UNIVERSITÀ CATTOLICA DEL SACRO CUORE

Campus of Milan

Faculty of Economics

Undergraduate programme in Economics and Management

An application of Smart Water Analytics for predicting water availability

Supervisor

Professor Marco De Ieso

Dissertation by

Chiara Lentsch

Id Number: 4901570

Academic Year 2021/2022

Table of contents

Introduction

The water supply industry is facing increasing challenges related to the aging of infrastructures, increasing population, environmental changes, and shortages of skilled workforce. All these aspects contribute to raising costs for utility companies. In recent years the advent of Smart Water Management gave the water sector the chance to use water resources in a more efficient and optimized way. Smart Water Management, in fact makes use of Big Data, IoT and analytics to solve water management problems through automation, data collection and data analysis.

Acea Group, a multiutility operator present in central Italy, embraced Smart Water Management by implementing mathematical models to efficiently manage its wide water network. This allows the firm to have a predictive maintenance approach and real time information on the network. Acea wants to broaden the use of Smart Water technologies and especially Smart Water Analytics also for waterbodies. In fact, given the increasing scarcity of water resources, it is becoming more and more essential to predict future availability to ensure that the firm can supply water in the next future. To address this issue Acea launched a competition on Kaggle (https://www.kaggle.com/c/acea-water-prediction) making available several datasets displaying a set of parameters for each waterbody. This paper will focus on Petrignano aquifer.

A possible solution to the problem which we will analyze in this thesis is the ARIMA (Autoregressive Integrated Moving Average) model by a Box & Jenkins, a forecasting algorithm for univariate time series. The objective of this model is to predict the future availability of the aquifer expressed by depth to groundwater.

The topic has been chosen to explore the world of Data Science, especially the time series and forecasting field. Also, it has been a way to apply some of the theoretical concepts covered along the course of Applied Statistics. The Acea competition allowed me to face and solve for the first time a real business problem. A great part of the knowledge required for this thesis has been acquired from personal research with the aid of my supervisor.

In the first chapter we will present the concept of Smart Water Management and the way it improved the water sector focusing on the case of Acea. The second chapter will be of more theoretical base and will display all the fundamental theories and notions behind time series. The third and last chapter is the core of the paper which describes all the steps preceding the construction of an ARIMA model, from problem framing and data understanding, up to the fitting of the model and the evaluation of its performance.

CHAPTER 1 – Smart Water Management

Water is a vital but scarce resource, and its management is a critical issue. Continuous ICT (Information Communication Technology) achievements allow to create smart solutions and applications for society's needs. In the water industry, these innovations are deployed to provide an alternative way to improve water management and its efficiency.

1.1 The Blue Gold

Water covers 70% of our planet but freshwater – the one we use in our daily life to bath, cook, irrigate and drink – is extremely rare. It represents only 3% of the world's water and only one third of that is available for use. As a result, according to the UN, 1.1 billion people worldwide lack access to water, and a total of 2.7 billion find water scarce for at least one month of the year.

Many of the water sources that keep the ecosystems flourishing and supply the growing human population are threatened. Rivers, lakes, and other water sources are disappearing or becoming too polluted to use. Climate changes exacerbate the scenario even more. Rainfalls are reduced, absent or excessively strong causing floods and glaciers are receding. At the same time water consumption is increasing due to the rising demand for agricultural and industrial products with an intensive use of water.

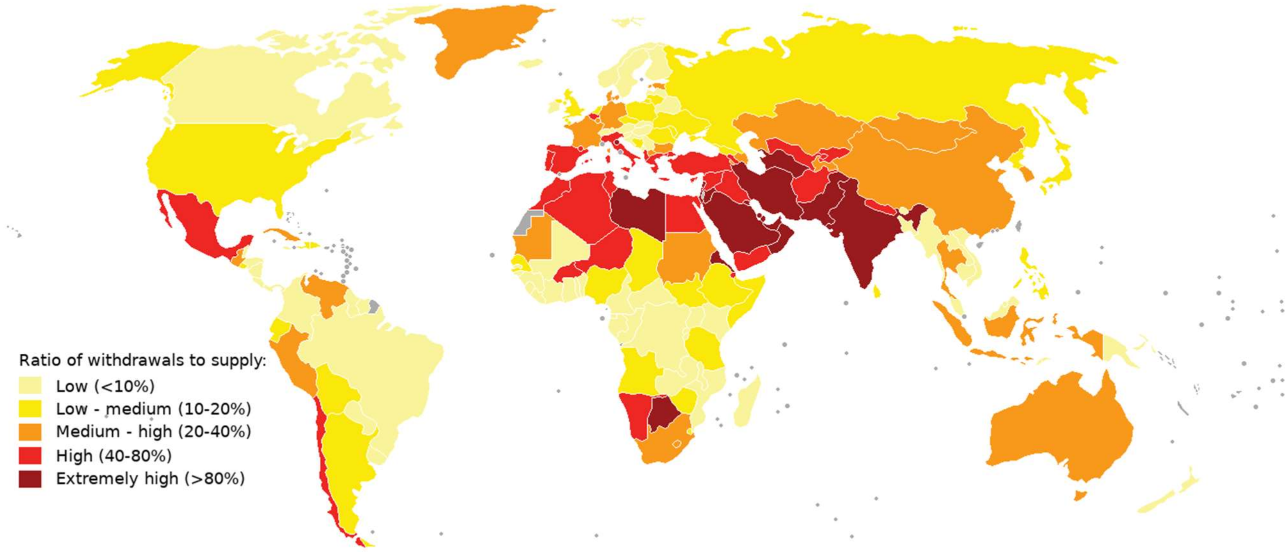


Figure 1.1. Water stress (ratio of water use relative to water availability) per country in 2019. Source: World Resources Institute

At the current pace the situation will soon turn into a proper emergency. According to the UN, by 2025 two thirds of the world's population will face water shortages and ecosystems around the world will barely survive. Individual countries and regions need to urgently deal with the problems related to water stress. Water is scarcer and scarcer and must be treated as such by focusing on managing demand and controlling the use of the sources.

1.2 Smart Water

Factors such as climate change, scarcity of water, growing demographics and consequent increase in water demand have pushed water providers to move towards a more effective and sustainable approach for the management and use of water in the residential sector. As a matter of fact, the water industry is constantly evolving and making use of available tools and technology to adapt to the new challenges posed by water scarcity.

The concept of Smart Water refers to a trend in the water sector that consists in involving emerging technologies (including hardware, software, and analytics) to solve water management problems through automation, data collection and data analysis. By means of ICT, it is now straightforward to trace water resources and construct precise models of availability and meteorological predictions.

According to US Water Alliance Smart Water addresses several challenges faced by municipalities such as the maintenance of the infrastructures, the detection of leaks in the network, the reduction of wastes, and the control of contaminants.

Water utilities as well as hardware and software companies are collaborating on Smart Water innovation since the collection and analysis of data proved to be effective for the optimization of water management. For example, Kamstrup – a world-leading supplier of energy and water metering solutions – cooperated with the IoT service provider Sigfox to develop a high-precision ultrasonic metering technology provided with remote reading and cutting-edge communication.

Smart Water technologies are mainly developed on Big Data, Internet of Things, and analytics. For example, Big Data enables the water sector to improve the detection of filtrations and leakages, monitor water use and obtain data on treatment processes. Furthermore, thanks to IoT communication it is possible to carefully monitor water information and optimize water resource use and management.

Smart Water Management (SWM) is the activity of planning, developing, distributing, and managing the use of water resources using Smart Water technologies to make more reasonable and sustainable the use of these water resources. SWM is becoming a concept of increasing interest as governments integrate smart principles into their urban, regional, and national strategies.

1.3 Practical Case: Acea Group

Acea is an Italian multiutility operator committed to improving resident's lives by offering high-quality everyday essential services in the water, energy distribution, production and sale, lightning, and waste treatment sectors.

1.3.1 Chronicle

Acea was founded in 1909 as a Municipal Electric Company of the Municipality of Rome to supply energy for public and private lightning. Soon afterwards the Rome governor assigned to the firm also the management of the municipal aqueducts and the construction and management of the Peschiera Capore aqueduct.

In 1953 the city Council approved Acea's plan to make the city electrically self-sufficient, improve the resident's water system and expand in the city's outer suburbs. This implied new electricity stations and receiver substations, waterpower plants, the completion of the Peschiera Capore aqueduct, the search for new aquifers and the construction of other aqueducts. In the following decades the company continuously evolved by taking over the management of the capital's wastewater treatment and by being admitted to listing in the Italian Stock Exchange.

In the early 2000s Acea went global thanks to the construction and concession of a water plant in Lima. Furthermore, the firm obtained the management of the integrated water service in multiple areas of Campania and Tuscany.

In 2017 Acea faced one of its most serious water crises. Its hard work in carrying out extraordinary maintenance on the water network guaranteed service continuity for citizens. Also, thanks to campaigns raising awareness on responsible use of water the company became a strong promoter of sustainability.

As established in the most recent business plan, the latest years marked a strong boost for infrastructural investments in both the water and the electricity sector. The aim is to develop resilient technology and innovation, with special attention to sustainability for both the environment and people.

In 2018 Acea and Open Fiber entered into an agreement to develop an ultra-broadband communication network in the city of Rome and the following year it entered the sector of gas distribution. Acea is currently working on the second line of the Peschiera Capore aqueduct, which will secure the water needs for the capital.

1.3.2 Integrated Water Management

Acea is the leading Italian operator in the sector of integrated water services for the number of residents served. It manages the entire water cycle from the spring to wastewater treatment for 9 million residents in five Italian regions – Lazio, Tuscany, Umbria, Molise, and Campania – through subsidiary companies (Acea Group, https://www.gruppo.acea.it).

As defined by the Global Water Partnership (GWP) the integrated management of water sources is "a process which promotes the coordinated development and management of water, land, and related resources, in order to maximize the resultant economic and social welfare in an equitable manner without compromising the sustainability of vital ecosystems". To concretely implement this process Acea inspects the quality of water throughout its journey, from catchment to distribution, with strong commitment to sustainability and preservation of the environment. All this is possible thanks to an efficient management of infrastructures, process digitalization and the operators' competence.

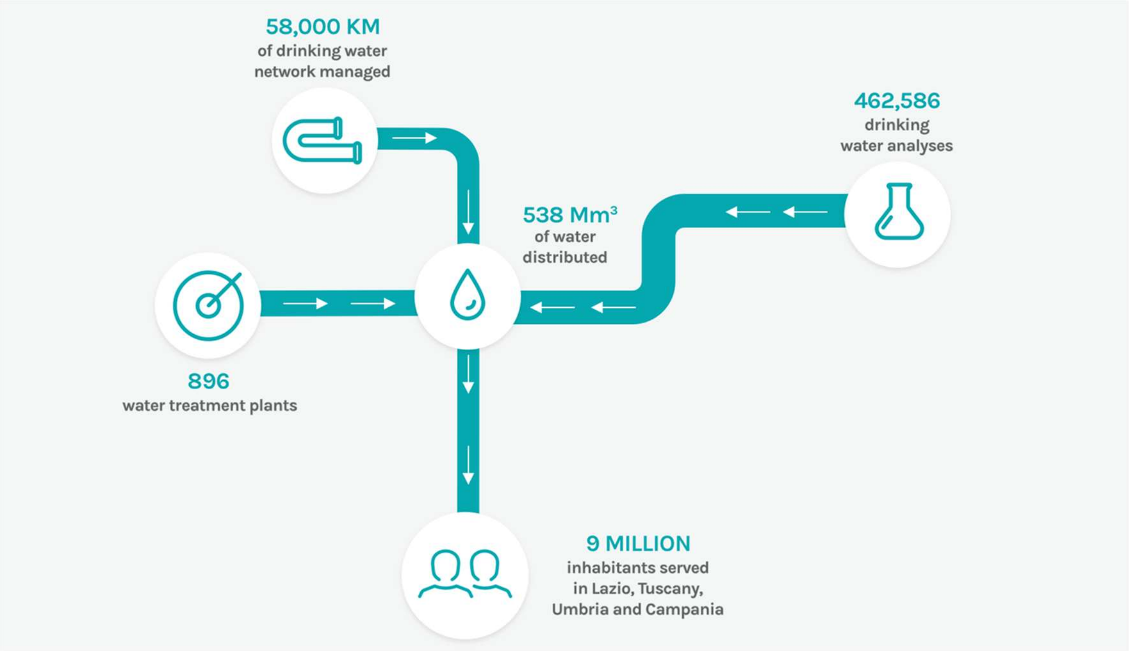


Figure 1.2. Source: Acea 2019 Sustainability Report and Consolidated Report

The waterbodies from which Acea draws water are placed in unpolluted areas and the firm takes care of the natural environment surrounding them. Actually, numerous areas are subject to full protection, such as the Capore springs (997,848 m^2) and the Peschiera springs (375,322 m^2). This allows Acea to guarantee the quality and integrity of the water resources from their origin. The resource catchment is the initial and crucial part of the supply chain during which water is taken from the waterbodies. In this context intervention plans are aimed not only at actions to optimize, detect and repair eventual leaks, but also at turning unused sources into operation and at securing existing springs.

The following step in the water journey is the distribution via an extensive network of aqueducts and pipelines. Most of the aqueduct network is managed thorough a remote-control system that provides information on the state of the network to plan maintenance, interventions, and upgrades. Acea aims at further improving the efficiency and quality of the service by installing smart water meters, reducing network leaks, rationalizing treatment plants, and dividing the network into districts.

The final phase is the sewage network and treatment system that consists in collecting wastewater via the sewage pipes and send it to the treatment plants to be turned into a new resource for further uses such as irrigation in agriculture. Acea manages these activities according to the concept of circular economy, another concrete application of its commitment to the protection of water resources.

