



Management And Optimisation Proposals For An Industrial Water Supply System

Author:

André Gomes Domingues

Company Supervisor:

Engineer Carlos Rafael Jesus Aleixo

Faculty Supervisors:

Professor José Joaquim da Mota Machado, Ph.D.

Professor Marco Paulo Lages Parente, Ph.D.

Master Dissertation

Master In Mechanical Engineering

Faculty of Engineering of the University of Porto

June 2023

Porto

Contents

Abbreviations And Symbols	xv
Abstract	xvii
Acknowledgments	xix
1 Introduction	1
1.1 Context	1
1.2 The Company	2
1.3 Key Objectives	3
1.4 Methodology	3
1.5 Dissertation Structure	4
2 Literature Review	5
2.1 Toyota Production System	5
2.1.1 Toyota Production System Pillars	6
2.1.2 Toyota Production System Management Tools	7
2.1.2.1 5S	8
2.1.2.2 PDCA - Plan-Do-Check-Act	10
2.1.2.3 The Seven Quality Tools	10
2.1.2.4 A3 Problem - Solving	11
2.2 Water - The Key To Life On Earth	12
2.3 Water Consumption In The Automotive Industry	14
2.4 Industry 4.0	18
2.5 The main driving pillars of I4.0	19
2.5.1 Big Data And Analytics	19
2.5.2 Autonomous Robots	20
2.5.3 Simulation	20
2.5.4 Vertical and Horizontal System Integration	21
2.5.5 The Cloud	21

2.5.6	Industrial Internet of Things	21
2.5.7	Augmented Reality	22
2.5.8	Cybersecurity	22
2.6	Water 4.0	22
2.6.1	Integrated Business Model	25
2.6.1.1	Technologies Component	25
2.6.1.2	Business Stack of Systems Component	27
2.6.1.3	Business Strategy/Management Component	28
2.7	Predictive Forecasting Techniques	28
2.7.1	Statistical Measures For Univariate Datasets	30
2.7.1.1	Measures of Location	30
2.7.1.2	Measures Of Dispersion	31
2.7.2	Statistical Measures For Bivariate Datasets	31
2.7.3	Normal Distribution	32
2.7.4	Data Pre-Processing And Preparation	35
2.7.4.1	Dealing With Missing Values	35
2.7.4.1.1	Deletion	36
2.7.4.1.2	Imputation	36
2.7.4.2	Outliers Detection In Linear Regression Analysis	36
2.7.4.3	Outliers detection using robust regression models	40
2.7.5	Stationarity, Trend And Seasonality Definitions In Time Series	42
2.7.6	Stationarity Analysis	43
2.7.6.1	Augmented Dickey-Fuller Test	43
2.7.6.2	Kwiatkowski–Phillips–Schmidt–Shin Test	44
2.7.7	Seasonality Analysis	45
2.7.7.1	Cycle Plot	46
2.7.7.2	Decomposition of the time series	46
2.7.7.3	Lagged Autocorrelation - ACF Test	49
2.7.7.4	Partial Lagged Autocorrelation - PACF	49
2.7.8	Univariate Time Series	50
2.7.8.1	Autoregressive Model	50
2.7.8.2	Moving Average Model	50
2.7.8.3	Autoregressive Integrated Moving Average Model	51
2.7.8.4	Seasonal Autoregressive Integrated Moving Average Model	52

2.7.8.5	Exponential Smoothing	52
2.7.9	Time Series Regression	54
2.7.9.1	Multiple Linear Regression	55
2.7.9.2	Multiple Nonlinear Regression	55
2.7.9.3	ARIMA with exogenous variables	55
2.7.9.4	SARIMA with exogenous variables	56
2.7.10	Artificial Neural Networks	56
2.7.10.1	Feed-Forward Neural Networks	58
2.7.10.2	Extreme Learning Machines	59
2.7.10.3	Multi-Layer Perceptron	61
3	The Industrial Water Supply System	65
3.1	The Water Supply System	65
3.1.1	Pumping Station	67
3.1.2	The Fire Protection System - RIA	67
3.1.3	Reverse Osmosis - Demineralized Water	70
3.2	System Analysis	71
3.2.1	<i>AS-IS</i>	71
4	Industrial Water Dashboard	73
4.1	Goal Question And Metric Approach	73
4.2	GQM Applied To The Industrial Water Dashboard	74
4.3	The Industrial Water Dashboard	76
4.3.1	System Requirements and UML Diagram	76
4.3.2	Data Pre-Processing And Preparation	78
4.3.3	<i>MySQL</i> Database	79
4.3.4	<i>PowerBI</i> Dashboard	79
4.3.4.1	Main Menu	79
4.3.4.2	Water Consumption Records Menu	79
4.3.4.3	Water Consumption Records Dashboard	80
4.3.4.4	Water KPIs Dashboard	82
5	Water Consumption Analysis and Prediction	83
5.1	Water Consumption Analysis In Car Production	84
5.1.1	Industrial Water Consumption Analysis	84

5.1.2	Demineralized Water Consumption Analysis	86
5.1.3	Analysis Conclusions	87
5.2	Water Consumption Analysis In The Paintshop	88
5.2.1	Industrial Water Consumption Analysis in Pre-Treatment	88
5.2.2	Demineralized Water Consumption Analysis in Pre-Treatment	90
5.2.3	Demineralized Water Consumption Analysis in KTL	91
5.2.4	Analysis Conclusions	93
5.3	Water Consumption Prediction Model	94
5.3.1	Data Pre-Processing	94
5.3.2	Outlier Detection With Hampel Filtering	96
5.3.3	Seasonality Analysis	98
5.3.3.1	Decomposition of the time series	98
5.3.3.2	ACF Test	99
5.3.3.3	PACF Test	99
5.3.4	Stationarity Analysis	100
5.3.4.1	Augmented Dickey-Fuller Test	100
5.3.4.2	Kwiatkowski–Phillips–Schmidt–Shin Test	100
5.3.4.3	Data Differencing	101
5.3.5	Model Training and Evaluation	102
5.3.5.1	Exogenous variables dataset	102
5.3.5.2	Metrics for model evaluation	102
5.3.5.3	SARIMA and SARIMAX models	103
5.3.5.4	Exponential Smoothing models	105
5.3.5.5	NNETAR models	107
5.3.5.6	ELM and MLP models	109
5.3.5.7	Conclusions	109
6	Digitalization of the Water Supply System	111
6.1	Improvement Proposals	111
6.1.1	Pumping Station And RIA	111
6.1.2	Factory Entrance And Surroundings	112
6.1.3	Water Proof Test And <i>Decapagem</i>	113
6.1.4	Canteen and Men’s Bathroom	113
6.1.5	Reverse Osmosis	113
6.1.6	Laundry Room	114

6.1.7	Paint Shop	114
6.2	<i>TO-DO</i> And <i>TO-BE</i>	114
6.2.1	Integrated Water Metering	115
6.2.2	Low-Power Wide Area Network Related Technologies	116
6.2.3	Criteria for LPWANs	117
6.3	Analysis of the potential digitalization of in-use water meters	120
6.3.1	Work partnership with <i>Go.Charge</i>	120
7	Conclusions And Future Work	121
7.1	Conclusions	121
7.2	Future Work	121
	Bibliography	130
A	Gantt Diagram	131
B	A3 Problem Solving Template	133
C	Industrial Water And RIA Tank Supply System Technical Drawing	135
D	First UML iteration	137
E	Water Consumption Analysis - Robust Regression Code	139
F	Water Consumption In The Paintshop - Robust Regression Code	141
G	Time Series Decomposition, Statistical Tests And SARIMAX Forecastings Code	143
H	Exponential Smoothing Forecastings Code	151
I	NNETAR (Neural Networks) Forecasting Code	155
J	A3 Solving of the water digitalization process	161
K	Water meter specifications and digitalization strategy	163

List of Figures

1.1	Annual Water consumption and vehicle production - from 2018 to 2022	2
1.2	Toyota Caetano Portugal, S.A	3
2.1	The main pillars of the TPS	5
2.2	5S as a starting point for multiple lean production techniques	8
2.3	5S Diagram	9
2.4	PDCA cycle approach	10
2.5	Ishikawa diagram used in the company	11
2.6	Europe's water demand	12
2.7	South Asia's water demand	13
2.8	Groundwater depletion rate	14
2.9	Water used in car production in the European Union	15
2.10	Timelapse of the industrial revolutions	18
2.11	The 6 V's of Big Data	20
2.12	Digitization in water management	23
2.13	Comparison of the four industrial development stages	24
2.14	IBM functioning scheme	25
2.15	Forecasting performance (sMAPE) of the ML and statistical methods included in the study	30
2.16	Error definition in the least square method	32
2.17	Normal Distribution	33
2.18	Symmetrical values of distribution $N(0,1)$	34
2.19	Types of outliers: (a) Point outliers, (b) Context outliers, (c) Collective outliers	37
2.20	Cycle plot of item sold by week day	46
2.21	Airline dataset plotted	47
2.22	Multiplicative Decomposition	47
2.23	Additive Decomposition	48
2.24	Neuron layout	56
2.25	Single Layer FNN	59

