are removing the trend and the stationary component at the same time. The *ADF* test give us a value of: -8.025942, and we reject the null hypothesis with a 99% confidence degree.



(a) TS with polynoliam fitting

(b) Correlogram

Figure 2: Polynoliam fitting trend estimation

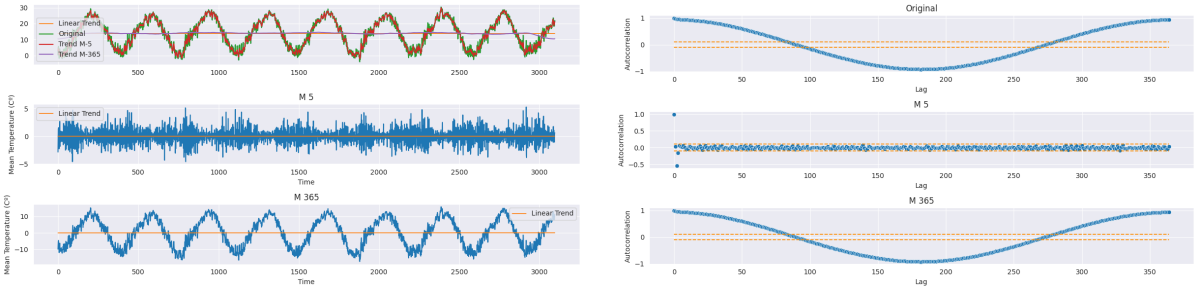
For a 1ª and 2ª order we obtained the values -2.153432 and -2.032724 for *ADF test* respectively, witch means the TS is not stationary. This method is not very useful. It does not give us a stationary TS when we try to remove the trend and to remove the stationary from a very long TS we need to increase the polynomial degree a lot which is very difficult to calculate.

4.2. MA Smoothing

As it is possible to see in figure 3a, for a windows of size $M = 5$, the model overfits the TS. We can say this because the trend has completely adjusted to the seasonality curves, and once removed will originate the erratic component. Well if we look at figure 3b the correlogram says the TS is completely stationary with a *ADF value* of -20.87.

On the other hand, with a M equals to the period of seasonal pattern, $M = 365$, the trend indeed follow the real long trend of TS, and that is why we don't see much of a difference from the original series.

So, if we want to assess the erratic component we can use a $M = 5$ and to assess the seasonal component we can use a $M = 365$.



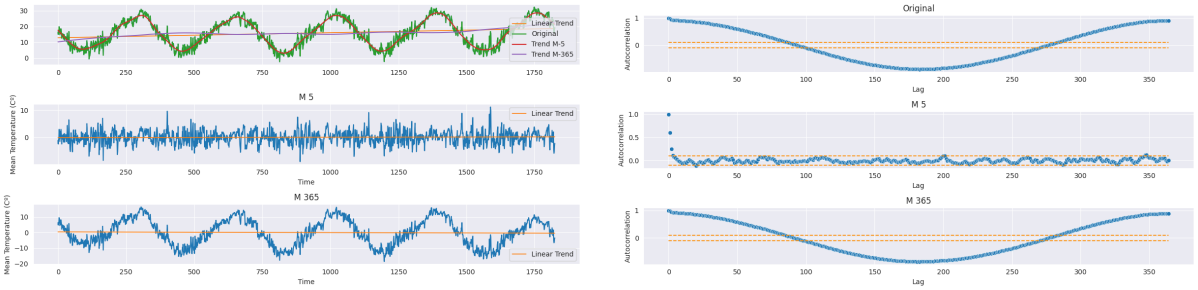
(a) TS with ma smoothing

(b) Correlogram

Figure 3: MA Smooth trend estimation

4.3. LOWESS

Our analysis of local smoothing techniques using moving average and LOWESS methods revealed that both methods obtained similar results. Below 4 we present the results for the LOWESS method:



(a) TS with local smoothing

(b) Correlogram

Figure 4: Local Smooth trend estimation

As the span of points used, we experimented with 5, a small window, that resulted in a good smooth but kind of overfitting, not perfectly identifying the main long trend. We also used a span of 365, as the yearly seasonality is present, where the trend is identified more precisely.

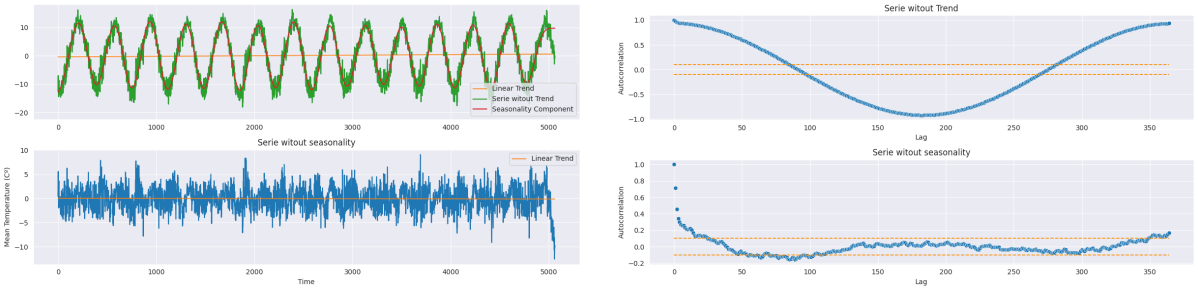
The *ADF statistic* for the the window of size 5 is: -9.94 and if you look to the correlograms, the *MA Smoothing* correlogram has more points closer to zero than *Local Smoothing Method*.