1.3.3 Water Digitalization

The digitalization of water sector – also known as Smart Water, discussed at length in the previous paragraphs – allows to protect and enhance water sources and for this reason Acea Group planned to invest €2.2 million over 2020-2024.

The integration between digital technologies and operating methods applied to supply networks and plants, it is feasible to improve the knowledge on the infrastructure, therefore improving its management and generating operational efficiency. Smart water technologies provide detailed information on the conditions of pipes, structures, and utilities, but also an overview of the whole water system.

The process of digitalization of the water sector starts with the collection and management of a large amount of data collected by smart water meters or through IoT sensors installed on the network – transmitting information such as flow rate or water pressure – or in proximity to waterbodies.

Then, the data collected are processed through mathematical models to have an insight on how complex water supply networks work and to make the management of the water resource increasingly efficient (Smart Water Management). This makes the maintenance preventive and predictive allowing to know in real time the state of the infrastructures and therefore intervene in a targeted and efficient way.

Acea carries out the process of water digitalization through two main techniques: the districtization and the mathematical modelling of the networks. The first one consists in dividing the whole network into smaller homogeneous areas with measured connection points to get a detailed quantitative balance of the water coming in and out from any district. It also allows to study performance variations in time for each area and help intervene promptly in case of leaks or other issues.

Once districts have been mapped, and their water balance monitored in real time, water digitalization is completed with mathematical modelling of the network to digitally reconstruct flows along thousands of kilometers of pipes. This practice permits to understand both the current functioning of the water networks and its potential behavior in case of interventions.

1.3.4 Predicting Water Availability

Italy is the first in Europe for consumption and waste of water and the last for reuse of wastewater. Furthermore, due to climate crisis rainfalls are rare and temperatures have increased by 1.65 degrees compared to the historical average. These factors pose a serious threat to water sources such that most of them are drying out and some lakes even disappeared.

As it is easy to imagine, Italian water supply companies are struggling with the need to forecast the water levels of waterbodies to handle daily consumption. During coldest seasons waterbodies are refilled, but during hotter periods of the year they start to drain. To help preserve the health of these waterbodies it is crucial to predict water availability, in terms of level and water flow for each period of the year.

To address this issue Acea launched a competition on Kaggle for the development of four mathematical models to predict the amount of water in each waterbody (aquifers, water springs, lakes, and rivers). The starting point is a series of distinct datasets containing measures such as rainfall, humidity, temperature, etc.

CHAPTER 2 – Time Series Analysis and ARIMA model

The datasets provided by ACEA for the construction of the models are made of periodical measurements of parameters related to the different waterbodies. These statistics are collected periodically over time and therefore can be mathematically seen as time series data.

This chapter will depict some of the basic theories behind time series analysis mainly extracted from Practical time series analysis by Avishek Pal, PKS Prakash and Time Series Analysis Forecasting and Control by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Greta M. Ljung.

2.1 Time Series

A time series is a sequence of quantitative observations about a system or process made at successive points in time. Commonly, the points in time are equally spaced. Examples of time series data are sales volumes, stock prices, weather attributes collected over a time spread of several years, months, days, hours, etc. The frequency of observation depends on the nature of the variable. For instance, gross domestic product, which is used to measure annual economic growth of a country, is reported every year. Instead, information about stock prices and weather attributes are available every second.

Time series data are typically characterized by a set of internal structures that require special formulation and techniques:

- General trend
- Seasonality
- Cyclical movements
- Unexpected variations

Most time series have one or more of these internal structures so that a time series can be expressed as $x_t = f_t + s_t + c_t + e_t$ which is a sum of the trend, seasonal, cyclical, and unexpected components in that order. Subscript t is the time index at which observations have been taken.

2.1.1 General Trend

The general trend occurs when a time series exhibits an upward or downward movement in the long run. A rapid way to detect the presence of a general trend is to plot the time series. However, it might not be that evident over the short run since short run effects such as seasonality and irregular variations can obfuscate the existence of a trend.

The presence of a general trend in a time series is due to fundamental shifts or systemic changes of the process or system it represents. For example, an upward movement of global temperatures can be attributed to global warming due to greenhouse effect.

The general trend can be modeled by setting up the time series as a regression against time and other known factors as explanatory variables. The regression or trend line can then be used as a prediction of the long run movement of time series. Residuals left by the trend line are explained by other properties like seasonality, cyclical behavior, and irregular variations.

2.1.2 Seasonality

Seasonality in time series data refers to a predictable pattern that occurs at a regular interval within one year. The easiest example of seasonality are temperature data since we always expect to have higher temperatures in summer and lower in winter.

Taking seasonality into consideration is fundamental in time series forecasting. In fact, the model that considers the seasonal effect will be much more accurate in terms of forecasting.

The simplest way to detect seasonality is through exploratory data analysis with the following graphical techniques: run sequence plot, seasonal subseries plot, and multiple boxplots.

2.1.3 Cyclical Changes

Cyclical changes are movements observed periodically over time but occur less frequently than seasonal fluctuations. The average periodicity for cyclical changes is usually in years, while the seasonal variations are observed within the same year and correspond to annual divisions of time such as seasons, quarters, and periods of holidays.

Time series related to economics often shows cyclical changes that correspond to macroeconomics cycles like periods of recessions followed by periods of boom within a few years of time span.

To identify cyclical changes, it may be useful to display a long run plot of the time series. They manifest through repetitive crests and troughs every few years. Therefore, identifying cyclical movements might require data that dates significantly back in the past.

2.1.4 Unexpected Variations

Referring to the model that expresses the time series as a sum of four components, it is important to note that after being able to account for the first three components we might be left with an irreducible error component that is random and does not exhibit dependency on the time index. This fourth element represents unexpected variations in the time series.

Unexpected variations are stochastic and cannot be framed in a mathematical model for future prediction. The presence of this component is due to lack of information on explanatory variables or to the presence of random noise

2.2 Time Series Analysis

Time series analysis aims at developing a mathematical model that can explain the behavior of a time series and forecast its future state. The model should be able to account for one or more of the internal components that might be present.

Time series analysis is applied in many fields like weather forecasting, finance, communication systems etc. It helps businesses both to understand the causes of trends or systemic patterns over time and to predict the likelihood of future events.

In the following section we will discuss two general models considered the building blocks of time series analysis.

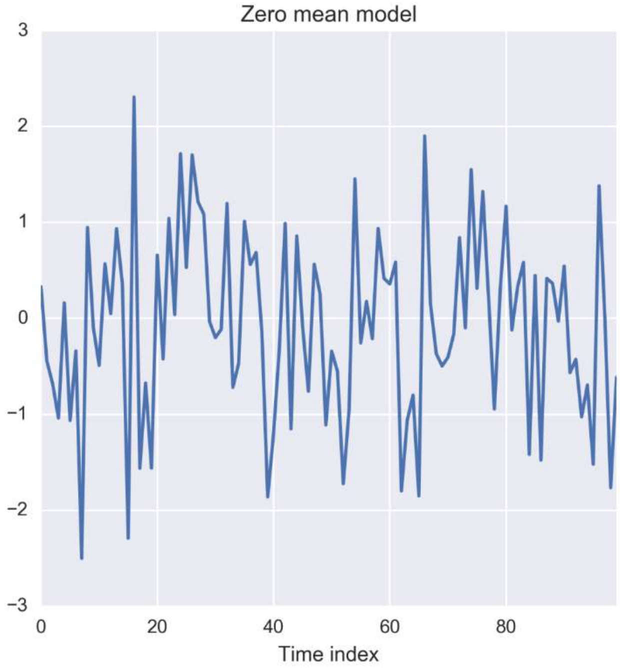
2.2.1 Zero-Mean Model

The zero-mean model – aka white noise – has constant mean and variance and shows no predictable trend or seasonality. Observations from this model are assumed to be independent and identically distributed (iid) and represent a random noise around a fixed mean, deducted from the time series as a constant term.

If we consider $X_1, X_2, \dots, X_n$ as the random variables corresponding to n observations of a zero-mean model and $x_1, x_2, \dots, x_n$ as n observations from the zero-men time series, then the joint distribution of the observations is give as a product of probability mass function for every time index as follows:

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = f(X_1 = x_1)f(X_2 = x_2) \dots f(X_n = x_n)$$

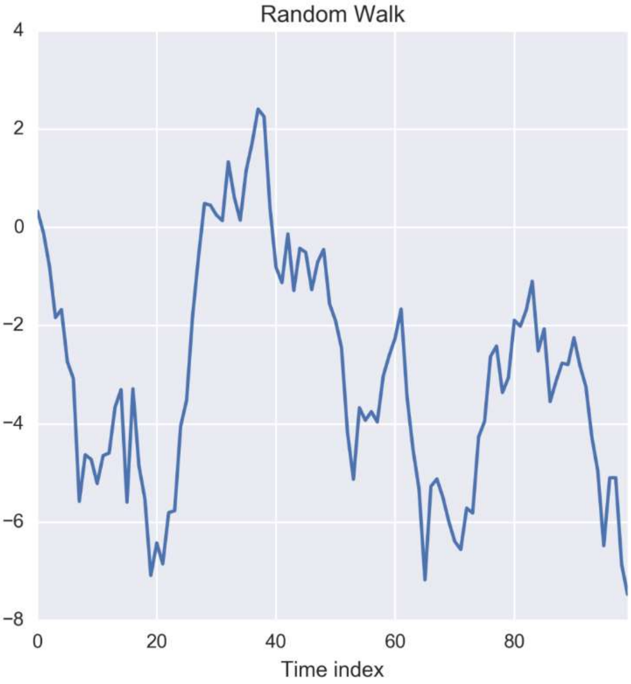
Commonly $f(X_t = x_t)$ is modeled by a normal distribution of mean zero and variance σ^2, which is assumed to be an irreducible error of the model and therefore treated like a random noise.



Graph 2.1. Zero-mean series of a normally distributed random noise of unit variance. Source: Practical Time Series Analysis

2.2.2 Random Walk

A random walk is a sum of n independent and identically distributed random variables with zero mean and constant variance. The realization of a random walk at time t is given by the sum $S = x_1 + x_2 + \ldots + x_n$.



Graph 2.2. Random walk obtained from iids varying according to a normal distribution of zero mean and unit variance

The random walk is key in time series analysis because if such behavior is found in a time series, it can be easily reduced to a zero-mean model by differencing the observations from two consecutive time indices. We will see this further when we will delineate the concept of differencing.

2.3 ARIMA model

"AutoRegressive Integrated Moving Average" is a forecasting algorithm based on the assumption that the information in the past values of the time series can be used to predict future values. It essentially creates a linear equation which describes and forecasts the time series data.

The ARIMA model is made up by three separate parts:

- AR (autoregression)
- I (integration or differencing)
- MA (moving average)

The model is usually represented as $ARIMA(p, d\,, q)$ where each of the three letters represent a parameter to be provided. Parameter p stands for the number of autoregressive (AR) terms, d determines the order of differencing, and q defines the number of moving average (MA) terms.

It is noteworthy to say that ARIMA model requires stationary time series. Therefore, before introducing all the components of the model, it is useful to discuss the concept of stationarity and the technique of differencing time series.

2.3.1 Stationarity

Stationarity is an assumption which requires the internal structure of the series to not change over time. Therefore, it implies mean, variance and autocorrelation to be invariant with respect to the actual time of observation.

A common example of stationary time series is the zero-mean series which is a collection of samples generated from a normal distribution with mean zero. Observations are assumed to be independent and identically distributed (iid) and represent the random noise around a fixed mean, which has been deducted from the time series as a constant term. The zero-mean series does not show any temporal pattern such as trend and seasonality.

However, most real-life time series are not stationary. Non-stationarity is mostly due to the presence of trend and seasonality that affect mean, variance, and autocorrelation at different points in time. In general, a time series with no predictable patterns in the long run is stationary.

The statistical test for objectively determining whether differencing is required to stationarize a time series are known as unit root tests. Unit root is a characteristic of a time series that makes it non-stationary. Technically, a unit root is said to exist in a time series of value of $\alpha = 1$ in below equation.

$$Y_t = \alpha Y_{t-1} + \beta X_e + \epsilon$$

where Y_t is the value of the time series at time t and X_e is an exogenous variable. The presence of a unit root means that the time series is non-stationary.

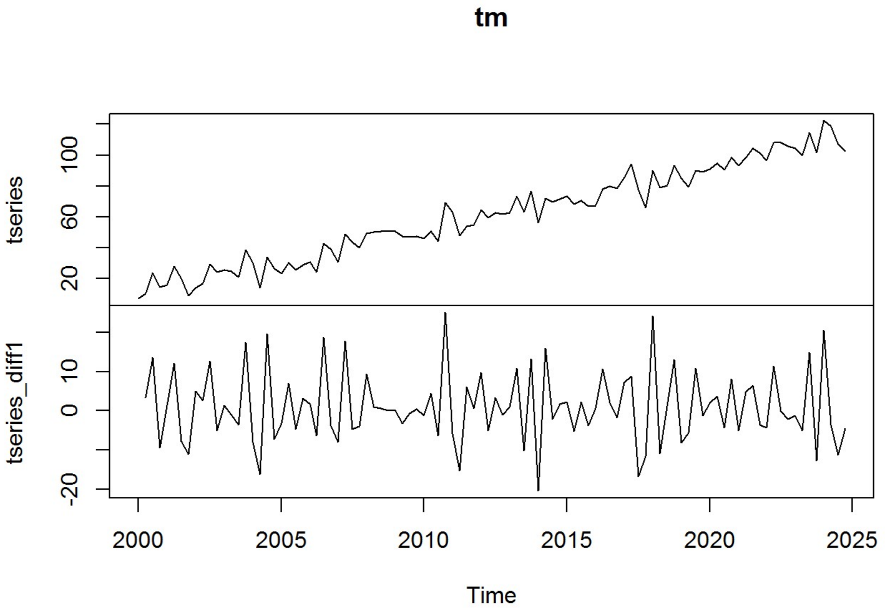
Augmented Dickey Fuller (ADF) test is one of the unit root tests. The null hypothesis (H_0) is presence of unit root or non-stationarity, while the alternative hypothesis (H_1) suggests stationarity of the data or no unit roots. There are two ways to reject the null hypothesis. The first way is having the p-value below a set significance level (default is 5%). The second way for which the null hypothesis can be rejected is if the test statistic (ADF statistic) is less than the critical value.