2.26	Multiple Layer FNN	59
2.27	ELM structure	60
2.28	Example of an MLP with three LTUs in the hidden layer and two LTUs in the output layer	62
2.29	Perceptron with 3 LTUs	62
3.1	Location of the Toyota Caetano Ovar plant - satellite top view	65
3.2	Location of the boreholes and ETAR	66
3.3	Location of the main equipment of the water supply system	66
3.4	EFAFLU AQUA Premium CFD system	67
3.5	Eco-line ECO315S-2 electrical pump	68
3.6	GOULDS 3408 8x6x12 F-M diesel pump	69
3.7	WBJP 5-16 jockey pump	69
3.8	Reverse osmosis process scheme	70
3.9	Reverse osmosis process in the plant	71
3.10	Water supply system Ishikawa diagram	72
4.1	GQM notation in three levels - G: Goal, Q: Question, M: Metric	74
4.2	GQM model to obtain the metrics for the industrial water dashboard	75
4.3	Final UML Diagram for the industrial water dashboard	78
4.4	Dashboard main page	79
4.5	Consumption records menu	80
4.6	Daily consumption records type menu	81
4.7	Monthly consumption records type menu	81
4.8	Water KPI dashboard	82
5.1	Data science process	83
5.2	Outlier detection in the industrial water dataset	84
5.3	Industrial Water Consumption Analysis	85
5.4	Outlier detection in the demineralized water dataset	86
5.5	Demineralized Water Consumption Analysis	87
5.6	Outlier detection in the industrial water dataset - PT	89
5.7	Industrial Water Consumption Analysis - PT section	89
5.8	Outlier detection in the demineralized water dataset - PT	90
5.9	Demineralized Water Consumption Analysis - PT	91
5.10	Outlier detection in the demineralized water dataset - KTL	92

5.11	Demineralized Water Consumption Analysis - PT	93
5.12	Total water consumption - 2018 to 2022	94
5.13	Annual water consumption - 2018 to 2022	95
5.14	Monthly water consumption - per year	95
5.15	Original industrial water dataset	97
5.16	Industrial water dataset after applying the Hampel filtering	97
5.17	STL decomposition of the industrial water consumption time series	98
5.18	ACF Test	99
5.19	PACF Test	100
5.20	Differenced industrial time series	102
5.21	Training of the SARIMAX - Temperature + Humidity model	104
5.22	Forecast of the SARIMAX - Temperature + Humidity model	105
5.23	Training of the ETS model	106
5.24	Forecast of the ETS model	106
5.25	Training of the NNETAR - Temperature + Humidity model	108
5.26	Forecasting of the NNETAR - Temperature + Humidity model	108
6.1	New water meter location in the pump station	112
6.2	Electrovalve placed next to the factory's entrance	113
6.3	Reverse Osmosis PT meter	114
6.4	AMR data flow chart	115
6.5	AMI data flow chart	116
6.6	Types of LPWANs	117
6.7	Spider chart of the three LPWAN technologies	119
7.1	Main outcomes from the developed work in the dissertation	121
A.1	Gantt diagram	131
B.1	A3 problem solving company template	133
C.1	Industrial Water And RIA Tank Supply System Technical Drawing	135
D.1	First handwriting UML iteration for the industrial water dashboard	137
J.1	A3 Solving of the water digitalization process	161
K.1	Water meter specifications and digitalization strategy - page 1	163

K.2 Water meter specifications and digitalization strategy - page 2 164

K.3 Water meter specifications and digitalization strategy - page 3 165

K.4 Water meter specifications and digitalization strategy - page 4 166

K.5 Water meter specifications and digitalization strategy - page 5 167

K.6 Water meter specifications and digitalization strategy - page 6 168

K.7 Water meter specifications and digitalization strategy - page 7 169

K.8 Water meter specifications and digitalization strategy - page 8 170

List of Tables

1.1	Critical equipment	1
2.1	Life cycle consumption of a passenger vehicle	16
2.2	Reported and calculated water use and consumption data by OEM on a per-unit basis	17
2.3	Main conclusions from the study	18
2.4	Main benefit of each technology for IBM	26
2.5	Main benefit of each business stack of systems for IBM	27
2.6	Main benefit of each business strategy/management for IBM	28
2.7	Conclusions from the ADF Test according with each equation	44
2.8	Formulas for recursive calculations and point forecasts using Exponential Smoothing	53
2.9	Additive ETS Models	54
2.10	Multiplicative ETS Models	54
2.11	Activation functions for ANNs	57
5.1	Results from the linear regression analysis of the industrial water consumption	85
5.2	Results from the linear regression analysis of the demineralized water consumption	86
5.3	Predicted value for water consumption according to the daily production . . .	88
5.4	Results from the linear regression analysis of the industrial water consumption	89
5.5	Results from the linear regression analysis of the demineralized water consumption - PT	90
5.6	Results from the linear regression analysis of the industrial water consumption	92
5.7	Predicted value for water consumption according to the daily production - PT and KTL	94
5.8	ADF test results	100
5.9	KPSS test results	101
5.10	Training and testing values for SARIMA and SARIMAX with single exogenous variable	103

5.11 Training and testing values for SARIMAX with three and two exogenous variables 104

5.12 Training and testing values for exponential smoothing models 105

5.13 Training and testing values for NNETAR and NNETAR with single exogenous variable 107

5.14 Training and testing values for NNETAR with three and two variables 107

6.1 Data range for each LPWAN technology (Adapted from Lalle et al. (2019)) . . 118

Abbreviations And Symbols

4IR	Fourth Industrial Revolution
ADF	Augmented Dickey-Fuller Test
AI	Artificial Intelligence
AR	Augmented Reality
ARIMA	Autoregressive Integrated Moving Average
CPPS	Cyber-Physical Production System
CRM	Customer Relationship Management
GQM	Goal Question And Metric
H_0	Null Hypothesis
H_A	Alternative Hypothesis
I4.0	Industry 4.0
IBM	Integrated Business Model
IDE	Integrated Development Environment
IoT	Internet of Things
KPI	Key Parameter Indicator
KPPS	Kwiatkowski–Phillips–Schmidt–Shin Test
ML	Machine Learning
PDCA	Plan - Do - Check - Act
SARIMA	Seasonal Autoregressive Integrated Moving Average
SARIMAX	Seasonal Autoregressive Integrated Moving Average With Exogenous Variables
SAP	Systems Applications and Products
SCM	Supply Chain Management
STL	Seasonal and Trend Decomposition Using Loess
TPS	Toyota Production System
UI	User Interface
UTS	Univariate Time Series
W4.0	Water 4.0

Abstract

The present master's thesis in a business environment was conducted in the Toyota Caetano Ovar plant in the maintenance department, with the main goal of developing new proposals for managing and optimising the water supply system.

The system had several problems, such as a lack of maintenance to the water supply system pipes, no careful and critical analysis of recorded consumptions, a lack of predictive consumption analysis methods and a missing poka-yoke (which means "mistake-proofing" or "error prevention") system to warn of possible leaks in the water supply system. All of these problems could endanger the factory's production because the water supply system is a piece of critical equipment.

To overcome these issues, the dissertation focused on three key objectives: creating a dashboard for water consumption recordings and data visualization, developing a predictive model for industrial water consumption (in R), and establishing a digitalization strategy for the water supply system.

All of these goals were achieved with the help of a comprehensive review of the literature on all relevant topics, research on the established prediction models and their use in R, and exploration of the available technology and potential applications for the factory's current water-related equipment.

Acknowledgments

I can't help but express my gratitude to those who helped carry out the work presented here.

First and foremost, I want to express gratitude to my family. I would not be the person I am now without everyone's understanding, support, and unconditional love. Thank you for making all of the sacrifices that allowed me to write my dissertation and for teaching me that nothing is impossible and that it is never too late to seek and fight for our dreams.

To my friends, who are more or less present, support me and contribute much to being the person I am and achieving my goals.

To everyone in the TCAP maintenance department, but especially to the three staff members. Engineer Antonio Veiga, for believing in my abilities to face this challenge, and Engineer Carlos Aleixo and Mr. Victor Dias, for guiding me on this long five-month journey. For all the accompaniment, advice, opportunities, willingness demonstrated in the development of this work, and for always being available to help me tirelessly.

