**Instituto Superior de Economia e Gestão**

**A red and white logo

Description automatically generated with medium confidence**

**Forecasting Management**

Lisbon Rainfall Forecasting Using Timeseries

2nd Semester of 2022/2023

**Professor:** Dr. Nuno Sobreira

**Authors:**  Khaja Zaffer 58409

May 2023

**LISBON RAINFALL FORECASTING USING TIMESERIES**

As an aspiring data scientist, it is essential to comprehend the methodology and application of statistical analysis to time-series data. R is a comprehensive programming environment renowned for statistical computing and graphics capabilities, rendering it an optimal choice for such analysis. This project exemplifies a beginner's approach to understanding time-series forecasting of rainfall data in Lisbon, Portugal, employing various R packages.

Forecasting time series data involves predicting future values based on historical observations. Specifically, this project focuses on forecasting rainfall data in Lisbon, a city known for its mild, rainy winters and dry summers. Understanding rainfall patterns and accurately forecasting them is critical to the region's agricultural planning, water resource management, and disaster risk mitigation efforts.

A brief step-by-step overview of the time-series forecasting process using R is as follows:

**Data Loading:** The first step involves importing the rainfall data into R. This data, usually in a machine-readable format such as CSV or Excel, consists of historical rainfall measurements for Lisbon.

**Exploratory Data Analysis (EDA):** This step involves a deep dive into the dataset. Utilizing time plots, autocorrelation plots, and decomposition techniques, the analysis identifies underlying patterns and trends in the rainfall data, such as seasonality and long-term trends.

**Data Preprocessing:** To prepare the data for modeling, preprocessing techniques are implemented. This step often includes handling missing values, data smoothing to reduce noise, and applying transformations to make the data more conducive to modeling. Imputation, moving averages, and differencing are common preprocessing techniques.

**Training and Test Sets Creation:** The preprocessed data is split into training and test datasets. The training dataset is used to fit the forecasting model, and the test set is used to evaluate its accuracy and performance.

**Model Selection**: R provides an array of models for time-series forecasting, including ARIMA (AutoRegressive Integrated Moving Average), exponential smoothing, and dynamic regression models. The selection of a model depends on the characteristics of the data and the specific forecasting problem at hand.

**Model Fitting:** The selected model is fitted to the training data, resulting in estimated model parameters.

**Model Evaluation:** The accuracy of the fitted model is evaluated using the test dataset. Key performance indicators for model evaluation include mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE).

**Forecasting:** Upon obtaining a satisfactorily fitted model, it is employed to forecast future rainfall patterns in Lisbon.

**Visualization:** The forecasts are visualized using R's comprehensive plotting functions, facilitating results interpretation and communication.

This project emphasizes the versatility of R for time-series forecasting, showcasing its numerous inbuilt packages for an efficient and detailed statistical analysis of rainfall data. Through these insights, stakeholders in Lisbon, Portugal can better prepare for potential impacts of rainfall patterns, thereby enhancing the region's resilience and planning efforts.

**Data Source**

The Prediction of Worldwide Energy Resources (POWER) initiative offers meteorological information derived from NASA's research efforts, aiding the advancement of renewable energy, enhancing building energy efficiency, and providing agricultural support. NASA maintains an Earth Science research program that employs satellite systems to gather vital data for climate study and understanding climate phenomena. This collection of data includes long-term climate averages and metrics on surface solar energy discharges.

**Data Collection**

**Data is sourced from the POWER Data Access Viewer. This meteorological data from POWER comprises predictions or observations generated by NASA's GMAO MERRA-2 assimilation model.**

**The data, collected on a monthly basis, is specific to a particular geographic location in Lisbon, spanning from the year 2000 to 2021. The dataset includes variables such as:**

**• Specific Humidity: It's a measure of the actual amount of water vapor in a particular sample of air.**

**• Relative Humidity: This is a measure of the amount of moisture in the air compared to the maximum amount the air could hold at that temperature.**