2.3.2 Differencing

If the data are not stationary but we want to use a model such as ARIMA which requires this characteristic, the data must be transformed. One way to make a non-stationary time series stationary is through differencing.

The basic idea of differencing is taking differences between successive occurrences of the time series $\Delta x_t = x_t - x_{t-1}$ such that Δx_t have constant mean and variance and hence can be treated as a stationary time series.

First-order differencing implies taking differences between successive realizations of the time series so that the differences Δx_t are irregular variations free from any trend or seasonality. The random walk model discussed in the previous section is a sum of subsequent random variations and it is given by $x_t = x_{t-1} + \epsilon_t$ where ϵ_t is a zero mean random number from a normal distribution. Random walks are characterized by long sequences of upward or downward trends which make them non-stationary. However, through first order differencing we can transform the random walk into a zero mean stationary series since the first differences of a random walk is equal to the random noise $\Delta x_t = \epsilon_t$. The transformed time series is denoted as $x_t = x_t - x_{t-1}$ and has $N - 1$ observations with zero mean and constant variance.



Graph 2.3. The differencing effect. Source: Time Series Transformation, Time Series Analysis in R

In some cases, first-order differencing is not enough to stationarize the time series and therefore data are differenced again. Hence, second-order differencing is generated as $x_t = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2})$. The resulting time series has $N - 2$ observations.

When a time series exhibits seasonality of a known time period of m time indices, it can be stationarized by taking seasonal differences between x_t and x_{t-m}.

2.3.3 Integration

The middle term I stands for "Integrated", and it denotes the application of a differencing step to the data. This because differenced values are generally much more stationary than the raw undifferenced ones and when performing time series modeling, we want our variables to be mean variance stationary.

In $ARIMA(p, d, q)$ parameter d – also known as degree of differencing – indicates the number of differences needed for stationarity.

2.3.4 Autoregression (AR) Model

In an autoregressive model the variable of interest is forecasted using a linear combination of its past values. Actually, the term autoregression indicates that it is a linear regression of the data in the current series against one or more past values in the same series.

An autoregressive model of order p can be written as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

where ϵ_t is white noise and c is defined by $c = (1 - \sum_{i=1}^{p} \phi_i)\mu$.

The order of an autoregression is the number of immediately preceding values in the series that are used to predict the value at the present time. Thus, for example, a first order autoregression or $AR(1)$ implies that the outcome variable at some point in time t is related only to time periods that are one period apart $(t - 1)$.

2.3.5 Moving Average (MA) Model

A moving average process states that the current value is linearly dependent on the current and past error terms. The error terms are assumed to be mutually independent and normally distributed like white noise.

A moving-average model of order q can be written as

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

The model expresses the present value as a linear combination of the mean of the series (μ), the present error term (ε), and the past error terms ($\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-q}$). The magnitude of the impact of each past error is expressed by the coefficient theta (θ).

The order q of the model determines the number of past error terms that affect the present value. Thus, for example, a first order moving average or $MA(1)$ implies that the outcome variable at some point in time t is related only to errors of time periods that are one period apart $(t - 1)$.

2.3.6 Autocorrelation Analysis

After the time series has been stationarized by differencing, the next step in fitting an ARIMA model is to determine which AR or MA terms are needed to correct any autocorrelation that remains in the differenced series.

Autocorrelation shows the correlation – or degree of similarity – between a time series and a lagged version of itself. Its value ranges from -1 to 1 and it can be either positive when the increase observed in a time interval leads to a proportionate increase in the lagged time interval, or negative when the increase observed in a time interval leads to a proportionate decrease in the lagged time interval. Autocorrelation can be computed for different time lags. For example, an autocorrelation of lag 1 measures the correlation between observations that are a one-time gap apart.

ACF plot is used to visualize the concept of autocorrelation and so to identify the correlation between the points up to and including the lag unit. Such plot is key in the identification of the terms of ARIMA model that we want to use for our time series. If there is positive autocorrelation at lag 1, we use AR model, if there is negative autocorrelation at lag 1, we use MA model.

After plotting ACF it is useful to move to Partial Autocorrelation Function (PACF). A partial autocorrelation is the correlation that results after removing the effect of any correlations due to the terms at shorter lags. If the PACF plot drops off at lag n, then use an AR(n) model and if the drop in PACF is more gradual then we use the MA term.

2.4 Evaluation Metrics

Measuring the performance of forecasting models is fundamental from both technical and business perspective, especially when business decisions are based on the insights generated by the forecasting model. There are several types of evaluation metrics depending on the model used and the results generated. Each has its pros and cons. We will cover the ones required by ACEA for the competition on Kaggle.

2.4.1 R-squared

The stationary R-squared is used in time series forecasting as a measure that compares the stationary part of the model to a simple mean model. It is defined as

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where SS_{res} is the sum of squared residuals from expected values and SS_{tot} is the sum of squared deviations from the dependent variable’s sample mean. It denotes the proportion of the dependent variable’s variance explained by the independent variable’s variance. A high R^2 value represents that the model’s variance is similar to that of the true values, while a low R^2 value suggests that the two values are not strongly related.

The main cons about R-squared is that it is more a measure of overall fitness than of forecast accuracy. In fact, it does not indicate whether or not the model is capable of making accurate future predictions but instead it shows whether or not the model is a good fit for the observed values as well as how good of a fit it is.

2.4.2 Mean Absolute Error (MAE)

The MAE is defined as the average of the absolute difference between forecasted and true values.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

y_i is the expected value and $x_i$is the actual value. The letter n represents the total number of values in the test set.

The MAE shows how much inaccuracy we should expect from the forecast on average. When $MAE = 0$ the expected values are correct, so the forecast is error-free. Hence, the lower the MAE value, the better the model.

However, MAE does not give an insight regarding the proportional scale of the error, so it can be difficult to distinguish between large and little errors. To solve this issue, MAE could be combined with other measures to see the magnitude of error. Also, MAE might obfuscate issues related to low data volume.

2.4.3 Root Mean Squared Error (RMSE)

RMSE is defined as the square root of mean square error and is an extension of it. MSE is the average of the error squares, and it is used to evaluate the quality of a forecasting model. RMSE formula is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

where Y_i is the actual value of a point for a given time period and $\hat{Y}_i$ denotes the fitted forecast value for the same time period. Letter n refers to the total number of values in the test set.

RMSE is always positive with lower values indicating higher performance. The advantage of this technique is that the result is in the same unit as the projected value which makes it easier to comprehend. However, like MSE, RMSE penalizes greater errors more.

CHAPTER 3 – Acea Smart Water Analytics

3.1 Introduction

This last chapter will be focused on the process of construction of the forecasting model to satisfy Acea's business needs. The issue has been presented through a Kaggle competition named "Acea Smart Water Analytics" which describes the desired output and the rules to be followed.

As for the evaluation submitted notebooks are graded by Acea according to three criteria:

- Methodology/Completeness: appropriacy of the statistical model given the data, number of machine learning models developed, MAE, RMSE
- Presentation: coherent narrative, helpful data visualization, easy to understand and present
- Application: ability to forecast water availability in a time interval of the year, possibility to apply the same methodology to other waterbodies

Further details on the competition can be found here https://www.kaggle.com/c/acea-water-prediction.

We will now go through each step of the model creation process starting from the identification of the business problem up to the actual application of the model and its evaluation.

3.2 Problem Framing

As previously mentioned Acea – as other companies in the water sector – is implementing smart water technologies to improve its efficiency and to face water availability issues posed by demographic and climate changes.

In particular, this competition is aimed at helping Acea in preserving precious waterbodies which are threatened during the hottest seasons. To pursue this aim it is key to predict water availability in terms of level and water flow for each day of the year. Gaining a better understanding of volumes will enable the firm to ensure water availability all over the year as well as in the future.

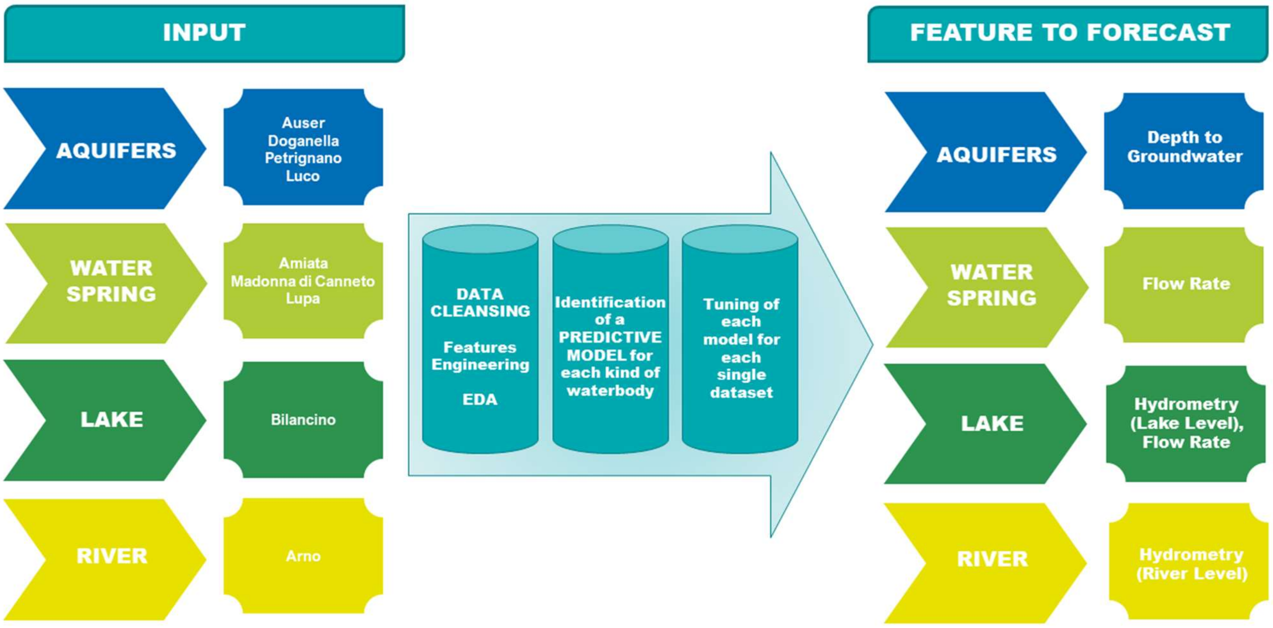
However, each waterbody presents unique characteristics, and their attributes are not linked to each other. Each waterbody has its own dataset which is completely independent from the other datasets. So, if we consider for instance a river, we notice that its features are different from those of a water spring due to the different behavior and characteristics of each waterbody. In fact, although they are both affected (in a different way) by temperature and rainfall, their water availability is expressed by dissimilar features. The river output variable is hydrometry (i.e., river level expressed in meters) while for the water spring it is the flow rate (i.e., volume of water in liters per second).

The Acea Group manages four types of waterbodies: water springs, lakes, rivers, and aquifers. The challenge is to determine how features like temperature, humidity, rainfalls, etc. influence the water availability of each. The output features to be forecasted for each type of waterbody are:

- Depth to groundwater for aquifers
- Flow rate for water springs
- Hydrometry (lake level) and flow rate for the lake
- Hydrometry (river level) for the river

In this thesis we will focus on one single waterbody: the Petrignano aquifer. As for the time interval the competition requires a weekly forecasting obtained as the mean of daily measures. Hence, the objective is to build a mathematical model for Petrignano aquifer for the forecasting of water availability expressed by depth to groundwater. The methodology used for building the model should be applicable also to other datasets.

Figure 4 shows the list of the nine waterbodies managed by Acea by category, the steps to follow, and the features to be forecasted by the model. The latter are simply the different units of measure for water availability for each category of waterbody that need to be forecasted.



INPUT
AQUIFERS
Auser Doganella Petrignano Luco
WATER SPRING
Amiata Madonna di Canneto Lupa
LAKE
Bilancino
RIVER
Arno
DATA CLEANSING
Features Engineering
EDA
Identification of a PREDICTIVE MODEL for each kind of waterbody
Tuning of each model for each single dataset
FEATURE TO FORECAST
AQUIFERS
Depth to Groundwater
WATER SPRING
Flow Rate
LAKE
Hydrometry (Lake Level), Flow Rate
RIVER
Hydrometry (River Level)

Figure 3.1. Source: Acea Analytics Competition

3.3 Available Data

Petrignano aquifer field is located in the alluvial plain between Ospedalicchio di Bastia Umbra and Petrignano and is fed by three underground aquifers separated by low permeability septa. The aquifer is considered a groundwater table and it is also fed by the Chiascio river. An aquifer is classified as groundwater table if the soil above it, up to the atmosphere, is permeable. This implies that the groundwater level is strongly affected by rainwater that, thanks to the fractures of the permeable superficial rocks, percolate in depth until it encounters a layer of impermeable soil.

Therefore, the Petrignano groundwater levels are influenced by rainfall, depth to groundwater, temperatures, drainage volumes, and level of the Chiascio river. Tables 3.1, 3.2 and 3.3 present each attribute in detail.