Last but not least, I must express my gratitude to my faculty supervisors, Professor José Machado and Professor Marco Parente, for their diligent assistance in the writing of this dissertation, even in a short period of time.

*“Sempre chega a hora em que descobrimos que
sabíamos muito mais do que antes julgávamos.”*

José Saramago

Introduction

1.1 Context

The present master's thesis in a business environment lasted 5 months. It occurred within the scope of the Master in Mechanical Engineering of the Faculty of Engineering of the University of Porto. It was conducted in the Toyota Caetano Ovar plant in the maintenance department, with the main goal of developing new proposals for managing and optimising the water supply system.

For the factory to maintain its productive process stable, several pieces of equipment are key to ensure that such is achieved. Those pieces of equipment are named critical equipment. According to the company, this type of equipment or machinery can be described as equipment or process, which in its operational absence, causes the plant to stop or jeopardizes the safety of the plant or its people. Alexis and Rounds (2016) also refer to critical equipment as any machinery or equipment that can severely hinder the capacity to fulfil corporate objectives safely, adversely affect quality levels and violate environmental standards of the organization when not fully operating. Table 1.1 shows which equipment is considered, according to the company, as critical equipment.

Table 1.1: Critical equipment

CRITICAL EQUIPMENTS			
Boiler	Cooling Group (Chiller)	Generator	Demineralised Water (DI)
Industrial Water	Heat System (HVAC)	Waste Water (ETAR)	Compressors
Cooling Towers	Substation	Busbar	Electric Switchboards

The water supply system is crucial for vehicle production since car painting requires demineralized and industrial water. Any interruption in delivering this sort of water would jeopardize the factory's productivity.

Figure 1.1 depicts annual water consumption as well as total vehicle manufacturing from 2018 to 2022. In only the previous three years, there has been a 46 percent rise in water usage while car manufacturing has decreased. This is a concerning statistic: how could a year, when the COVID-19 was raging, have a higher production and nearly half the usage when compared to 2022?

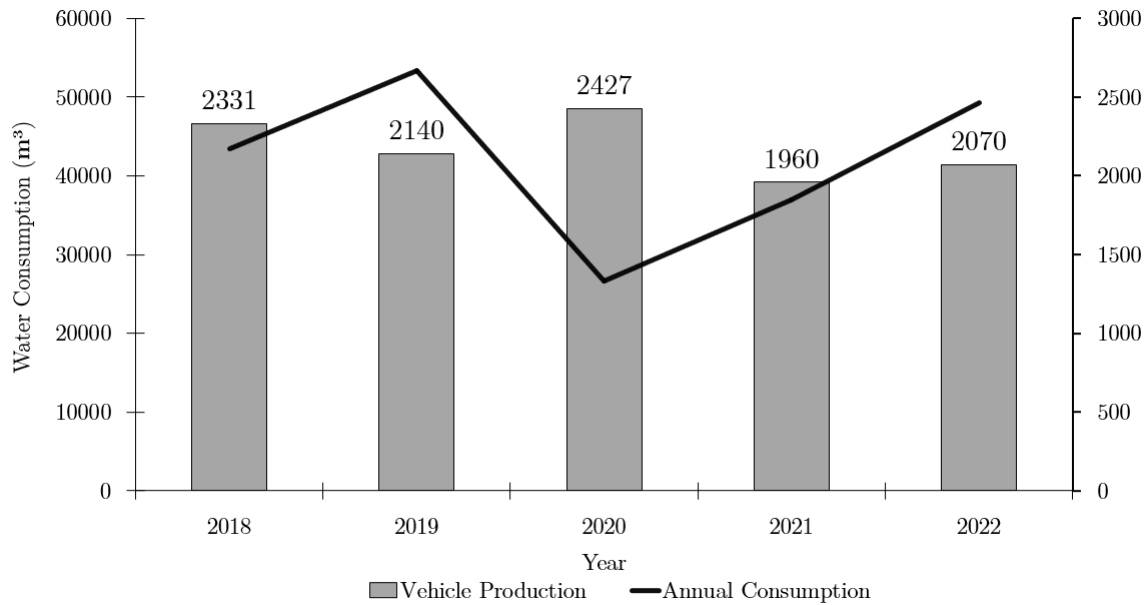


Figure 1.1: Annual Water consumption and vehicle production - from 2018 to 2022

In addition, the board of directors has determined that water consumption must be reduced for the corporation to accomplish the goals specified by the Toyota Environmental Challenge 2050: minimize water waste and consumption to the greatest extent possible.

The current state of the water consumption emphasizes the significance of this master's dissertation in establishing innovative methods for managing and optimising the water supply system.

1.2 The Company

Salvador Caetano is a Portuguese automaker founded in 1946 and based in Gaia, Portugal. The company is known for its partnerships with major automakers, including Toyota, and for producing vehicles under license for this brand (Salvador Caetano, 2021).

The Salvador Caetano factory in Ovar, Portugal, has a long history of producing vehicles for the Portuguese market. The plant was established in the 1970s and has operated ever since. Over the years, the factory has produced a wide range of vehicles, including cars, buses, and commercial vehicles. Throughout its history, the factory has been a major contributor to the Portuguese economy, providing jobs and helping to spur economic growth in the region. The plant has also been at the forefront of technological innovation in the automotive industry, and it continues to play a key role in developing Portugal's automotive sector (Toyota Portugal, 2008). Currently, the main production consists of the Land Cruiser model - LC70, mainly exported and widely popular in South Africa.



Figure 1.2: Toyota Caetano Portugal, S.A (Toyota Portugal, 2008)

1.3 Key Objectives

The main objectives to be achieved at the end of this master’s dissertation are:

- Analyse and evaluate the current state of the industrial water supply system.
- Creation of a dashboard for visualization and monitoring of the water meter record and KPIs related to industrial water.
- Creation of a predictive model for industrial water consumption in the factory.
- Development of a digitalization plan for integrated water metering recording.
- Proposal of new ideas for optimization of the water supply system regarding the facilities.

1.4 Methodology

An initial literature review was conducted for each subject to achieve the key objectives outlined in the preceding section.

As the author had no prior knowledge of *PowerBI*, he underwent a quick online tutorial on the foundations of the software to develop the dashboard. A similar approach was taken for learning the basics of coding and data analytics in R, as it was needed to create predictive models.

Several management and quality tools were employed to analyse the current state of the industrial water supply system, such as Genba Walks, Ishikawa, and 5S, among others.

Annex A depicts the Gantt chart of this master dissertation.

1.5 Dissertation Structure

This master dissertation will be divided into the following structure:

- **Chapter 1:** Brief introduction and context about the company, as well as the presentation of the objectives and methodology of the dissertation.
- **Chapter 2:** Literature review provided in accordance with the dissertation's main subjects.
- **Chapter 3:** Presentation, analysis and critical evaluation of the industrial water supply system (with an *AS-IS* approach).
- **Chapter 4:** Development of the industrial water dashboard is explained.
- **Chapter 5:** Research on industrial water consumption prediction models (model training, testing and evaluation according to the established metrics).
- **Chapter 6:** Presentation of potential enhancements to the industrial water supply system (using a *TO-DO*, *TO-BE* strategy) as well as the supply system's digitization plan.
- **Chapter 7:** Conclusions and future work.
- **Annexes:** Presentation of all the necessary documentation and code for a full understanding of the work done in this dissertation.

Literature Review

In this chapter, a literature review will be discussed in accordance with the main subjects of this master's dissertation.

2.1 Toyota Production System

Toyota Production System (TPS) is a management concept and set of techniques created by Toyota in order to minimize waste and maximize efficiency. It is a socio-technical system that organizes manufacturing and logistics to reduce costs and waste. It is sometimes referred to as "Lean Manufacturing" or "Just In Time" production (Toyota Europe, 2022). In the 1950s and 1960s, Taiichi Ohno oversaw the creation of the TPS, which is maintained and enhanced via iterations of standardized work and kaizen based on PDCA - Plan, Do, Check, Act (Lean Enterprise Institute, 2009). The main pillars of the TPS are presented in Figure 2.1.