In short, we can say *MA Smoothing Method* might work better than *Local Smoothing Method*.

4.4. Filtering

In order to remove the seasonality we applied a low-pass filter to remove the frequencies on TS, we previously used the *MA Smoothing Method* with $M = 365$ to remove the trend. This method was chosen because it is the one that has the lowest *ADF values* in the end of the experiment.

In figure 5a we can see that the seasonality component fits the TS. The correlogram is not perfectly close to zero, nonetheless we can say the TS is stationary with a confidence bound of 99% for a *ADF value* of -12.03.



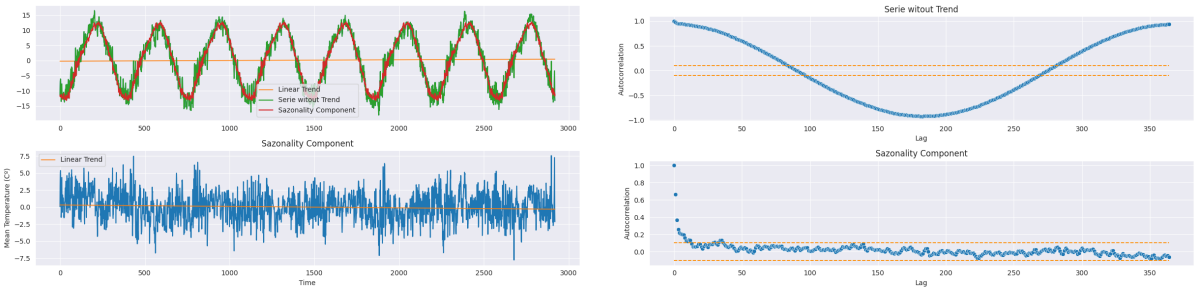
(a) TS with Filtering method

(b) Correlogram

Figure 5: Filtering seasonality estimation

4.5. Epoch Averaging

We also assessed the seasonality using epoch averaging in the same conditions, achieving a similar outcome. In this method, we defined the seasonal cycle to be 365 days, as the yearly periodicity of the data was evident. By averaging the values over each cycle, we obtained an estimate of the seasonality, which we then removed to isolate the erratic component of the time series. We can see these results in figure 6a.



(a) TS with epoch averaging method

(b) Correlogram

Figure 6: epoch averaging seasonality estimation

Similarly to filtering, the correlogram is not perfectly close to zero, but we can say that the TS is stationary with a confidence bound of 99% for a *ADF value* of −14.46.

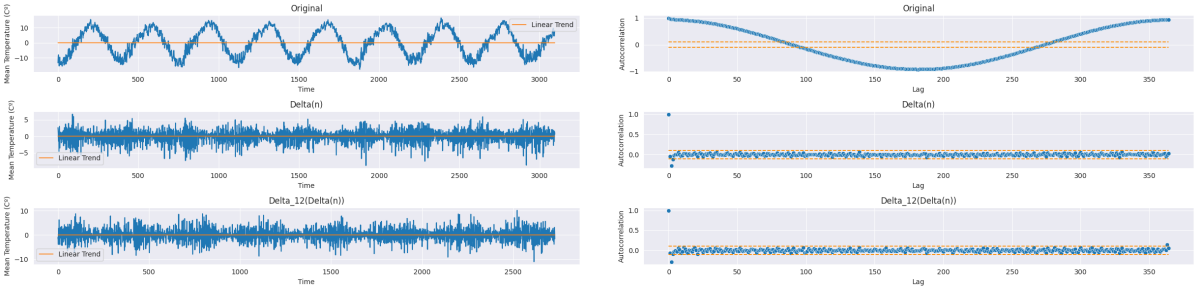
The low-pass filter approach provided a smoother result, while the epoch averaging method fitted a bit more. In the end both retained the same overall outline.

The remaining irregular fluctuations, in bottom figures 6a and previous 5a, represent the random noise and short-term variations, now more visible without much trend and seasonality influence.

4.6. Trend and Seasonality removal by differencing

To apply differencing, we first took a one-step difference, followed by a second difference with a lag equal to the frequency rate, based on the previously differenced values.

As you can see in figure 7a, when we applied the first order differencing, it was expected by us to remove the trend but at the same time we removed the seasonality too. This behaviour already happened before and occurred again. In the correlogram you can see the TS is stationary with a *ADF value* of −7.49.



(a) TS with differencing method

(b) Correlogram

Figure 7: Filtering trend and seasonality estimation

We notice is this method outputted a similar result to *MA smoothing method* for a $M = 5$. This maybe happened because the samples are too closed one each others and the doing the difference result in a similar value for the mean of the neighbours.

5. Discussion & Conclusion

The project demonstrates that various time series analysis methods can effectively achieve stationarity and isolate the primary components of the dataset. Among the methods, Moving Average (MA) smoothing and low-pass filtering with larger windows proved most effective for long-term trend isolation and seasonal decomposition. While polynomial fitting and differencing achieved stationary results, their computational demands and the risk of overfitting render them less practical for lengthy time series. Epoch averaging provided a useful confirmation of seasonal patterns but did not outperform filtering in terms of smoothness and residual control.