Field	Format	Description	Values
Date	Datetime	Uniquely identifies a day (Primary Key)	Daily value
Rainfall_X	Float	It indicates the quantity of rain falling, expressed in millimeters (mm), in the area **X**	Daily mean
Depth_to_Groundwater_Y (target)	Float	It indicates the groundwater level, expressed in ground level (meters from the ground floor), detected by the piezometer **Y**	Daily mean
Temperature_Z	Float	It indicates the temperature, expressed in **°C**, detected by the thermometric station **Z**	Daily mean
Volume_K	Float	It indicates the volume of water, expressed in cubic meters (**mc**), taken from the drinking water treatment plant **K**	Daily value
Hydrometry_H	Float	It indicates the groundwater level, expressed in meters (m), detected by the hydrometric station **H**	Daily mean

Table 3.1. Source: our elaboration.

Code	Values
X	Bastia_Umbra
Y	P24, P25
Z	Bastia_Umbra, Petrignano
K	C10_Petrignano
H	Fiume_Chiascio_Petrignano

Table 3.2. Source: our elaboration

Field	Min Value	Max Value
Rainfall_X	67.3	0
Depth_to_Groundwater_Y (target)	- 33.71	- 19.1
Temperature_Z	- 3.7	33
Volume_K	- 41890.176	- 16058.304
Hydrometry_H	1.8	4.1

Table 3.3. Source: our elaboration

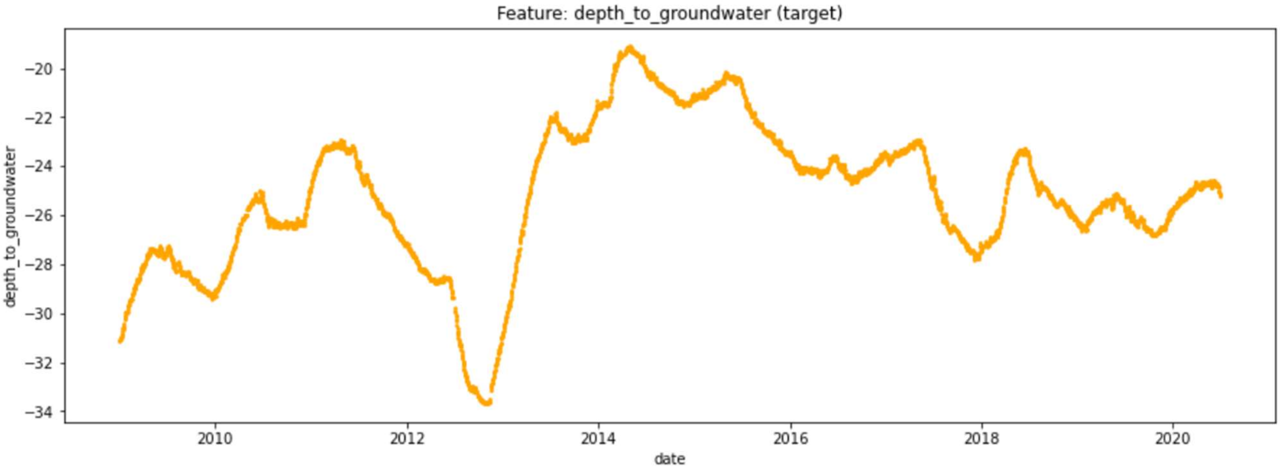
It is important note that some features like rainfall and temperature do not affect immediately the other features. This means, for instance, temperature of a specific day does not affect the other features right the same day but sometime later. Since we have no clue on how many days/weeks/months later temperature affects these features, this is another aspect to keep into consideration when analyzing the dataset.

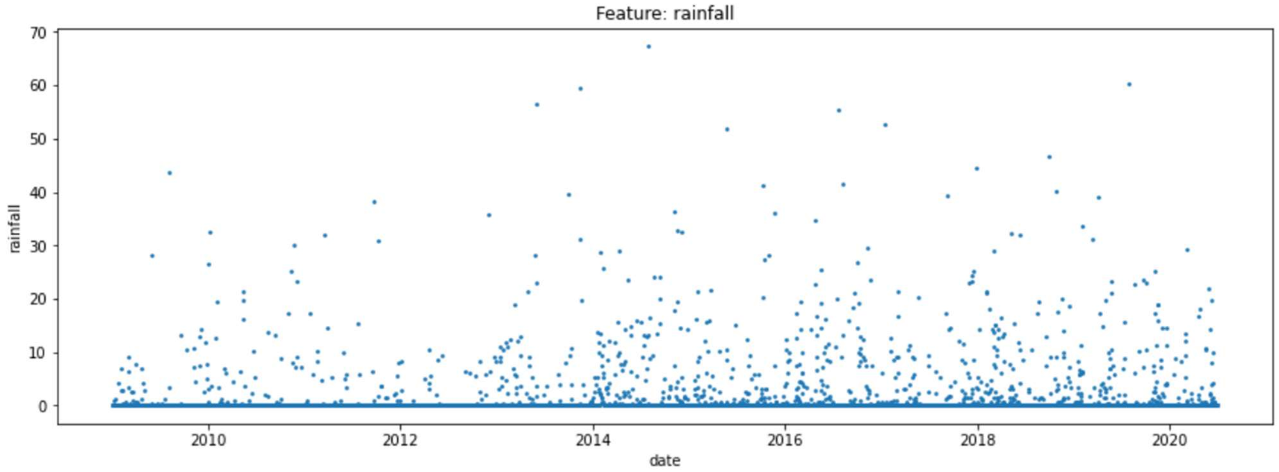
3.4 Data Understanding

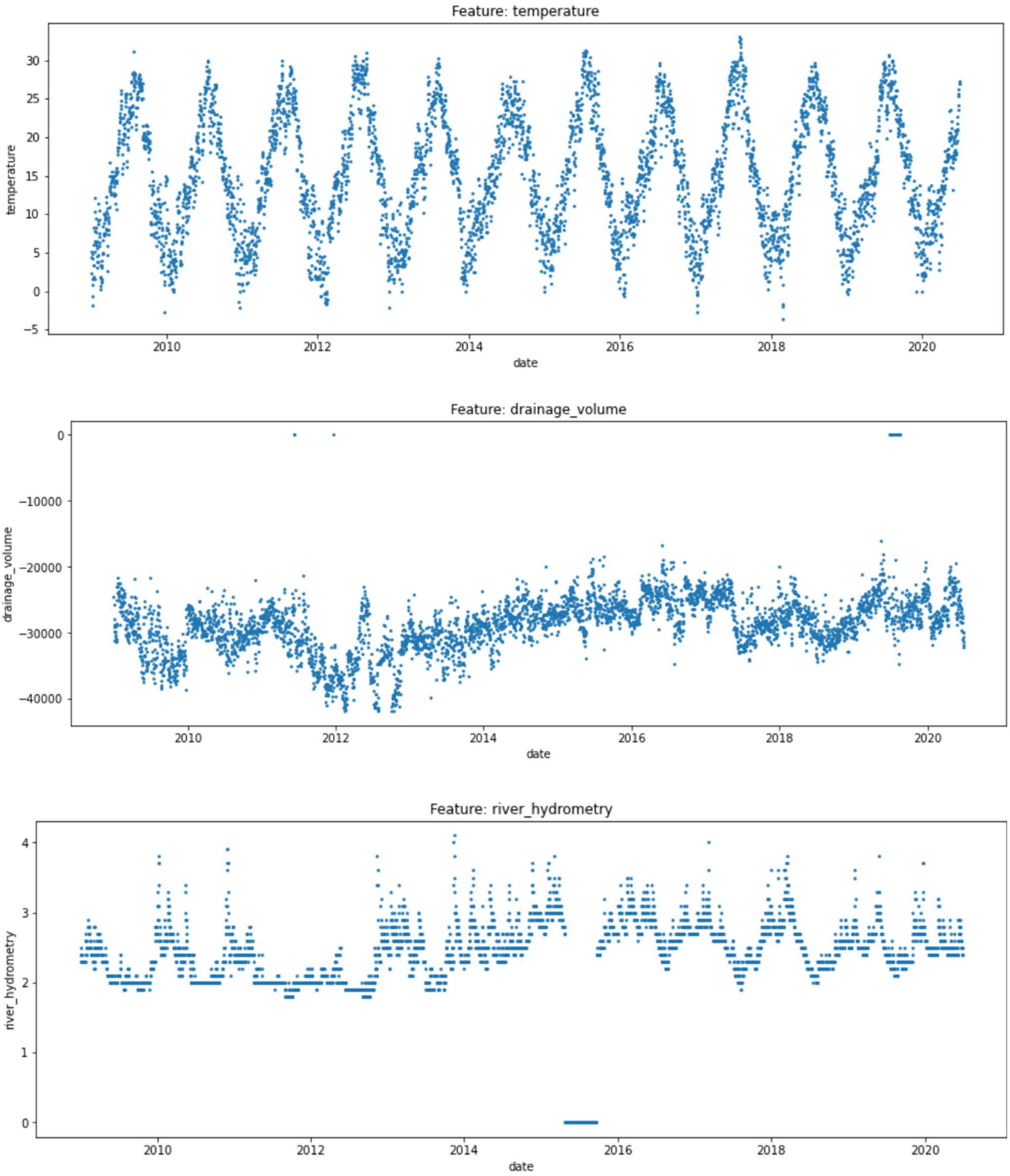
As already anticipated, although the dataset contains multiple waterbodies, we will only be looking at the `Aquifer_Petrignano.csv` file. In this example, we have the column `Date` which uniquely identifies a day. The time interval is one day, and the data is already in chronological order.

3.4.1 Data Visualization

The first useful step to get an insight on our time series is to plot each attribute. This will help us in understanding how the variable behaves and whether the data have to be processed before starting the construction of the model.



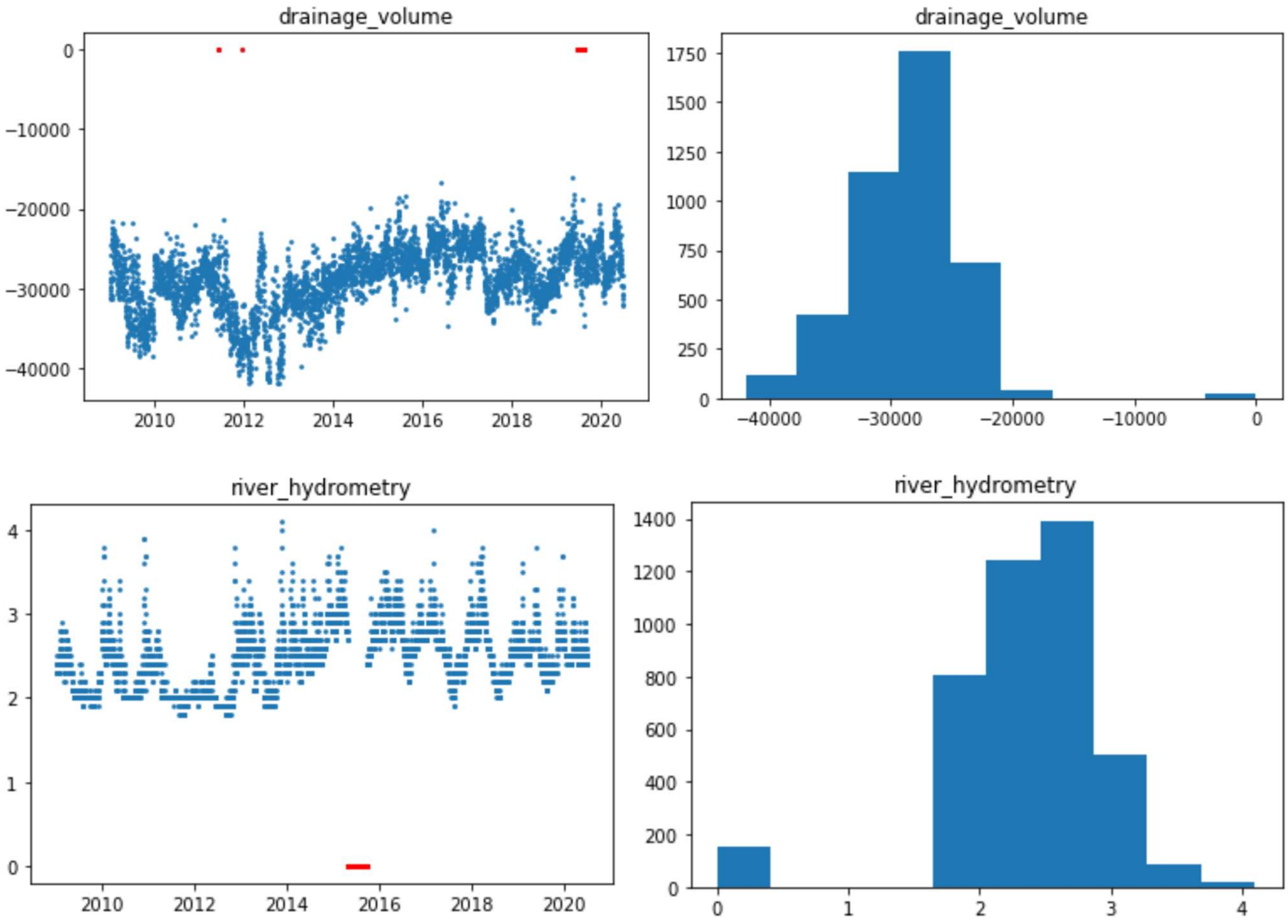




Graph 3.1. Source: our elaboration.

3.4.2 Zero Values

As we can easily notice the dataset present some unusual zero values in `drainage_volume` and `river_hydrometry`. We assume that these values are NaNs since they are uncommon values for the series that may result from measurement errors.