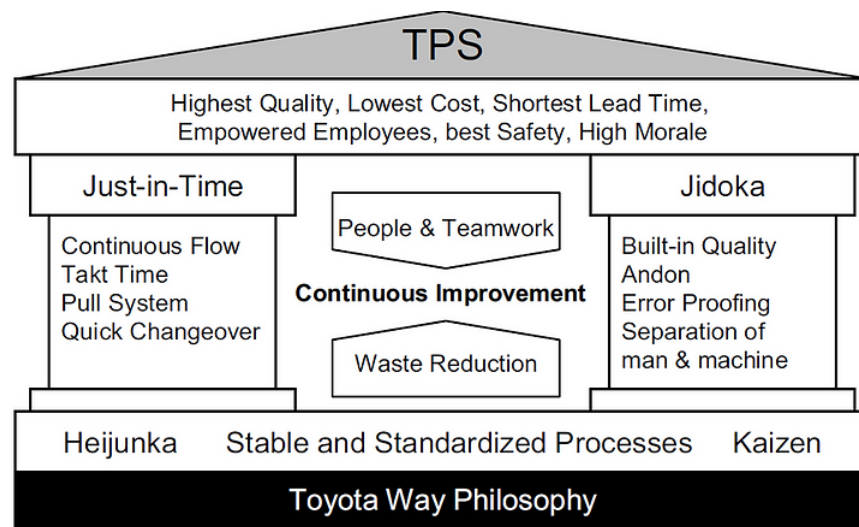


Figure 2.1: The main pillars of the TPS (S.J.M, 2021)

2.1.1 Toyota Production System Pillars

The TPS is maintained by twelve important pillars of production management:

- **Just-In-Time:** Just-in-Time (JIT) manufacturing entails producing only what is required, when, and in the required amount. It is necessary to keep the production line stocked with all the necessary parts, and the Kanban (Signboard) system provides an automated, real-time means of replenishing parts at the line side (Toyota Europe, 2022).
- **Jidoka:** Jidoka is a method of identifying issues and acting quickly to remedy them. If there is an issue on the shop floor, employees have the authority to halt the line and relay information to the andon (signal) display board. Automated processes and machinery are designed to identify problems and safely halt, allowing for examination and modifications. This helps keep problems from recurring and maintains quality and productivity (Toyota Europe, 2022).
- **Kaizen:** Kaizen is a concept that promotes optimum quality, waste elimination, and increased efficiency in equipment and work practices. It entails constantly following processes and spotting issues quickly, resulting in enhanced output (Toyota UK, 2013g).
- **Konnyaku Stone:** In the Toyota Production System, a small but crucial instrument called the Konnyaku Stone is employed to remove minor flaws from body panels before painting (Toyota UK, 2016c).
- **Poka-Yoke:** Poka-Yoke, from the Japanese words poka (to avoid) and yokeru (mistakes), is any step in the production process that aids a Toyota employee in avoiding missteps. Its goal is to remove flaws by stopping, fixing, or pointing out problems as they happen. It may also be used to describe any behaviour-shaping restriction incorporated into a process to prevent user error (Toyota UK, 2016d).
- **Hansei:** Hansei is the process of recognizing errors and taking the necessary steps to prevent a recurrence. Toyota understands the need for hansei-kai, or reflection meetings, to identify problems encountered along the road and develop precise strategies for the next initiatives. A failure to see problems is interpreted as a sign of failing to stretch to meet or surpass expectations, failing to analyze problems critically or objectively, or lacking modesty and humility. No issue is an issue in and of itself (Toyota UK, 2016b).
- **Andon:** Andon is a tool that visually illustrates where the action is required. It is used to implement the Jidoka principle, which calls for identifying an issue as it arises in order to take quick action to address it and stop reoccurring. It is an illuminated indication that alerts others to a problem in the quality control or production streams. It got its name from the term for a paper lantern (Toyota UK, 2016a).

- **Heijunka:** Heijunka is a method for levelling production across all departments and suppliers over time to support JIT manufacturing. It is crucial for production sequencing as well (Toyota UK, 2013c).
- **Genchi Genbutsu:** Genchi Genbutsu is the most effective method for finding rapid and effective solutions to issues. High levels of management must be present on the manufacturing floor for any issues to be properly understood before being resolved. The expression is more about a personal awareness of the entire effects of any activity within an area than physically visiting an area (Toyota UK, 2013b).
- **Nemawashi:** Nemawashi is the initial stage of the decision-making process, during which a corporation asks employees for their opinions to include them in the process. It requires information exchange and consensus building (Toyota UK, 2013f).
- **Kanban:** Kanban is a mechanism for communicating information across operations and automatically ordering parts as they are depleted. A simple, sign-based scheduling system initiates and manages the production logistical chain. Kanban displays may transmit a wide range of data, from stock levels to production numbers. In its most basic form, a kanban board will display goods in, products in production, and goods out (Toyota UK, 2013d).
- **Muda:** Muda is a non-value-added activity that is eliminated in the Toyota Production System's major pillar, the Just-In-Time system. It includes excessive finance charges, storage fees, and obsolete inventory (Toyota UK, 2013e).
- **Muri:** Muri, which refers to the JIT system's equally spread production in assembly processes, aims to reduce equipment and personnel overuse (Toyota UK, 2013e).
- **Mura:** One of the key tenets of the JIT system, the cornerstone of the Toyota Production System, is the concept of mura, which consists in removing unevenness or anomalies in the production process (Toyota UK, 2013e).
- **Genba:** In management, the term refers to the workplace or the location where value is added. It often refers to the shop floor in manufacturing (Toyota UK, 2013a).

2.1.2 Toyota Production System Management Tools

To achieve the best production efficiency and continuous improvement, the TPS relies on several management tools to maintain its process stable. These are going to be reviewed in the next section.

2.1.2.1 5S

The 5S operating model is a tool of the Lean concept, which is widely recognized as the cornerstone of lean manufacturing, as it is the basis for many lean production techniques, as seen in Figure 2.2. It assists manufacturing companies in establishing the system and operational stability required for effective continuous improvement initiatives. To increase workplace efficiency, the 5S pillars strive to identify and remove wastes connected with industrial systems (Singh and Ahuja, 2015). This method consists of 5 pillars, presented in Figure 2.3.

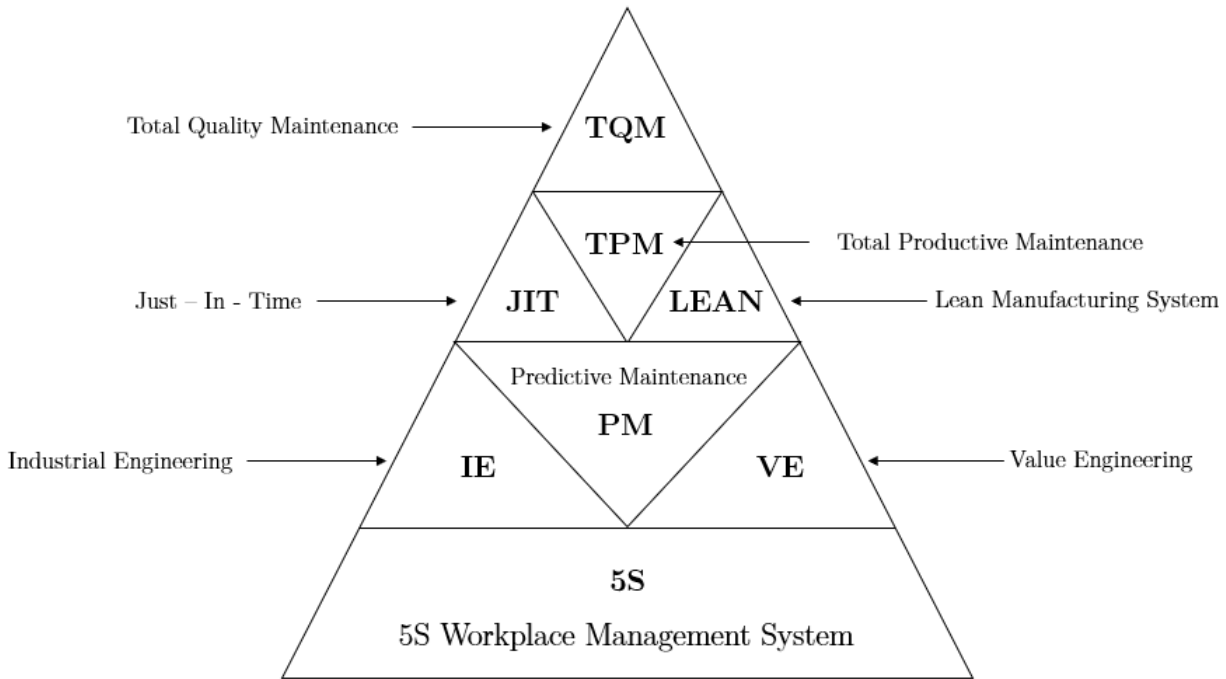


Figure 2.2: 5S as a starting point for multiple lean production techniques (Adapted from Singh and Ahuja (2015))

- **Seiri:** Seiri refers to the organization which involves arranging things in accordance with certain standards or principles. This entails discriminating between important and extraneous things in order to establish an efficient system. Unneeded things are either kept elsewhere or removed, resulting in fewer risks and less clutter (Singh and Ahuja, 2015).
- **Seiton:** Seiton, which means neatness, attempts to put items in the appropriate area or arrangement so that people may swiftly access or use anything they need. It entails prioritizing the requirement and significance of goods/equipment to maximize the convenience of location. This activity entails ensuring that all things in the workplace

have specified areas, allowing staff to control activities efficiently, and assisting them in planning materials, supplies, or tool requirements (Singh and Ahuja, 2015).