**• Temperature: This records the atmospheric temperature at the specific location.**

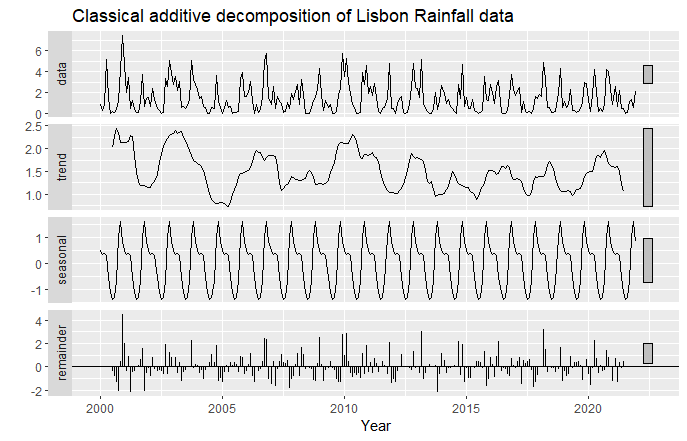
**• Precipitation: This data records the monthly cumulative rainfall at the location.**

A picture containing text, screenshot, plot, diagram

Description automatically generated

* Plot the time series data from the year 2000 to 2021
* **Classical Decomposition of data**

A time series data is composed of

* + Trend: Trend is the overall direction of the data.
  + Seasonlity: Seasonality is a periodic component which repeats itself within a particulat time period.
  + Residuals: the residual is what’s left over when the trend and seasonality have been removed. Residuals are random fluctuation
* 

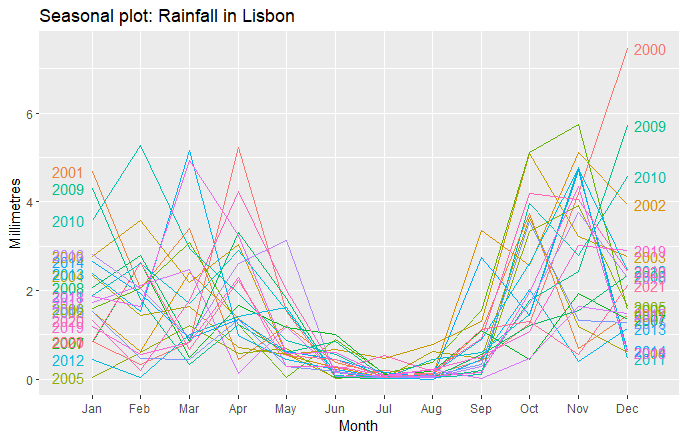
Null Hypothesis: Data is Stationary

Alternative Hypothesis: Data is not stationary

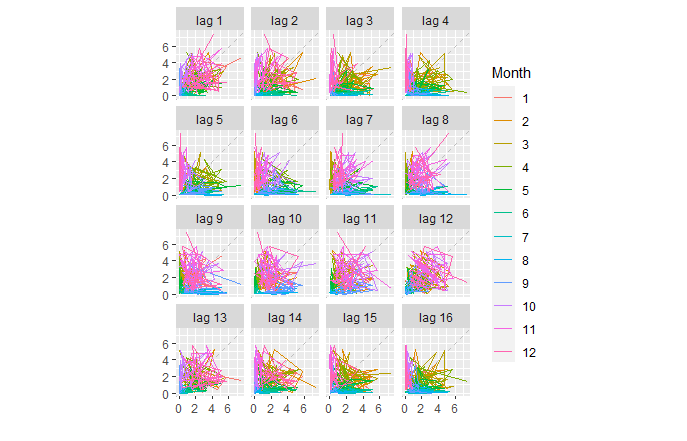
From KPSS test we got p-value = 0.14 and P-value =0.1 which means:

We fail to reject null hypothesis.