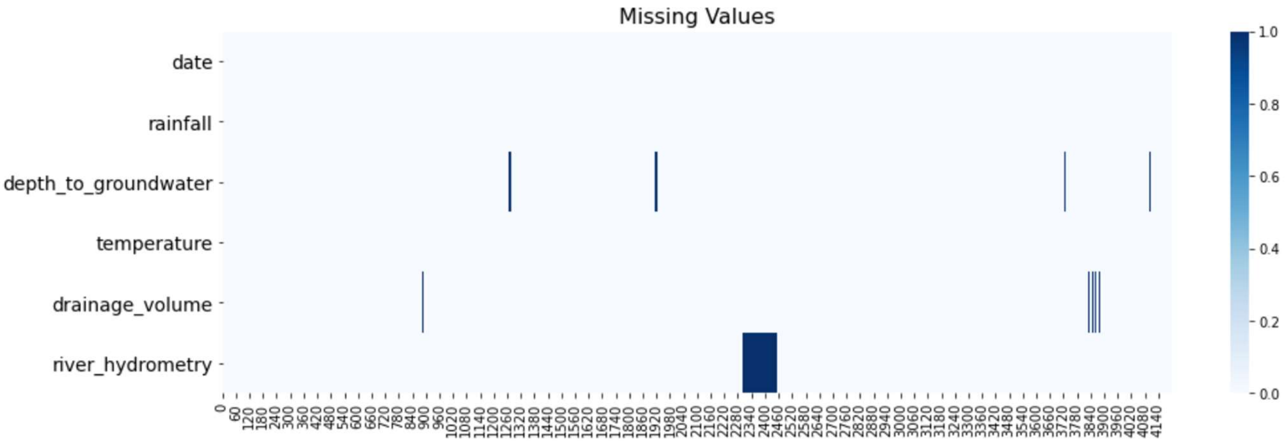


Graph 3.2. Source: our elaboration.

Hence, we proceed by substituting zero values with NaNs and subsequently count the overall number of NaN values present in each attribute. We will deal with the replacement of missing values in the preprocessing phase.

```
date                     0
rainfall                 0
depth_to_groundwater    27
temperature              0
drainage_volume         26
river_hydrometry       150
```

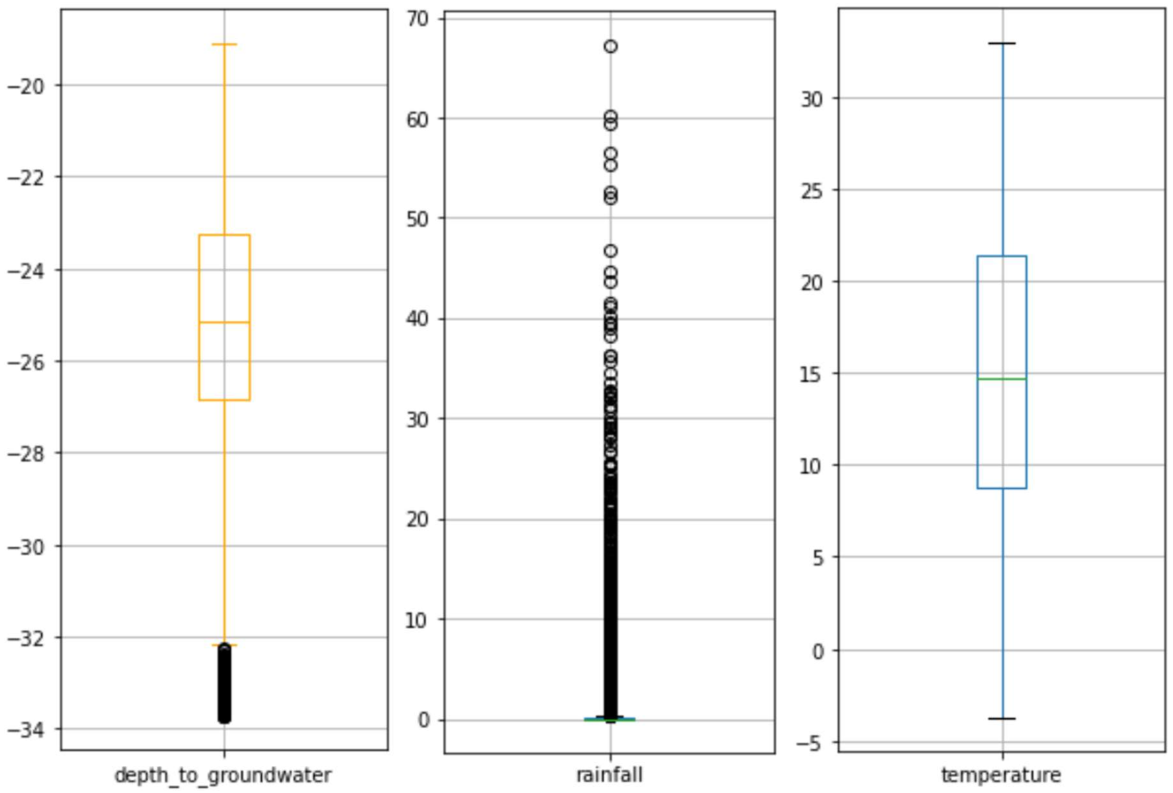
Constructing a heatmap can be useful to have a clue on how NaN are spread in each feature. In `depth_to_groundwater` they are spread randomly, in `drainage_volume` one is placed at day 900 while the others are concentrated between day 3800 and day 3900, and in `river_hydrometry` NaNs are consecutive and located between day 2300 and day 2400 approximately. Consecutive NaNs are harder to be replaced because it implies guessing a span of time which could lead to biased results.

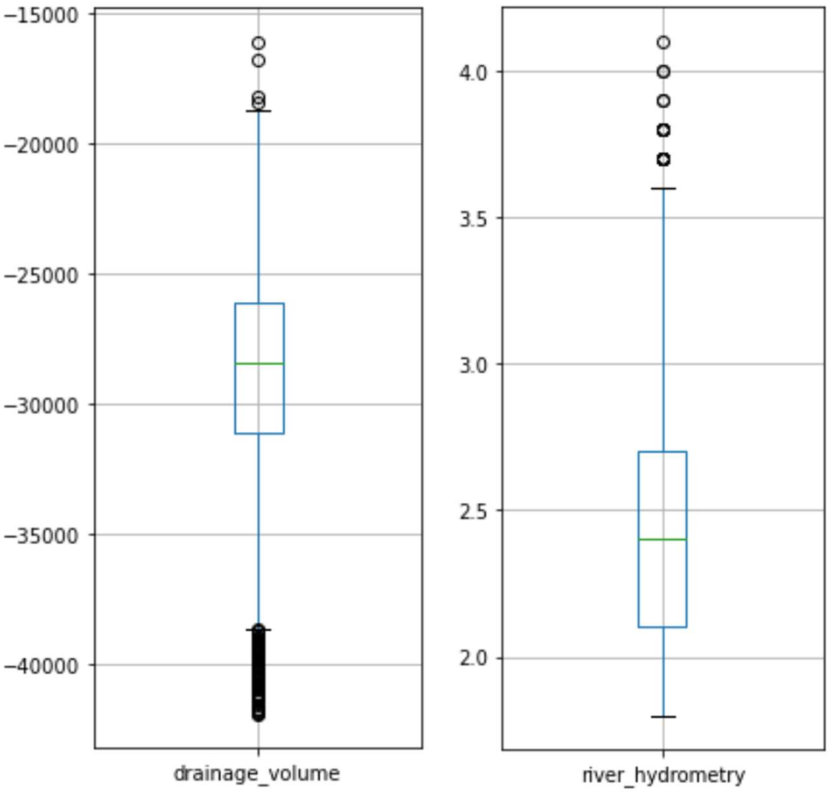


Graph 3.3. Source: our elaboration.

3.4.3 Distribution

At this point, it is useful to display how the values in each attribute are spread out. To do so we will produce boxplots that are a standardized way to represent a distribution based on a five-number summary. We will display them separately and not in a single plot since they have different scales.

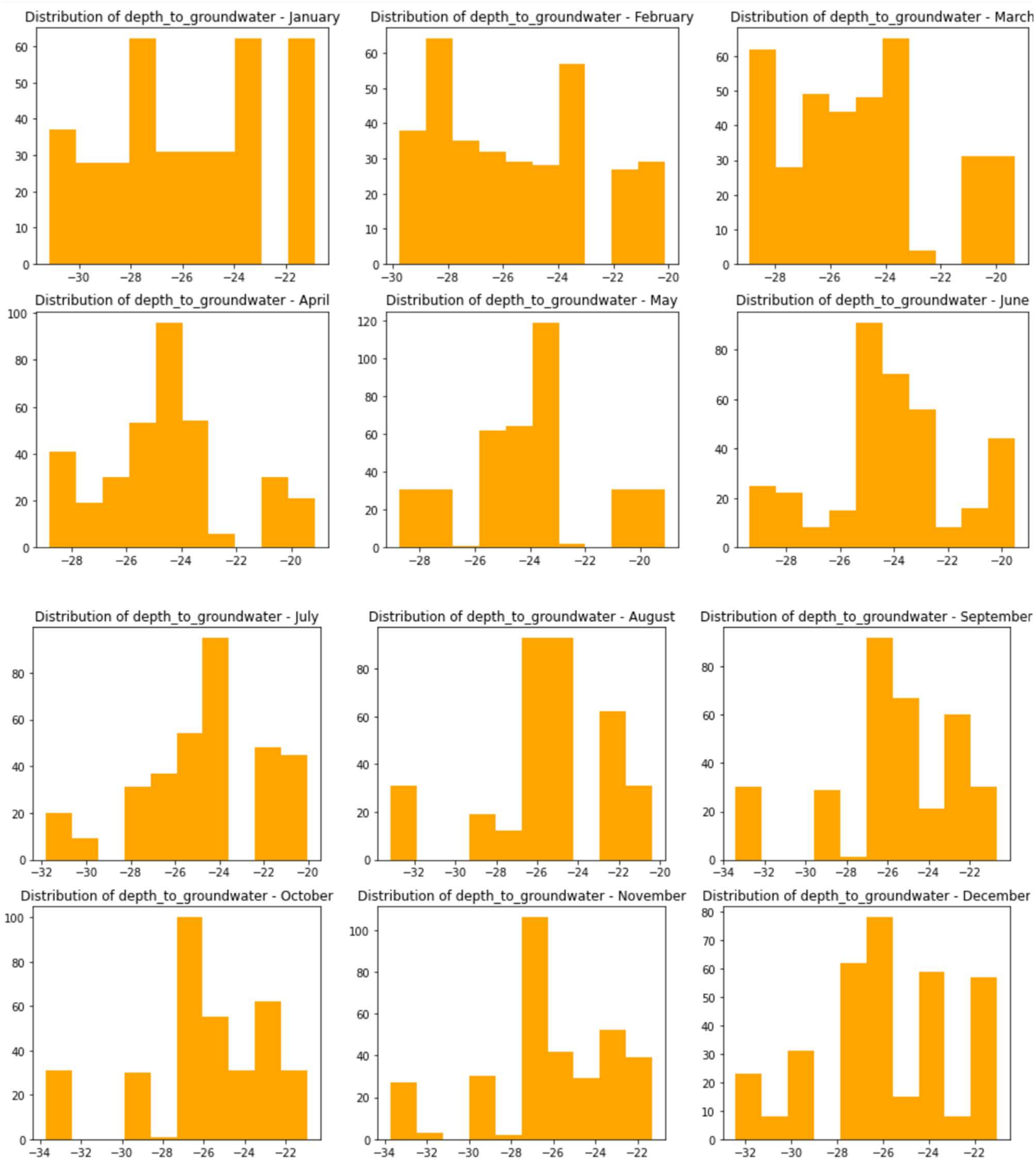




Graph 3.4. Source: our elaboration.

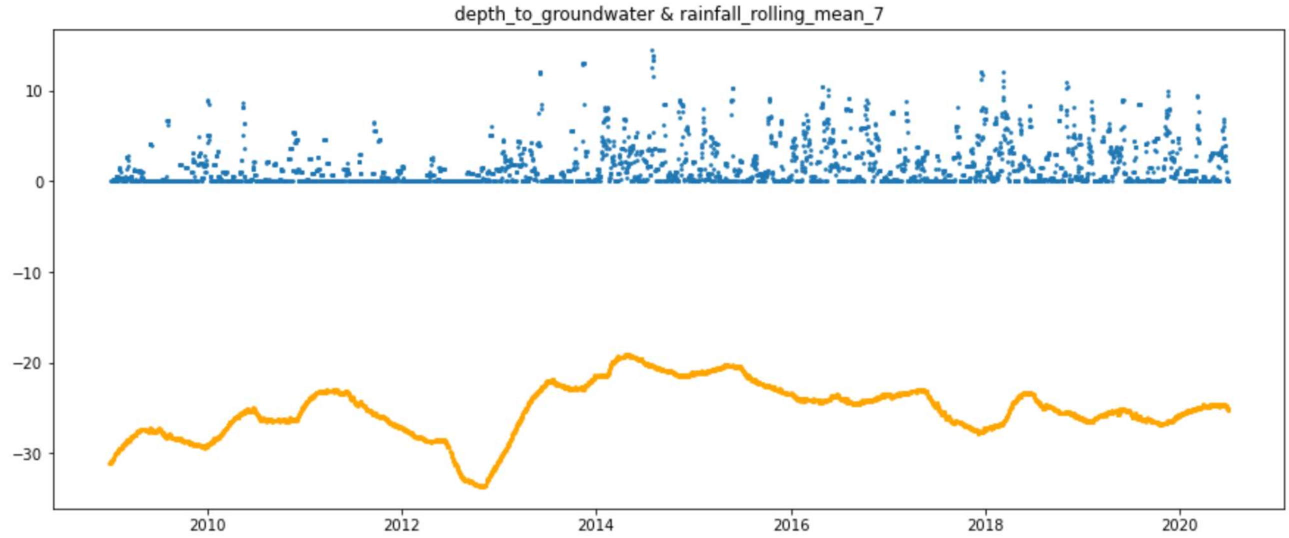
- `depth_to_groundwter` distribution is left skewed due to the presence of many outliers. Its value range between – 34 and – 19 meters from the ground floor.
- `rainfall` is extremely right skewed. Its values are normally close to zero (no rain) but there is a large number of outliers that coincide with rainy days.
- `temperature` is normally distributed, and its values range from – 4 to 33 degrees Celsius
- `drainage_volume` is left skewed and presents outliers on both sides. Its values range between – 40,000 and – 15,000 cubic meters.
- `river_hydrometry` is right skewed with the presence of some outliers. The 75% of values is concentrated between 1.7 and 2.7 meters.