- **Seiso:** Seiso is a self-inspection, cleanliness, and optimal workplace method. It consists of three key activities: cleaning the workplace, preserving its attractiveness, and implementing preventative measures. Shining at work removes dirt, dust, fluids, and other contaminants. 5S teams may paint or coat work surfaces, equipment, floors, and walls to demarcate different activity zones. Active machine self-inspection and incorporating quality into each work process extend machine life and performance while avoiding critical breakdowns or downtime in operations (Singh and Ahuja, 2015).
- **Seiketsu:** The fourth S, Seiketsu, is for standardization and refers to repeating Seiri-Seiton-Seiso to maintain a comfortable and effective workplace. Creating a standard operating procedure and standard work instructions, including colour coding, flowcharts, checklists, and labelling, are part of the implementation phase. This supports a consistent strategy throughout the organization (Singh and Ahuja, 2015).
- **Shitsuke:** Shitsuke, which means sustain, is vital for 5S implementation and necessitates a constant auditing procedure to maintain the program's performance improvement (Singh and Ahuja, 2015).

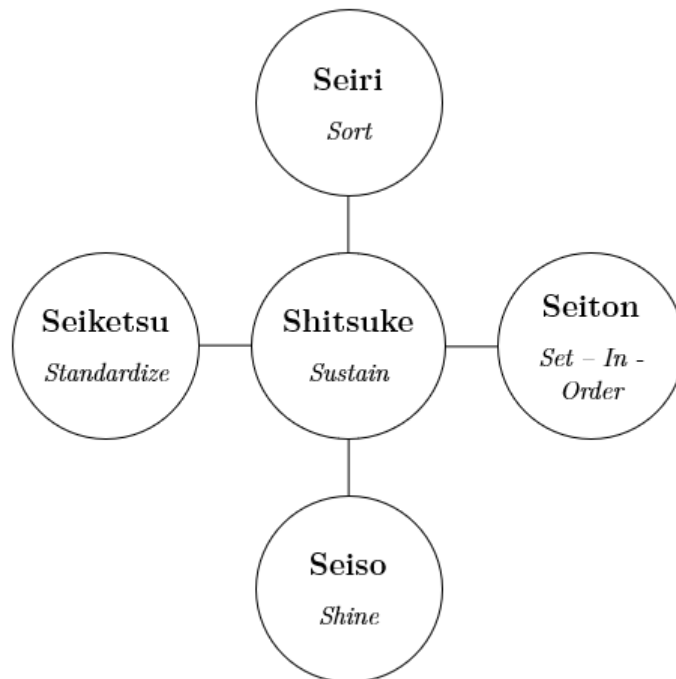


Figure 2.3: 5S Diagram (Adapted from a Rhaffor et al. (2019))

2.1.2.2 PDCA - Plan-Do-Check-Act

Dr Edwards Deming introduced the Plan-Do-Check-Act (PDCA) cycle, presented in Figure 2.4, as a quality management approach in 1950. It is utilized in companies and industrial fields, including manufacturing, services, offshore, project areas, and manufacturing. The PDCA approach, which is more future-focused, adaptable, rational, and reasonable, is used to continuously progress without pausing. It is more adaptable, rational, and sensible and describes every component of the drafted plan (Isniah et al., 2020).

This problem-solving cycle consists of four stages that are repeated until there's a final solution (Isniah et al., 2020):

Plan: A plan establishes objectives and procedures to accomplish particular results.

Do: Put the new method to the test by attempting to attain the goals.

Check: The inspection process stages have been monitored and evaluated according to specifications.

Act: Actions are made to enhance outcomes and achieve or meet criteria.

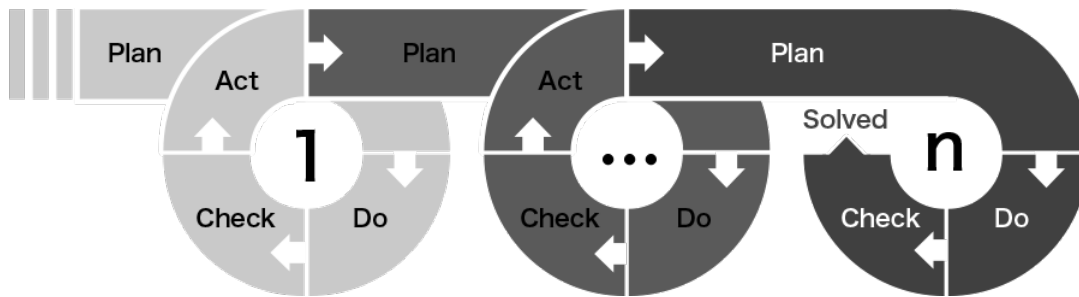


Figure 2.4: PDCA cycle approach (Scott Curtis, 2022)

2.1.2.3 The Seven Quality Tools

When it comes to quality control, seven tools are employed very often in the TPS:

- Ishikawa/Fishbone Diagram.
- Pareto charts.
- Control Charts.
- Scatter diagrams.
- Histograms.
- Flowcharts.

- Checklists.

One important tool that was mostly employed during this dissertation was the Ishikawa diagram, which was created by Kaoru Ishikawa, an engineering professor at the University of Tokyo. It helped establish quality-management practices in the Kawasaki shipyards in the 1960s. He invented Ishikawa diagrams in 1943 as a cause-effect analysis aid. The summative impact or result of the data shown in the fish skeleton is represented by the head of the fish in the fishbone graphic. Each of the fish's ribs stands for a reason that influences the outcome. Each item reason might be followed by a series of specific instances on the horizontal branches. Sub-branches can be used to list the causes or traits of each example. Fishbone diagrams can therefore be more intricate images with several levels of detail or rather basic visual organizers (Clary and Wandersee, 2010).

It must be noted that the company follows an adapted variant of this diagram, most known as the 4S Ishikawa, as seen in Figure 2.5. The 4S diagram only comprises four fishbones instead of the usual six: Man, Material, Method, Machine, Measurement, and Environment. Only Man, Material, Method and Machine are employed in the company's diagram.

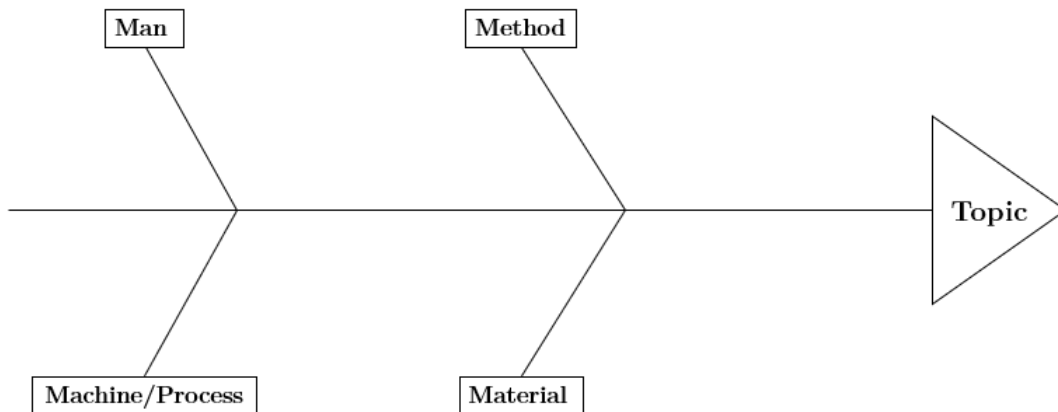


Figure 2.5: Ishikawa Diagram used in the company

2.1.2.4 A3 Problem - Solving

An A3 Report is a Toyota-invented method of writing out the issue, analysis, remedial measures, and action plan on a single big A3 paper. It is a common way of summarizing problem-solving exercises, status reports, and planning activities like value-stream mapping. It directs debate and analysis by defining the existing situation, the nature of the problem, the range of viable solutions, the best solution, the means to put it into action, and proof that the problem has been handled (Lean Enterprise Institute, 2022). Annex B of this dissertation presents a template of the A3 used in the company.