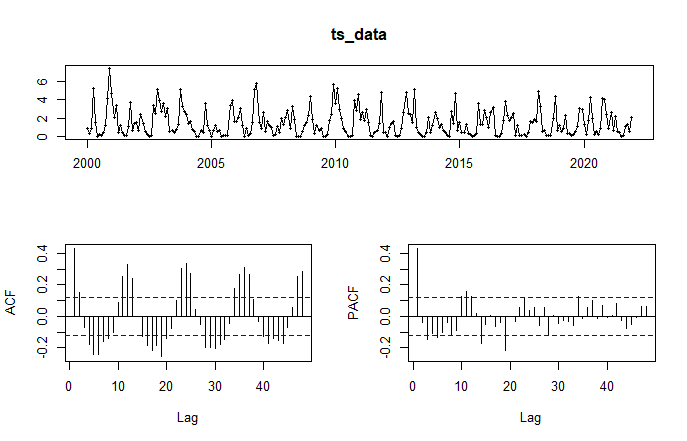
* To check the seasonality we plot:



* Now to see the lag: we can see there is good correlation at lag 16 which is expected also.



* ACF and PACF plot:



Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are important tools for analyzing time series data. These plots help to identify the presence of autocorrelation and provide insights into the order of the Autoregressive (AR) and Moving Average (MA) components of a time series.

Exponential Smoothing

* Exponential smoothing is generally used to make short-term forecasts
* More recent observations are given larger weights by exponential smoothing
* Exponential Smoothing Methods
  + Simple or Single Exponential Smoothing
  + Double Exponential Smoothing
  + Triple Exponential Smoothing

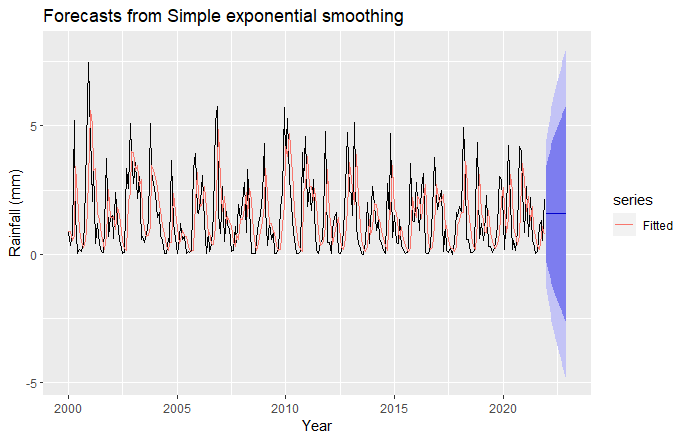
**Single Exponential Smoothing**

* If the data has no trend and no seasonal pattern, then this method of forecasting the time series is essentially used.
* This method uses weighted moving averages with exponentially decreasing weights.

ME RMSE MAE MPE MAPE MASE ACF1 Theil's U

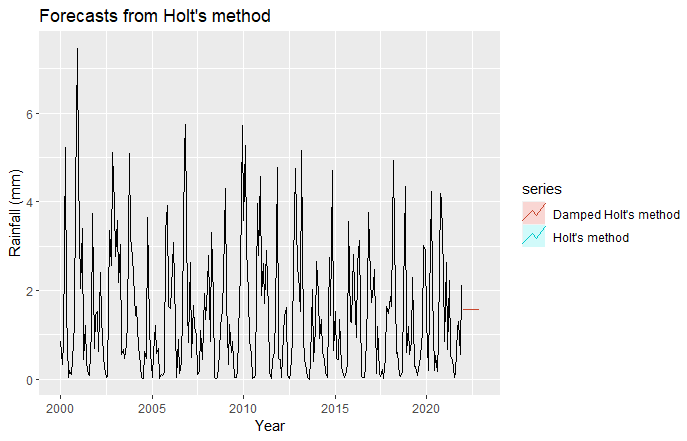
Training set -0.004011745 1.2580398 0.9393768 -Inf Inf 0.8301132 -0.01059559 NA

Test set -0.481882477 0.9146883 0.7555315 -288.1718 298.731 0.6676519 -0.28995556 1.798872