To get a clearer view of the distributions we display 5 distinct histograms for the target variable, one for each month of the year. We notice that during warm months the `depth_to_groundwater` is higher than in colder months. This may suggest that weather features do not immediately impact the target value.



Graph 3.5. Source: our elaboration.

A further step in data understanding is putting into relation different features to see how their behaviors are related. For instance, we can plot on the same graph our target variable together with a smoothed version of `rainfall` (rolling mean of window 7). As expected, the values of `depth_to_groundwater` increase as values of `rainfall` increase. This because rainwater percolates into the acquifer and the groundwater level rises.



Graph 3.6. Source: our elaboration.

3.5 Data Preparation

3.5.1 Introduction

According to the CRISP-DM methodology data preparation is one of the most important and often time-consuming aspects of data mining. In fact, it is estimated that data preparation usually takes 50-70% of a project's time and effort. A good effort in the earlier stages of business understanding and data understanding can minimize this overhead, but you still need to spend a good amount of energies preparing and packaging the data for mining (IBM Corporation, 2018).

As defined by IBM CRISP-DM guide, the data preparation typically involves the following tasks:

- Merging data sets and/or records
- Selecting a sample subset of data
- Aggregating records
- Deriving new attributes
- Sorting the data for modeling
- Removing or replacing blank or missing values
- Splitting into training and test data sets

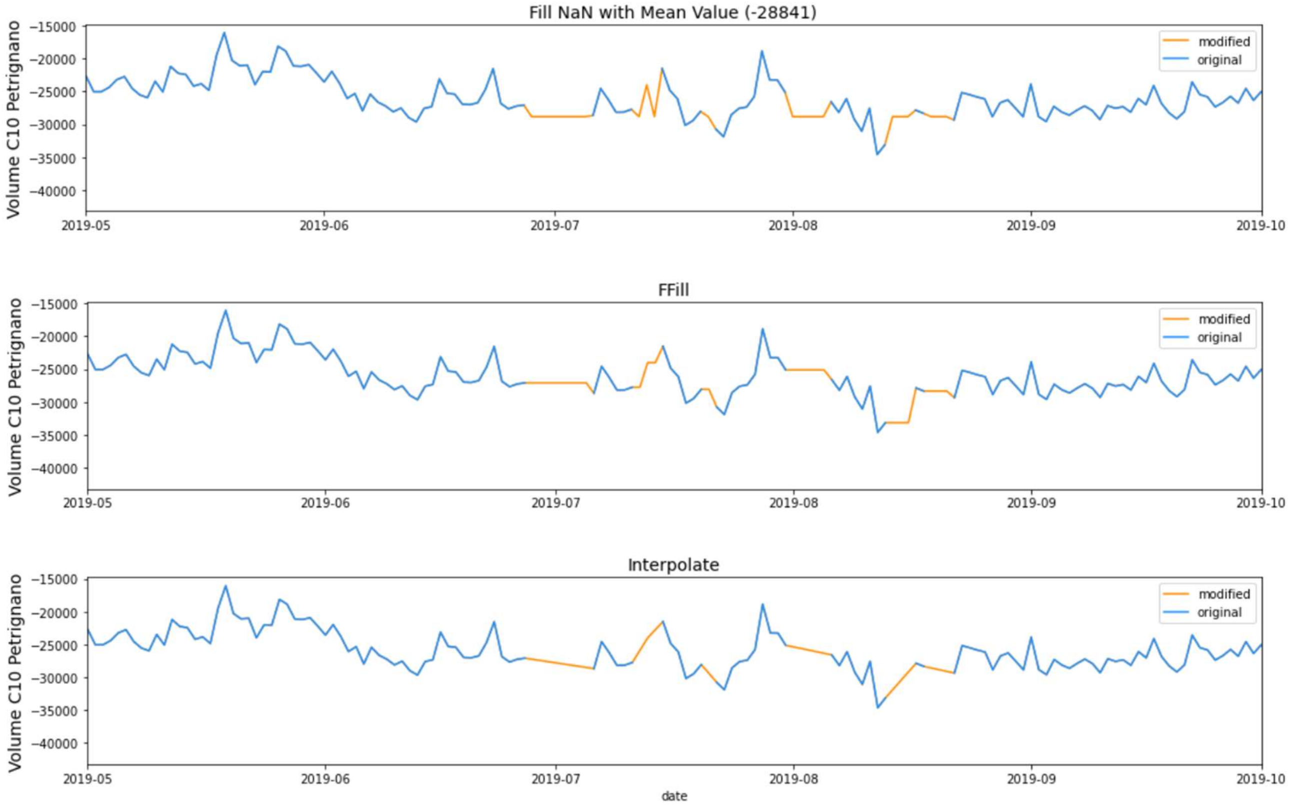
3.5.2 Handle Missing's

In the data-understanding phase null values have been replaced with NaN. Now, we will proceed by filling them with the most suitable values. We consider three options to deal with missing values:

1. Fill NaN with mean value

2. Fill NaN with last value with `.ffill()`
3. Fill NaN with linearly interpolated value with `.interpolate()`

Let’s analyze each option graphically in graph 3.7.


Fill NaN with Mean Value (-28841)

FFill

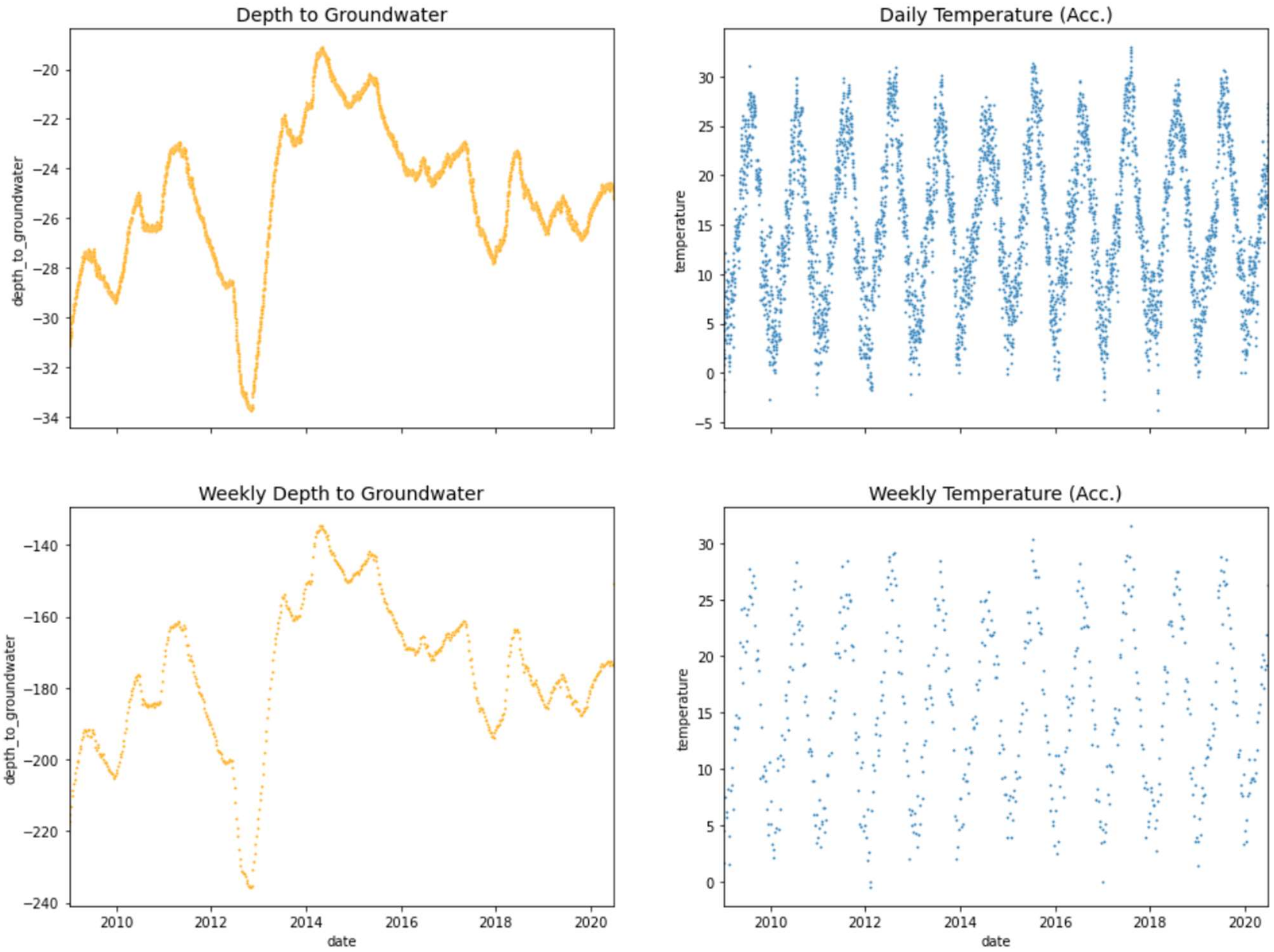
Interpolate

Graph 3.7. Source: our elaboration.

Filling NaN’s with the interpolated values is the best option even though it requires knowledge of the neighboring values. Thus, we proceed by interpolating missing values also for `depth_to_groundwater` and `river_hydrometry`.

3.5.3 Resampling

There are two types of resampling: upsampling (frequency is increased) and downsampling (frequency is decreased). Since the competition imposes weekly frequency of the forecasting, will opt for downsampling our data. This procedure will be done for every attribute by computing the weekly value as the mean of daily values. The following figure shows the graphical effect of downsampling on the target variable and `temperature`.



Graph 3.8. Source: our elaboration.

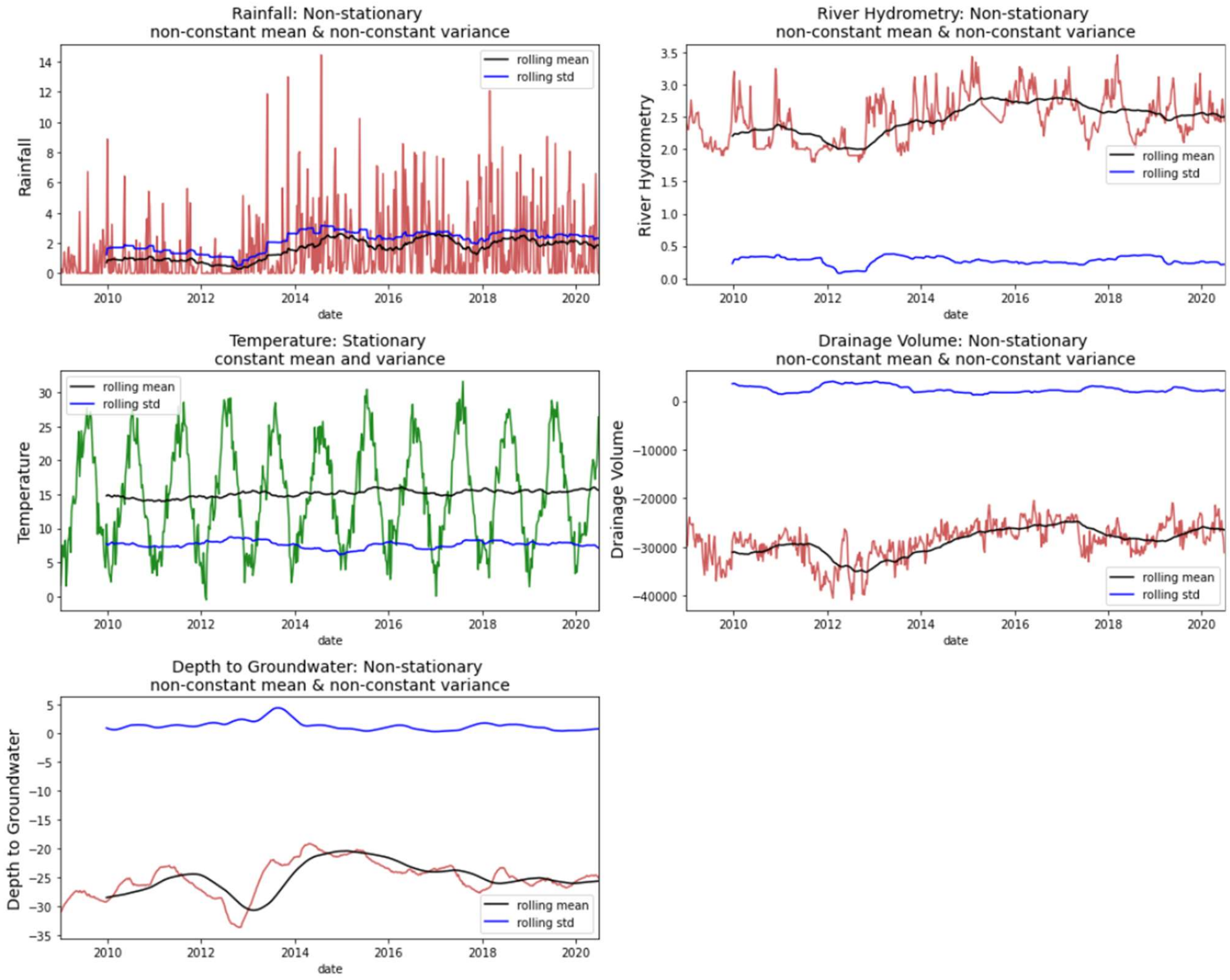
3.5.4 Stationarity

Some time series models, such as ARIMA, assume that the underlying data is stationary. This because if a time series has a specific (stationary) behavior over a given time interval, then it can be assumed that the time series will behave the same at a later time. Time series with trend and/or seasonality are not stationary. Remind that trend indicates that the mean is not constant over time and seasonality indicates that the variance is not constant over time (Pal, Prakash, 2017).