2.2 Water - The Key To Life On Earth

Water is an essential element that is vital to maintaining life on Earth. It covers nearly 71% of the planet's surface, has no taste or smell, and is transparent. Water also makes up the majority of cells, tissues, and body fluids, making it the fundamental component of all living things.

For the sustainable management and preservation of this essential resource, it is important to comprehend the qualities, distribution, and conservation of water. For the welfare of both the present and future generations, preserving water, enhancing water quality, and providing fair access to water resources is imperative.

According to the latest report published by UNESCO, *The United Nations World Water Development Report 2023*, over the past 40 years, water use has been rising globally at a rate of about 1% annually. Most of this growth is centred on developing economies and middle and lower-income nations. Population expansion, socioeconomic advancement, and shifting purchasing habits contributed to this trend (United Nations Educational, Scientific and Cultural Organization, 2023).

The evolution of water demand is extremely site-specific and reflects changing usage patterns in the three main water-using sectors: agriculture, industry, and municipalities. The development of water supply and sanitation services is the main driver of growth in the municipal industry. Water-intensive operations, such as manufacturing and energy generation, are what fuel the industrial water demand. Irrigation, which can vary according to the soil, weather variables, crop type and field uses, is what drives the need for water in agriculture (United Nations Educational, Scientific and Cultural Organization, 2023).

However, significant regional and local differences are hidden by the global statistics, as Figures 2.6 and 2.7 demonstrate: for example, Europe's demand is much more concentrated in its industries, while South Asia is keener on the water for its agricultural demands.

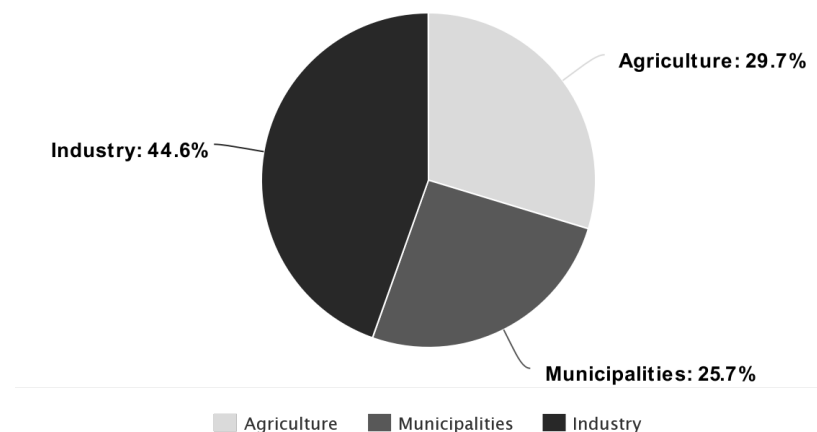


Figure 2.6: Europe's water demand (United Nations Educational, Scientific and Cultural Organization, 2023)

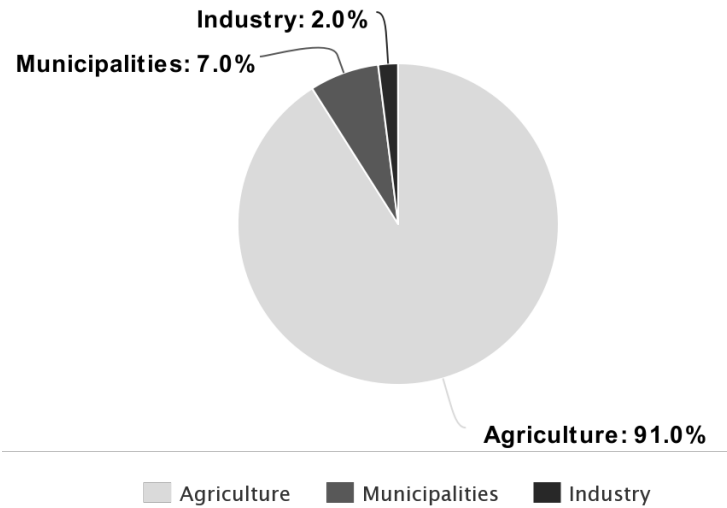


Figure 2.7: South Asia's water demand (United Nations Educational, Scientific and Cultural Organization, 2023)

One of the main problems regarding water supply is its availability. According to geological and climatic conditions, availability varies greatly on a local and regional level. Due to the local effects of physical water stress and the escalation and spread of freshwater contamination, water scarcity is starting to become endemic (United Nations Educational, Scientific and Cultural Organization, 2023). Figure 2.8 reassures such a claim.

The main effect of shortage is increased groundwater consumption and depletion. Groundwater storage depletion is expected to occur at a rate of 100 to 200 km³/year, accounting for 15 to 25% of total groundwater withdrawals (United Nations Educational, Scientific and Cultural Organization, 2023).

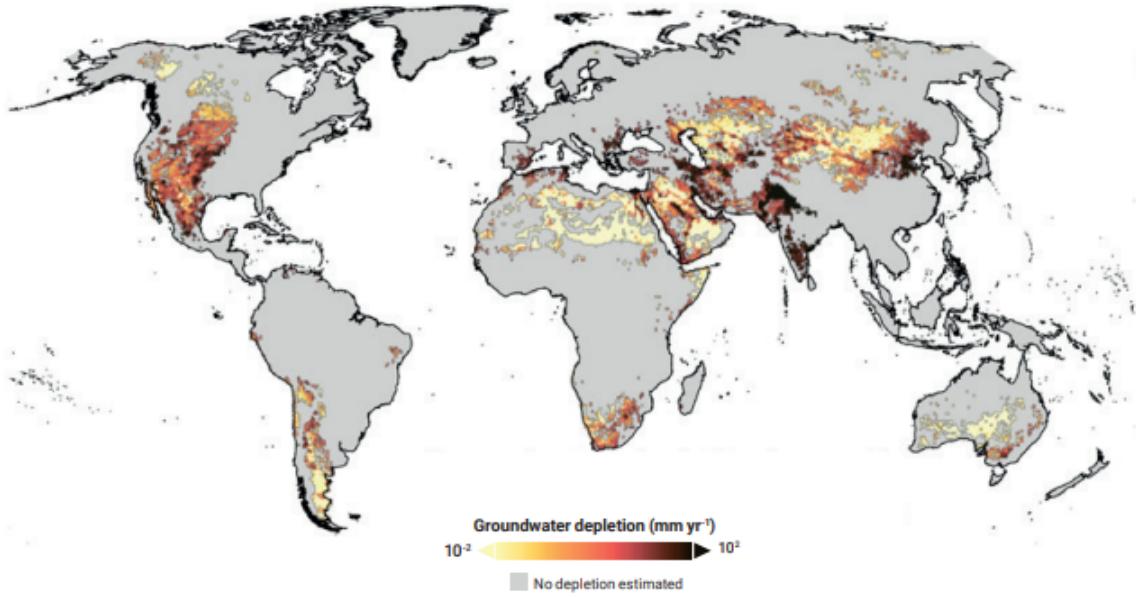


Figure 2.8: Groundwater depletion rate (United Nations Educational, Scientific and Cultural Organization, 2023)

These figures reveal a dangerous prospect regarding the future of water supply and reiterate the importance of taking measures much faster than what is being done nowadays by top governmental and international entities.

2.3 Water Consumption In The Automotive Industry

The worldwide automobile sector consumes significant water in numerous manufacturing processes. According to some estimates, the manufacture of an automobile consumes more than 150 m³ of water, and whether tyre production is included varies depending on the estimate (Isaiah, 2014).

Metal finishing activities are known for high water consumption among the different automobile production processes that consume water, but water consumption is especially significant in the paint shop. While the painting process itself consumes a significant amount of water, for water-based automotive coatings and paints, processing equipment must be cleaned with water on a regular basis (Isaiah, 2014).

The United Kingdom's (UK) Society of Motor Manufacturers and Traders (SMMT) declared in its 2014 automobile sustainability report that water usage in automobile manufacturing increased by 5.1% in the country last year, owing mostly to increased output. In the

UK, per-vehicle water consumption has increased by 3.9%. Painting process improvements account for most water use in automobile production, and the development of closed-loop systems has since greatly decreased the quantity of water needed per car (Isaiah, 2014).

The European Automobile Manufacturers Association (ACEA), in an infographic presented in Figure 2.9, published in 2022, evaluated the water used in car production in the European Union from 2006 to 2021. Their conclusion was staggering: there was a decrease of 30.5% in the water per car produced and almost 50% in the total water consumption (ACEA, 2022). Nonetheless, one must note the slight increase in the water per unit produced from 2019 to 2021.