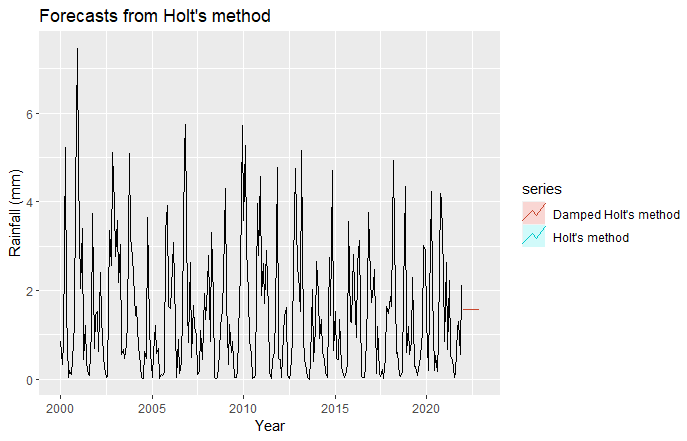
**Double Exponential Smoothing**

* This method is also called as Holt’s trend corrected or second-order exponential smoothing.
* This method is used for forecasting the time series when the data has a linear trend and no seasonal pattern.
* It introduces a new smoothing factor beta that addresses trend in data

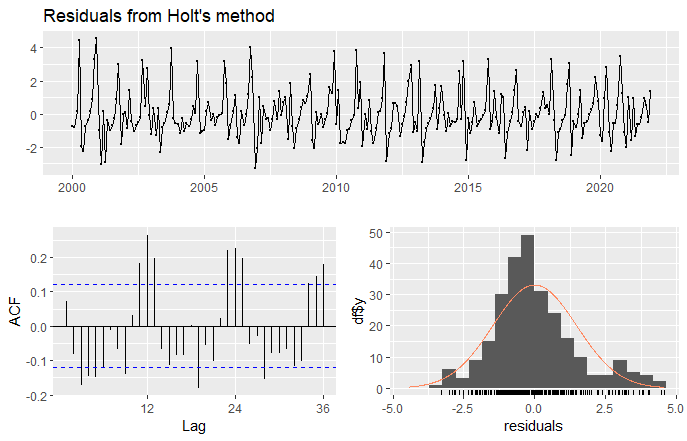


**Triple Exponential Smoothing**

* In this method, exponential smoothing applied three times.
* This method is used for forecasting the time series when the data has both linear trend and seasonal pattern.
* This method is also called **Holt-Winters exponential smoothing.**
* It introduces a new smoothing factor γ that addresses seasonality in data



Residual Plot for HW method

 There are only couple of spikes outside the upper and lower bound range in ACF plot, that too after a significant lag.

Which mean there is no residual error.

**ETS Model:**

the ETS models are an important addition to the time series forecasting toolbox. Introduced by Rob J. Hyndman and Anne B. Koehler, ETS (Error, Trend, Seasonality) models extend the capability of traditional forecasting models, making them a popular choice among data scientists and statisticians.

In ETS models, a time series is decomposed into three components: Error, Trend, and Seasonality, as the acronym suggests.

Error (E): This component represents the random, irregular fluctuations in the data, or the residuals, which cannot be attributed to the trend or seasonal components. These residuals are the "noise" in our data that doesn't follow any predictable pattern.

Trend (T): This component signifies the overall direction in which the time series data is moving. Trends can be upward, indicating that the data is increasing over time; downward, showing a decrease over time; or it can be stable, indicating no change over time. The trend component does not account for fluctuations due to seasonality.

Seasonality (S): This component represents repetitive fluctuations in the time series data that occur within a specific time frame, such as daily, monthly, or annually. These are predictable changes that recur over a certain period.