Stationarity is verified for each attribute because in case of multivariate models of forecasting the future behavior of the target variable is explained by forecasted behaviors of other variables that influence it. To predict future values of these influencing features we need them to be stationary.

The check for stationarity can be done both visually by plotting the time series and check for trend and seasonality and by means of the Augmented Dickey Fuller test.

Let's start with the visual check. We can note that all features except `temperature` have non-constant mean and non-constant variance. Therefore, none of them seems to be stationary.



Graph 3.9. Source: our elaboration.

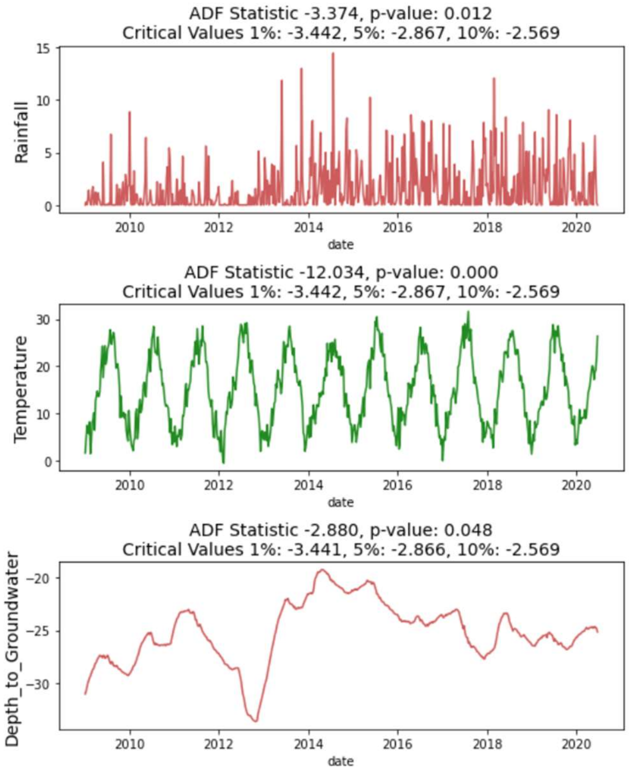
The Augmented Dickey-Fuller (ADF) test, which we already presented in chapter 2, is a type of statistical test called a unit root test (Pal, Prakash, 2017). Unit roots are a cause for non-stationarity.

- Null Hypothesis (H0): Time series has a unit root. (Time series is not stationary).
- Alternate Hypothesis (H1): Time series has no unit root (Time series is stationary).

If the null hypothesis can be rejected, we can conclude that the time series is stationary. There are two ways to rejects the null hypothesis:

- p-value is below a set significance level (defaults significance level is 5%)
- test statistic is less than the critical value

We can note that according to the ADF test `Temperature` and `river_hydrometry` are stationary.


ADF Statistic -3.374, p-value: 0.012
Critical Values 1%: -3.442, 5%: -2.867, 10%: -2.569

ADF Statistic -4.824, p-value: 0.000
Critical Values 1%: -3.441, 5%: -2.866, 10%: -2.569

ADF Statistic -12.034, p-value: 0.000
Critical Values 1%: -3.442, 5%: -2.867, 10%: -2.569

ADF Statistic -3.010, p-value: 0.034
Critical Values 1%: -3.441, 5%: -2.866, 10%: -2.569

ADF Statistic -2.880, p-value: 0.048
Critical Values 1%: -3.441, 5%: -2.866, 10%: -2.569

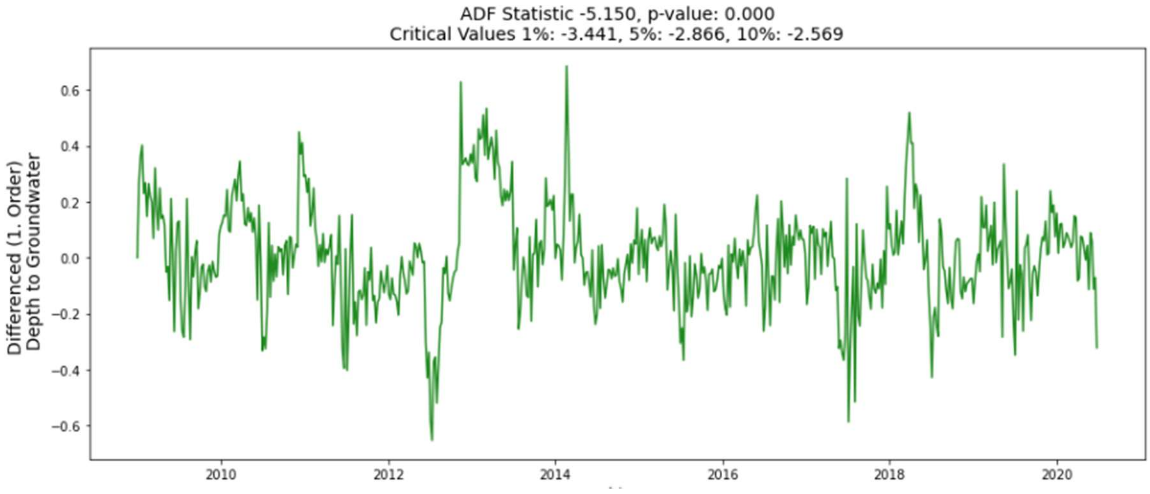
Graph 3.10. Source: our elaboration.

If the data is not stationary but we want to use a model that requires this characteristic, the data has to be transformed. As anticipated in the previous chapter, the most common method to achieve stationarity is differencing which implies subtracting the current value from the previous.

Differencing can be done in different orders:

- First order differencing: linear trends with $x_t = x_t - x_{t-1}$
- Second order differencing: quadratic trends with $x_t = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2})$
- and so on...

In our example with first order differencing, we reach stationarity.


ADF Statistic -5.150, p-value: 0.000
Critical Values 1%: -3.441, 5%: -2.866, 10%: -2.569

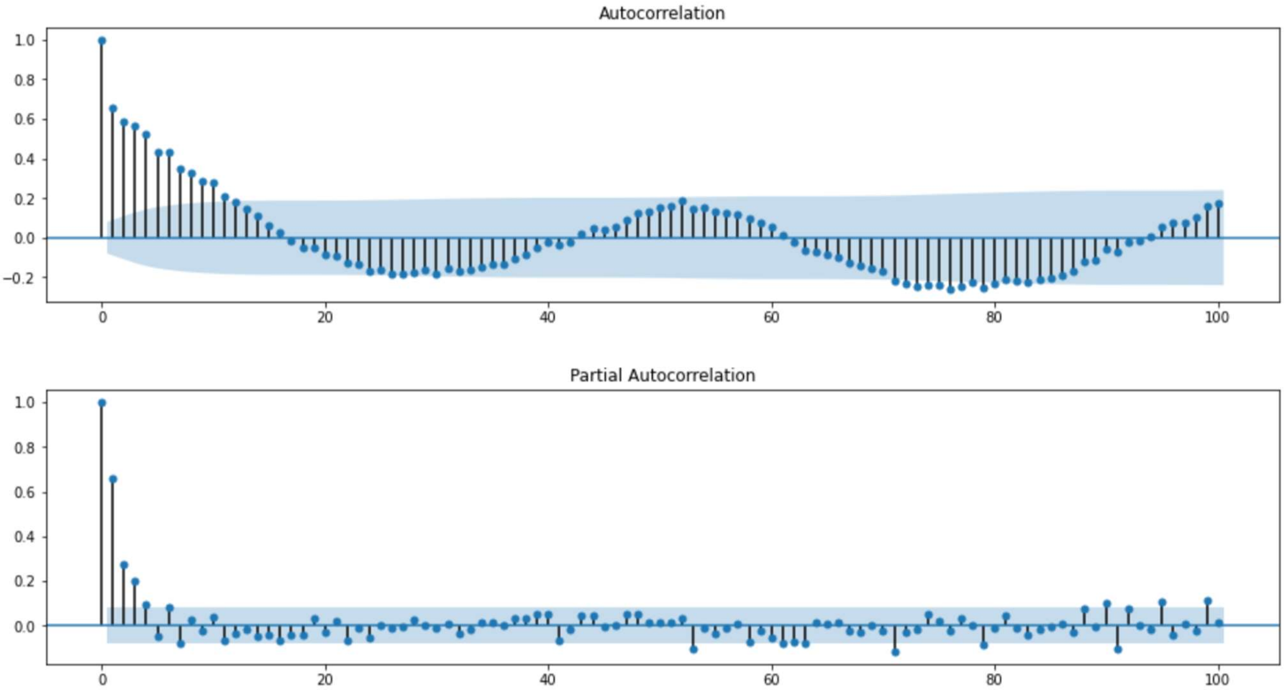
Graph 3.11. Source: our elaboration.

3.6.1 Autocorrelation Analysis

After the time series has been stationarized, the following step for fitting an ARIMA model is to determine AR and/or MA terms needed to correct any remaining autocorrelation in the differenced series.

The Autocorrelation Function shows the correlation between time series with a lagged version of itself. The Partial Autocorrelation Function, instead, explains the correlation that results after removing the effect of any correlations due to the terms at shorter lags.

By looking at the two plots we can see that both ACF and PACF show slow and gradual decrease. Hence, $ARMA(1,1)$ model could be appropriate for the series. Again, the PACF also shows a sharp drop after two significant lags which indicates that also an $AR(2)$ would be a good alternative. Therefore, we should experiment with both models and select the optimal one based on performance metrices.



Graph 3.12. Source: our elaboration.

3.7 Cross Validation

Cross-validation (CV) is a technique to evaluate learning models and assess how the statistical analysis generalizes to an independent data set. When we train a model, we split the dataset into two main sets: training and testing. The training set represents all the examples that a model is learning from, while the testing set simulates the testing examples. CV splits the dataset into k random sections, then trains the model on k-1 sections and test model on remaining one. These steps are repeated until each section behaved as test set for the model and the final metrics will be average of scores obtained in every section. This allows to grant generalization of the model and evaluate model performance in a more robust way than simple train-test split.

However, this technique is not suitable for time series where the temporal dependency between observations must be preserved. We cannot choose random samples and assign them to either the test set or the train set because it makes no sense to use the values from the future to forecast values in the past.

An alternative technique suitable for time series is cross-validation on a rolling basis. It begins by training the model with a small subset, forecast for the second subset of data points, and then checking the accuracy of the forecasting. The data points used for testing are then included as part of the next training dataset and subsequent data points are forecasted.

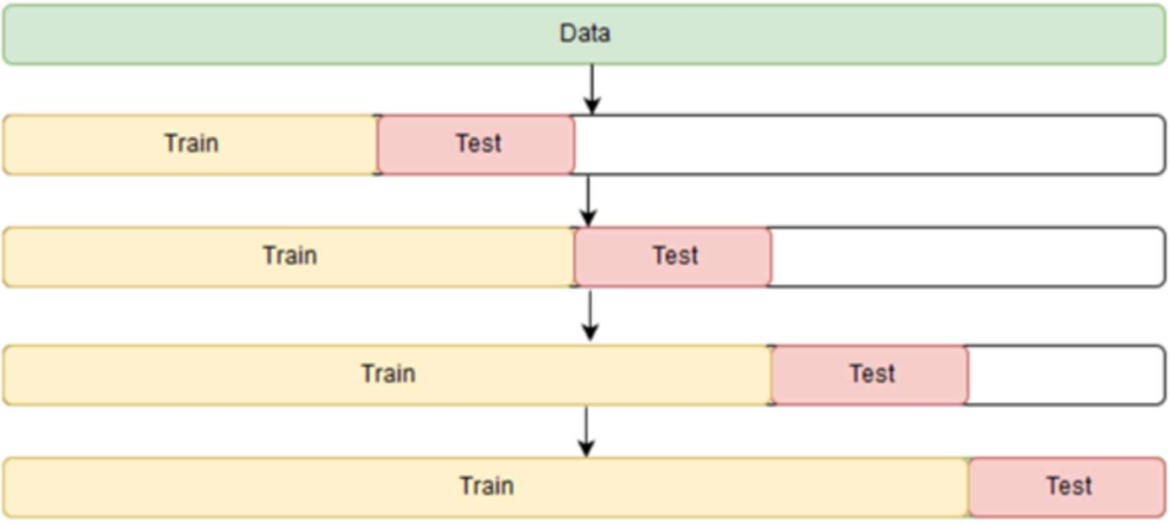


Figure 3.2. Source: Cross Validation in Time Series, Medium

To perform this technique, we use `TimeSeriesSplit` library from Scikit-learn which splits the training data into multiple (k) segments. It uses the first segment to train the model with a set of hyper-parameters and test it with the second. Then the model is trained on the first two chunks and tested on the third part of the data. In this way we perform the cross-validation technique $k - 1$ times.