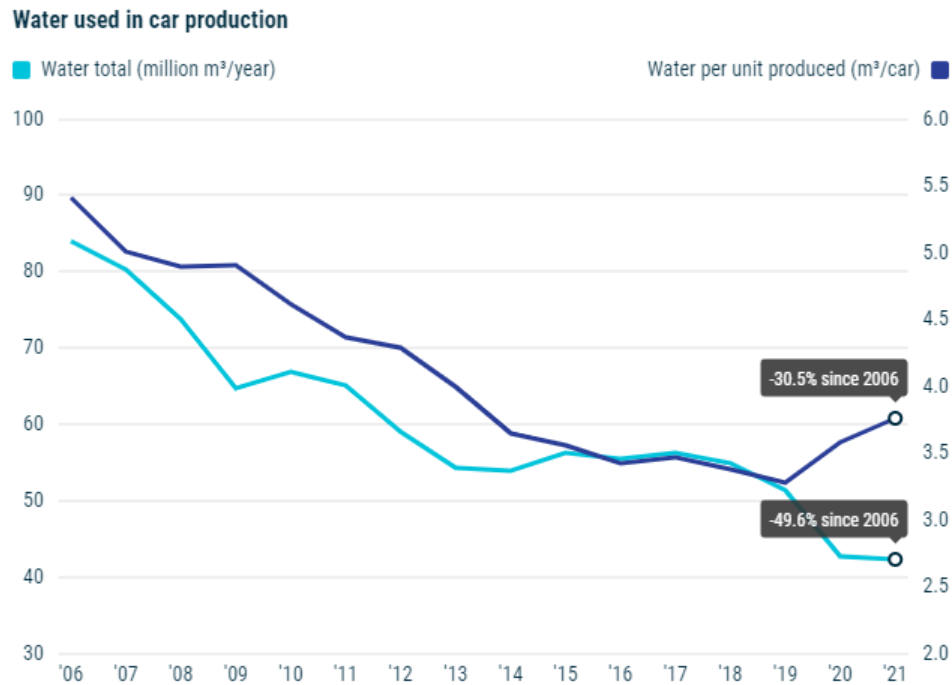


Figure 2.9: Water used in car production in the European Union (ACEA, 2022)

Semmens et al. (2014) assessed the life cycle of the water consumption of a passenger vehicle and determined that most of the water is utilized in the "Use Phase" of the vehicle. Table 2.1 shows the average values in litres, for each life cycle phase, of water consumed.

Table 2.1: Life cycle consumption of a passenger vehicle (Semmens et al., 2014)

Life Cycle Phase	Water Consumption (l)	Percentage
Material Production	5569	9.4
Parts Production	902	1.5
Vehicle Assembly	670	1.1
Use Phase	51965	87.5
End Of Life	259	0.4
Total	59365	100

Semmens et al. (2014) also developed a study to analyse the water-related data of automotive manufacturers retrieved from their sustainability reports. Seven manufacturers were involved in this report: BMW, Daimler, Ford, GM, Kia, Toyota and Volkswagen. The authors establish, before analyzing and evaluating the data, the following definitions:

- **Water Use:** defined as all water entering a system.

$$Water_{Use}[m^3/Vehicle] = \frac{Water_{Input}}{Production} \quad (2.1)$$

- **Water Consumption:** Freshwater withdrawals are evaporated or absorbed into goods and waste and hence do not return to the source of origin.

$$Water_{Consumption}[m^3/Vehicle] = \frac{Water_{Input} - Water_{Discharge}}{Production} \quad (2.2)$$

- **Direct Water Consumption:** Similar to direct materials or prices, it may be easily connected to a single industrial activity or process.
- **Indirect Water Consumption:** Refers to water consumed by items or processes within the Original Equipment Manufacturers' (OEM) control but outside of the vehicle production process.

After analyzing all the environmental reports, the authors summarised each brand's water use and consumption values in Table 2.2.

2.3. Water Consumption In The Automotive Industry

Table 2.2: Reported and calculated water use and consumption data by OEM on a per-unit basis (Semmens et al., 2014)

OEM	Use (m ³ /vehicle)					Consumption (m ³ /vehicle)					Activity
	2006	2007	2008	2009	2010	2006	2007	2008	2009	2010	
BMW	2.99 ^a	2.61 ^a	2.56 ^a	2.56 ^a	2.31 ^a	–	–	–	–	0.67	Mfg
Volkswagen	5.44 ^a	5.38 ^a	5.34 ^a	5.40 ^a	5.01 ^a	1.72 ^a	1.38 ^a	1.46 ^a	1.48 ^a	1.19 ^a	Mfg
Nissan	6.90	6.65	6.16	5.29	6.38	1.83	1.69	1.45	1.76	1.50	Mfg
Chrysler	–	–	–	–	6.44	–	–	–	–	–	Mfg
	–	–	–	–	0.36	–	–	–	–	–	Nonmfg
	–	–	–	–	6.80	–	–	–	–	2.31	Co-wide
Honda 2011	–	–	–	–	–	–	–	2.96	3.92	3.21	Co-wide
Honda 2010	–	–	–	–	–	–	–	–	–	7.16	Mfg
	–	–	–	–	–	–	–	–	–	1.98	Nonmfg
	–	–	–	–	–	4.99	5.79	6.77	11.22	9.14	Co-wide
Toyota 2011	–	–	–	–	–	–	3.60 ^a	3.80 ^a	3.60 ^a	3.70 ^a	Co-wide
Toyota 2010	–	–	–	–	–	4.10 ^a	4.10 ^a	4.60 ^a	4.40 ^a	–	Mfg
Toyota Australia	–	–	–	–	–	4.23	3.72	3.02	3.56	3.84	Mfg
	–	–	–	–	–	–	0.38 ^a	0.23 ^a	0.43 ^a	0.27 ^a	Nonmfg
	–	–	–	–	–	–	4.10	3.25	3.99	4.11	Co-wide
Kia	–	–	–	–	–	–	–	4.80 ^a	4.90 ^a	4.90 ^a	Co-wide
Mazda	22.41	16.24	11.81	15.10	11.68	6.38	9.77	6.12	7.67	5.86	Co-wide
Daimler Cars	–	–	–	–	–	–	6.20 ^a	5.80 ^a	6.00 ^a	–	Mfg
Daimler 2008	–	–	–	–	–	7.26	7.24	–	–	–	Mfg
Daimler 2010	–	–	–	–	–	–	7.73	7.16	8.24	–	Mfg
Daimler 2011	–	–	–	–	–	–	–	7.29	8.16	7.26	Mfg
Daimler Trucks	–	–	–	–	–	–	13.10 ^a	12.20 ^a	18.10 ^a	–	Mfg
General Motors	–	–	–	–	4.70 ^a	–	–	–	–	–	Co-wide
Ford	7.60 ^a	6.20 ^a	5.60 ^a	5.20 ^a	4.80 ^a	–	–	–	–	–	Mfg
Hyundai	6.73	5.65	4.01	3.67	5.42	–	–	–	–	–	Co-wide

Mfg manufacturing, *Nonmfg* nonmanufacturing, *Co-wide* company-wide

^a Directly reported water data on a per-unit basis

At the end of the study, the authors concluded that the manufacturing average value for water use is 5.20 m³/vehicle, with a standard deviation of 2.83 m³/vehicle (1.25 m³/vehicle if Daimler and Mazda values not included). On the other hand, the manufacturing average value for the water consumption stands at 2.83 m³/vehicle, with a standard deviation of 3.47 m³/vehicle (0.52 m³/vehicle if Daimler and Mazda values not included). Table 2.3 summarizes the main conclusions from the report.

Table 2.3: Main conclusions from the study (Semmens et al., 2014) - values in parenthesis do not account Daimler and Mazda data

Parameter	Water Use ($m^3/vehicle$)	Water Consumption ($m^3/vehicle$)
Manufacturing Average	5.20	2.83 (1.25)
Manufacturing Standard Deviation	1.42	3.17 (0.52)
Company-Wide Average	9.11 (5.95)	4.86 (4.29)
Company-Wide Standard Average	5.55 (1.20)	1.86 (1.56)
Estimated Nonmanufacturing Average	3.91 (0.75)	2.03 (3.04)

2.4 Industry 4.0

The term Industry 4.0 (I4.0) was first introduced at the Hannover Fair in 2011 (Vogel-Heuser and Hess, 2016), and it can be considered the new industrial revolution of the 21st century. There are several definitions of I4.0. Nevertheless, according to the German Federal Government, it can be defined as a structure in which CPPS can be extensively used to promote a faster and automatic global exchange of information with enhanced communication capabilities, where production and business are matched (Vaidya et al., 2018). Figure 2.10 depicts a time-lapse of all the industrial revolutions since the 1800s and their main outcomes.