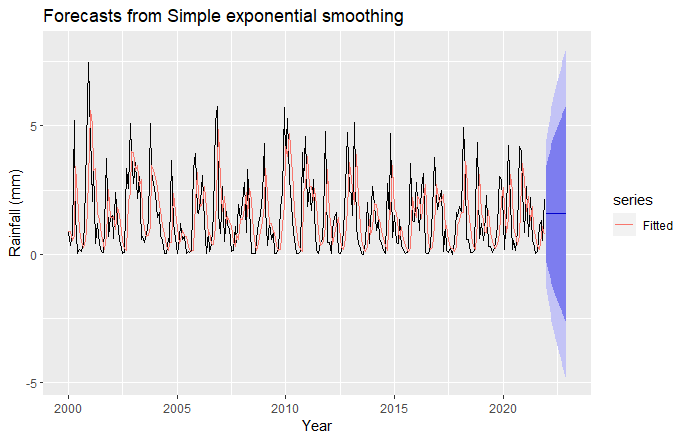
ETS models are highly flexible and can handle various combinations of these components, making them suitable for a wide range of time series patterns. For example, an ETS model could be multiplicative or additive in its error, trend, or seasonality components, and the trend could be either damped or not. The model parameters are typically estimated using maximum likelihood estimation.

In the context of forecasting rainfall in Lisbon, Portugal, the ETS model might be particularly useful. Rainfall often shows clear seasonality (e.g., more rain in winter than summer), and there may also be a long-term trend due to climate change. The error term would capture the random fluctuations in rainfall not explained by these components. By fitting an ETS model to the historical rainfall data, we can provide accurate forecasts of future rainfall

ME RMSE MAE MPE

Training set -0.07047845 1.1238801 0.8091575

Test set -0.39205034 0.9735604 0.6960504



**SARIMA (Seasonal Autoregressive Integrated Moving Average)** :

The Seasonal Autoregressive Integrated Moving Average (SARIMA) is an extension of the ARIMA model, designed specifically for handling time series data that showcases both trending and seasonal characteristics. This powerful forecasting model decomposes the data into distinct elements to better capture its underlying patterns.

SARIMA is specified through three core components:

**Seasonality (S):** This component captures the repeating patterns within the data over a specific time frame, such as a year or a quarter. It's described by the notation (P, D, Q) x s, where 'P', 'D', and 'Q' represent the orders of the seasonal autoregressive, differencing, and moving average parts, respectively. 's' denotes the length of the seasonal cycle.

**Non-Seasonal Part (ARIMA):** This component captures the trends and patterns in the data that aren't seasonally recurring. It's represented by the notation (p, d, q), where 'p', 'd', and 'q' stand for the orders of the autoregressive, integrated, and moving average parts, respectively.

**Trend Component:** This is an essential element that describes the overall direction or trend in the data over time. This can be incorporated into either the seasonal or non-seasonal parts of the model, depending on the specificities of the data.

In essence, the SARIMA model offers a comprehensive approach to understanding time series data, taking into account the various facets of trend and seasonality, thereby producing more accurate and reliable forecasts. This model's flexibility and robustness make it a popular choice among data analysts and forecasters for dealing with time series data.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Series: train\_data

ARIMA(1,0,0)(2,0,0)[12] with non-zero mean

Coefficients:

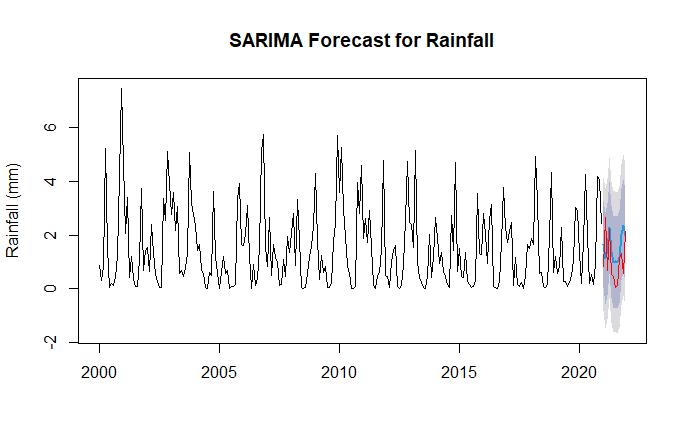
ar1 sar1 sar2 mean

0.3177 0.2165 0.2000 1.5197

s.e. 0.0647 0.0666 0.0704 0.1895

sigma^2 = 1.608: log likelihood = -416.42

AIC=842.84 AICc=843.08 BIC=860.49



**Model Non\_seasonality\_order Seasonal\_order RMSE AIC**

**sarima1 1,0,1 1 ,1,1 0.9978773 780.2243**

**sarima2 1,0,0 1,1,2 0.9767126 779.7006**

**sarima3 1,0,1 1,1,2 0.9752486 781.0603**

**sarima4 2,0,0 1,1,1 0.9976747 780.1698**

**sarima5 2,0,0 1,1,2 0.9839597 781.0675**

**sarima6 1,0,0 0,1,1 1.0059258 777.0293**

**auto.arima 1,0,2 0,12,0,0 0.9146883 842.8392**

**sarima1 1,0,1 1,1,1 0.9978773 780.2243**

**sarima2 1,0,0 1,1,2 0.9767126 779.7006**

**sarima3 1,0,1 1,1,2 0.9752486 781.0603**

**sarima4 2,0,0 1,1,1 0.9976747 780.1698**

**sarima5 2,0,0 1,1,2 0.9839597 781.0675**

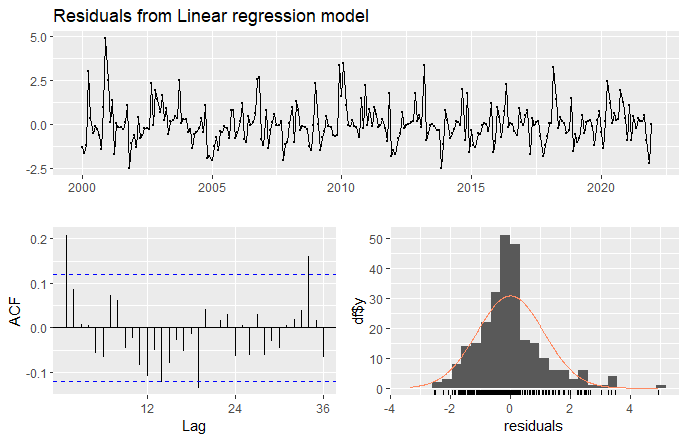
**sarima6 1,0,0 0,1,1 1.0059258 777.0293**

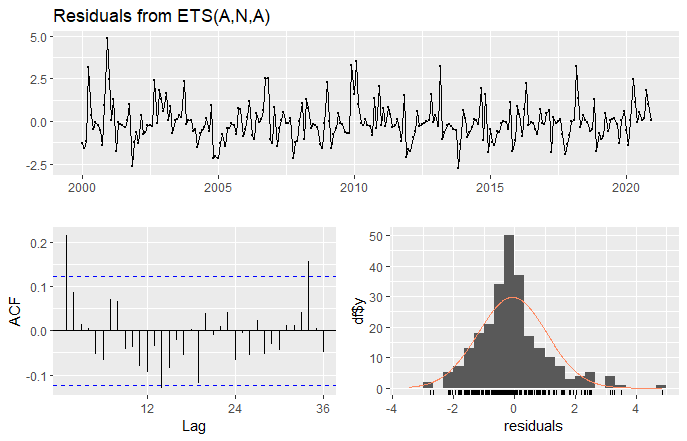
**auto.arima 1,0,2 0,12,0,0 0.9146883 842.8392**

**Linear regression time series forecasting**:

It is a method of using historical data to make predictions about future values of a time series. It involves fitting a linear regression model to the historical data and then using that model to make predictions about future values.

The basic idea behind linear regression time series forecasting is to use past observations of a variable to predict future values of that variable. This is done by fitting a linear regression model to the historical data, where the independent variable is time and the dependent variable is the variable being forecasted. The resulting model can then be used to predict future values of the dependent variable.



* Checking the residual plot for our best fitted model
* 
* The plot shows almost no error as the spike is after significant lag of 1
* Forecast for next 12 periods.