3.8 Modelling and Evaluation

3.8.1 Introduction

Time series can be either univariate or multivariate:

- Univariate time series only has a single time-dependent variable

- Multivariate time series have a multiple time-dependent variable

Our example originally is a multivariate time series because it has multiple features that are all time dependent. However, by only looking at the target variable `depth_to_groundwater`, we can convert it to a univariate time series. In this thesis we will focus on univariate time series analysis.

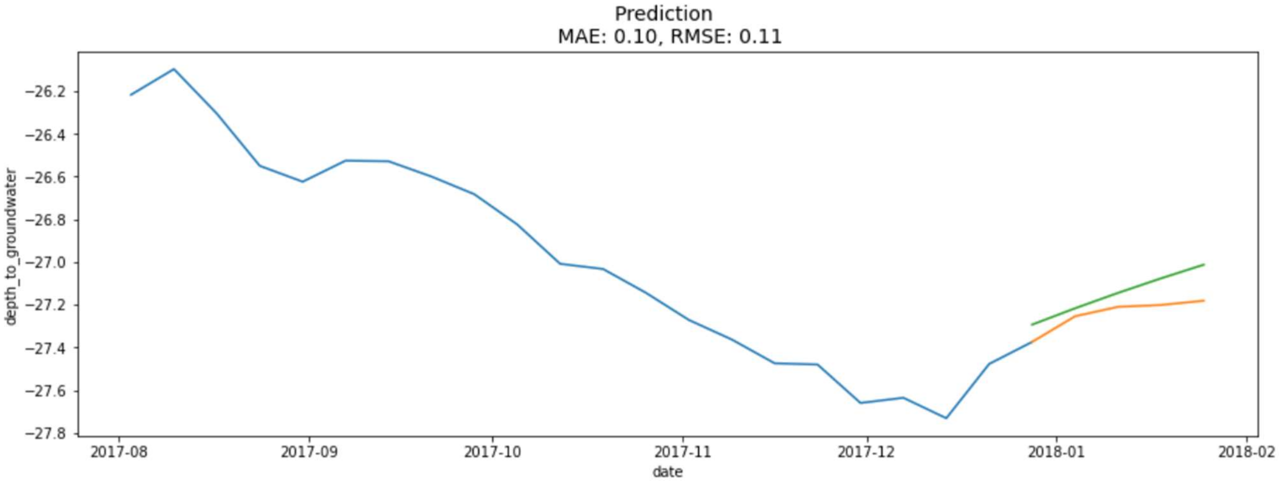
3.8.2 Univariate Time Series (ARIMA)

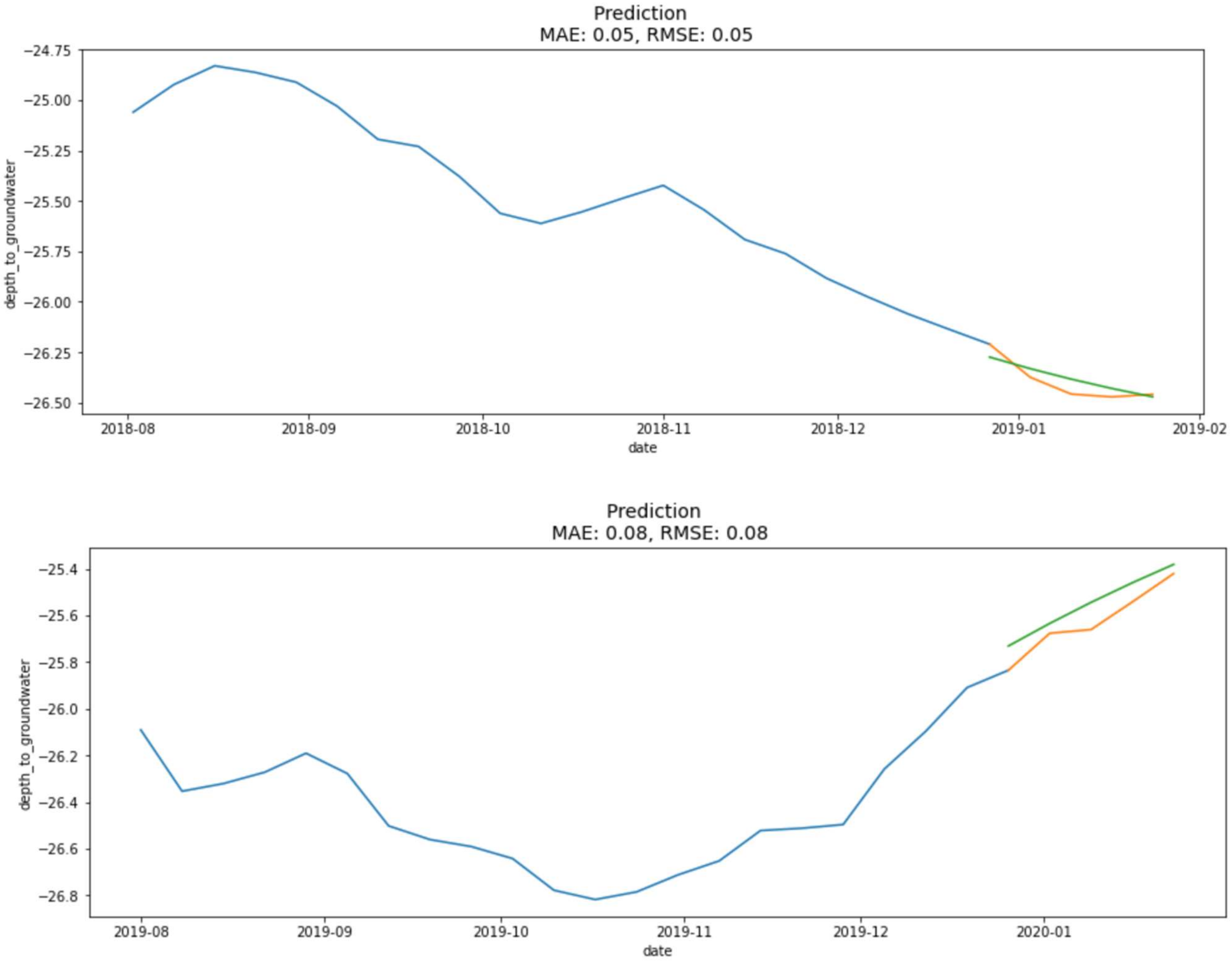
The Auto-Regressive Integrated Moving Average (ARIMA) model describes the autocorrelations in the data. The model assumes that the time-series is stationary. The ARIMA model is made up by three separate parts:

- AR (autoregression)
- I (integration or differencing)
- MA (moving average)

The model is usually represented as $ARIMA(p, d\ , q)$ where each of the three letters represent a parameter to be provided. Parameter p stands for the number of autoregressive (AR) terms, d determines the order of differencing, and q defines the number of moving average (MA) terms (Pal, Prakash, 2017). In our example we use $ARIMA(1,1,1)$ since we differentiated by order 1 and we estimated the other parameters by looking at ACF and PACF.

Following the technique of cross validation on a rolling basis we will repeat the training and test process of ARIMA for every split of the time series. We also compute for each iteration the MAE and RMSE to evaluate the performance of our model. Clearly, the performance of the model improves as it trains on a larger set of values.





Graph 3.13. Source: our elaboration.

3.9 Conclusions

In this chapter we have constructed an $ARIMA(1,1,1)$ model to forecast future availability of water for Petrignano aquifer. The water availability is expressed with the variable depth to groundwater and the frequency of forecasting is weekly. According to MAE and RMSE evaluation metrics we can state that the model is accurate.

We used an approach based on ARIMA model for univariate time series setting aside exogenous variables like temperature, rainfall, etc.

We could envision as a next step of this thesis a multivariate modelling approach which consider the provided exogenous variables, thus expecting a further improvement of our results. There are several activities that should be planned and performed in order to achieve this.

First of all, we should construct a set of covariates starting from the meteorological prediction of the exogenous variables, such as the average of each selected variable in the k days before the prediction (k to be fixed). Then, we need to study if there is any multicollinearity between the constructed variables, and eventually implement a strategy that consider a set of orthogonal factors (i.e., Principal Component Analysis). After these preliminary steps, it is necessary to construct a model that relates the exogenous variables, which will be given as prediction by using meteorological data, and the target variable. The choice of the modeling technique depends upon the assumptions of each algorithm, and the relation existing between the dependent variable and the independent variables. A careful evaluation should be done in order to study the residuals. In case there is still autocorrelation we suggest constructing a second modelling layer which is built by estimating a time series algorithm to predict the residual.

The further expiring that we are suggesting does not guarantee an improvement of performance, but the data understanding we did in this chapter shows that this approach could be promising.

Conclusion

This paper aimed at finding a forecasting model for predicting water availability of Petrignano aquifer for Acea Group. A deep understanding of each feature of the dataset through data visualization and distribution analysis allowed to plan the data preparation steps to undergo to make data ready for being processed. These steps were related with handling of missing values, resampling, stationarity, and autocorrelation analysis. Most of the variables, and especially the target variable, turned out to be non-stationary and were subject to differencing to be stationarized. Also, through autocorrelation analysis we were able to set the hyperparameters for our model.

The solution to the problem proposed in this thesis consists in an $ARIMA(1,1,1)$ model for univariate time series which forecasts future values of the target variable based on its own past behavior. Also, according to the evaluation metrics we can state that the constructed model is accurate and therefore a valid solution to the problem.

However, it would be interesting to deepen my research by considering also exogenous variables for the prediction of the target variable. This could be done with multivariate models and in this way, we would construct a two-step modeling. The first step consists in applying a model which relates an appropriate subset of explicative variables with the target variable. Then, if the there is still autocorrelation in the residuals, an autoregressive ARIMA model could be applied to the residuals.

This paper has contributed to expand my statistical knowledge as well as to explore the time series and forecasting branch. Furthermore, it has been my first practical resolution of a business problem and my first implementation of the CRISP-DM methodology. It has been very challenging and stimulating as well as an opportunity to learn new topics that go beyond those foreseen by my university course.

References

Bibliography

Gerasimos Antzoulatos, Christos Mourtzios, Panagiota Stournara, Ioannis-Omiros Kouloglou, and others (2020, Water Science & Technology). *Making urban water smart: the SMART-WATER solution*.

George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Greta M. Ljung (2015, Wiley). *Time Series Analysis Forecasting and Control.*

Rob J Hyndman and George Athanasopoulos (2018, Texts). *Forecasting: Principles and Practice. A comprehensive introduction to the latest forecasting methods using R.*

Avishek Pal, PKR Prakash (2017, Packt). *Practical Time Series Analysis. Master Time Series Data Processing, Visualization, and Modeling using Python.*

Jake VanderPlas (2016, O'Reilly). *Python Data Science Handbook. Essential tools for working with data*.

Sitography

Acea Group (2019). *Technology and digitalization for water networks.* https://www.gruppo.acea.it/en/stories/technology-digitalization/digitalization-water-network-contain-water-leaks

Acea Group (2018). *Our History and About Acea*. https://www.gruppo.acea.it/en/about-acea

AIM (2021). *A Guide to Different Evaluation Metrics for Time Series Forecasting Models*. https://analyticsindiamag.com/a-guide-to-different-evaluation-metrics-for-time-series-forecasting-models/

IBM (2021). *Introduction to CRISP-DM*. https://www.ibm.com/docs/en/spss-modeler/18.2.0?topic=guide-introduction-crisp-dm

Kaggle (2021). Acea Smart Water Analytics. https://www.kaggle.com/c/acea-water-prediction

Kaggle (2021). Competition Notebook for Acea Smart Water Analytics. *Time Series Analysis, a Complete Guide*. https://www.kaggle.com/code/andreshg/timeseries-analysis-a-complete-guide

Kaggle (2021). Competition Notebook for Acea Smart Water Analytics. *Intro to Time Series Forecasting*. https://www.kaggle.com/code/iamleonie/intro-to-time-series-forecasting

Left (2015). *Nel 2040 il mondo senz'acqua. Il rapporto del World Resources Institute*. https://left.it/2015/08/31/nel-2040-il-mondo-senzacqua-il-rapporto-del-world-resources-institute/

Medium (2020). *Cross Validation in Time Series.* https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4

Medium (2020). *Why is Augmented Dickey–Fuller test (ADF Test) so important in Time Series Analysis*. https://medium.com/@cmukesh8688/why-is-augmented-dickey-fuller-test-adf-test-so-important-in-time-series-analysis-6fc97c6be2f0

Mordor Intelligence (2022). *Smart Water Management Market: Growth, Trends, Covid-19 Impact, and Forecasts*. https://www.mordorintelligence.com/industry-reports/smart-water-management-market

Towards Data Science (2021). *Defining the Moving Average Model for Time Series Forecasting in Python*. https://towardsdatascience.com/defining-the-moving-average-model-for-time-series-forecasting-in-python-626781db2502

Towards Data Science (2020). *Time Series Analysis: Identifying AR and MA using ACF and PACF Plots.* https://towardsdatascience.com/identifying-ar-and-ma-terms-using-acf-and-pacf-plots-in-time-series-forecasting-ccb9fd073db8

Towards Data Science (2020). *What is Cross-Validation? Testing your machine learning models with cross-validation.* https://towardsdatascience.com/what-is-cross-validation-60c01f9d9e75

Water and Wastes Digest (2021). *What is smart water technology?* https://www.wwdmag.com/editorial-topical/what-is-articles/article/10939511/what-is-smart-water-technology

Acknowledgements

Firstly, I’d like to thank my supervisor, Prof. Marco De Ieso, for his professionalism and dedication to this thesis. He gave me the possibility to deepen my knowledge on Data Science and provided me with precious lessons and advice which I will treasure for my future professional career.

Also, I’d like to thank my family and friends for supporting me, especially my grandmother Carla who taught me the art of being strong even in toughest times.