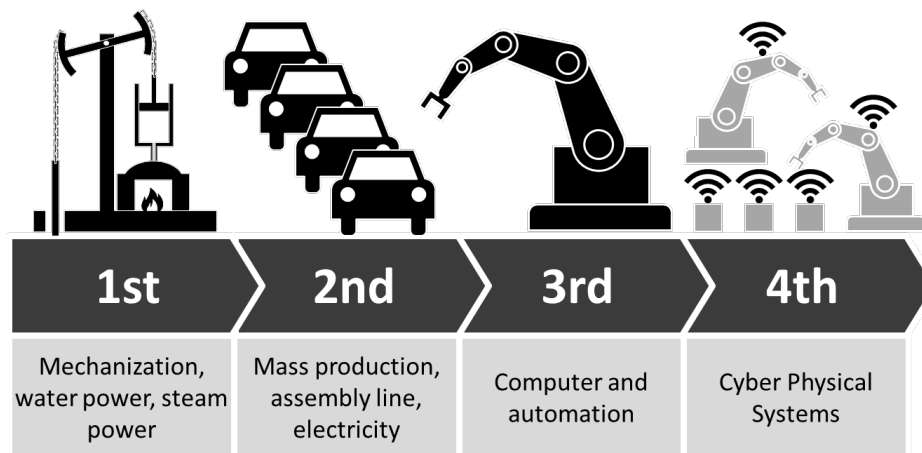


Figure 2.10: Timelapse of the industrial revolutions (Roser, 2015)

2.5 The main driving pillars of I4.0

According to Rüßmann et al. (2015) and Erboz (2017), the I4.0 is based on the following main driving pillars.

2.5.1 Big Data And Analytics

Ishwarappa and Anuradha (2015) define Big Data as "a collection of massive and complex data sets and data volume that include the huge quantities of data, data management capabilities, social media analytics and real-time data" (p.319). There are six dimensions in this type of data, as seen in Figure 2.11, mostly known as the 6V's of Big Data:

- **Volume:** Nowadays, the amount of data produced surpasses the threshold capacity of the typical gathering data tools, thus demanding a new approach (Witkowski, 2017).
- **Variety:** The information obtained mostly originates from transactional systems and social networking. Due to their dynamic nature and unstructured ability, conventional analytic approaches cannot be used to analyze them (Witkowski, 2017).
- **Velocity:** Due to the constant data flow, the new prediction algorithms must be able to analyze and conclude accurately in near real-time (Witkowski, 2017).
- **Value:** Analytics must be able to separate the most significant data from a large collection of it. This will ultimately result in improved outcomes (Witkowski, 2017).
- **Veracity:** A strategy for determining the trustworthiness of a large amount of data must be conceived since it's coming from uncontrolled environments (Andreu-Perez et al., 2015).
- **Variability:** One must evaluate the consistency of the data retrieved throughout the data analysis (Andreu-Perez et al., 2015).

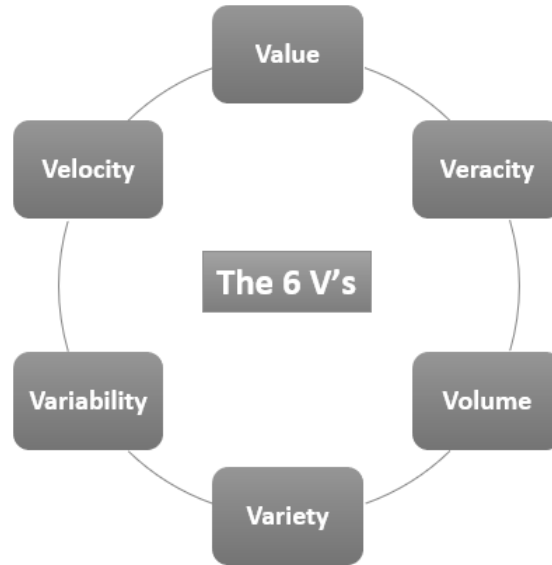


Figure 2.11: The 6 V's of Big Data

2.5.2 Autonomous Robots

Erboz (2017) indicates the importance of the new wave of autonomous robots regarding the continuous improvement policies in the I4.0. These could be implemented in several areas, such as production, logistics or distribution activities, with human-robot cooperation ability (Erboz, 2017).

Rüßmann et al. (2015) regards human-robot cooperation as one very positive outcome of I4.0. It will provide a much more flexible and enhanced working capability, decreasing cost compared to today's standards.

2.5.3 Simulation

Another important aspect regarding the I4.0 is the ability to simulate different aspects of the industry with the aid of the new digital simulation tools.

These new digital tools, with the capability of self-configuration, would allow the users to, for example, promote a dynamic analysis of a production system or even enable an effective shop-floor management (Erboz, 2017) (Uhlemann et al., 2017).

At the end of the day, the main goal is to promote an extension of the tools used in engineering development stages to the plant floors. One great example comes from Siemens, which created a virtual machine that can mimic component machining using data from an actual machine. This allowed a decrease in 80% of the total setup time for the machining process (Rüßmann et al., 2015).

2.5.4 Vertical and Horizontal System Integration

Vertical integration can be considered an entailing linking of all business divisions and operations inside a factory or firm. With this connectivity in place, data flows between business units and is made available to all of them. This encompasses all the departments of the company (McLaughlin, 2023).

This type of integration enables companies to have a smart factory with self-organization and a type of production that becomes client-specific. Also, maintenance management becomes much more efficient since every part or equipment is easily traceable, using smart sensors (Deloitte, 2014).

On the other hand, horizontal integration entails linking all components of the supply chain. This enhanced alignment boosts visibility, flexibility, and efficiency while increasing automation levels (McLaughlin, 2023). According to Deloitte (2014), this type of integration allows each part or product's history to be documented and accessible at any moment, providing continual traceability, mostly known as "product memory".

2.5.5 The Cloud

With the increasing need for data sharing related to production, with both systems inside and outside of the company, cloud computing has become a hot topic related to I4.0 (Rüßmann et al., 2015). The main assumption regarding cloud systems is the ability to facilitate the management and automation of the supply chains, for example (Erboz, 2017).

There are three types of cloud computing systems:

- **Software as a Service (SaaS):** Access is determined by the customer's purchase, such as Enterprise Resource Planning (Erboz, 2017).
- **Platform as a Service (PaaS):** Customers, even software developers, can access their cloud-based apps (Erboz, 2017).
- **Infrastructure as a Service (IaaS):** provides the fundamental functions, such storage capacity (Erboz, 2017).

2.5.6 Industrial Internet of Things

Pursuant to Erboz (2017), the Internet of Things (IoT) can be comprehended as the next technological evolution that provides solutions for computations and analytics while relying on cloud-based systems. This enables the engagement and communication between field devices and centralized controllers when needed and real-time responses by decentralizing analytics and decision-making (Rüßmann et al., 2015).

2.5.7 Augmented Reality

Augmented Reality (AR) is a variation of Virtual Environments, allowing the user to see the real world with virtual objects superimposed upon or composited with it (Azuma, 1997).

One of the main barriers to adopting AR was the lack of cost-affordable hardware. However, nowadays, with the established technology (smartphones, tablets, etc.), this is no longer a problem. Mobile devices provide the sensors and processing units needed to develop and deploy AR applications, and technology innovations lead to new products, most known as "wearables" (De Pace et al., 2018). This technology enhances human-machine interaction, remote control, and visual inspection by combining computer-generated graphics and physical objects, important aspects of I4.0. It also gives users motion control by using sensor technology (Erboz, 2017).

2.5.8 Cybersecurity

IoT has been integrated into I4.0's physical entry terminals, rendering them susceptible to hackers. This makes the network attractive for cybercrimes, which can have disastrous consequences like lost earnings, a decline in profitability, permanent fire damage, or a severe hazard to persons and assets. Value chains that combine IT-I4.0 with operational technology must have a comprehensive plan for handling cyber risk. Nonetheless, the cyber problems plaguing today's internet technologies provide consumers with significant challenges and hurdles. I4.0 might not fulfil its potential if these problems aren't addressed reasonably (Khalid et al., 2020).

2.6 Water 4.0

Water 4.0 (W4.0) is a project created by the German Water Partnership (GWP) in 2016 that defined the future paradigm for Germany's water management. The foundation of this project lies in the tools provided by the fast-paced development of I4.0, as seen in Figure 2.12, with the final goal of creating a new water management system that satisfies the needs according to the new global challenges, in an effective and efficient way (Bufler et al., 2019).

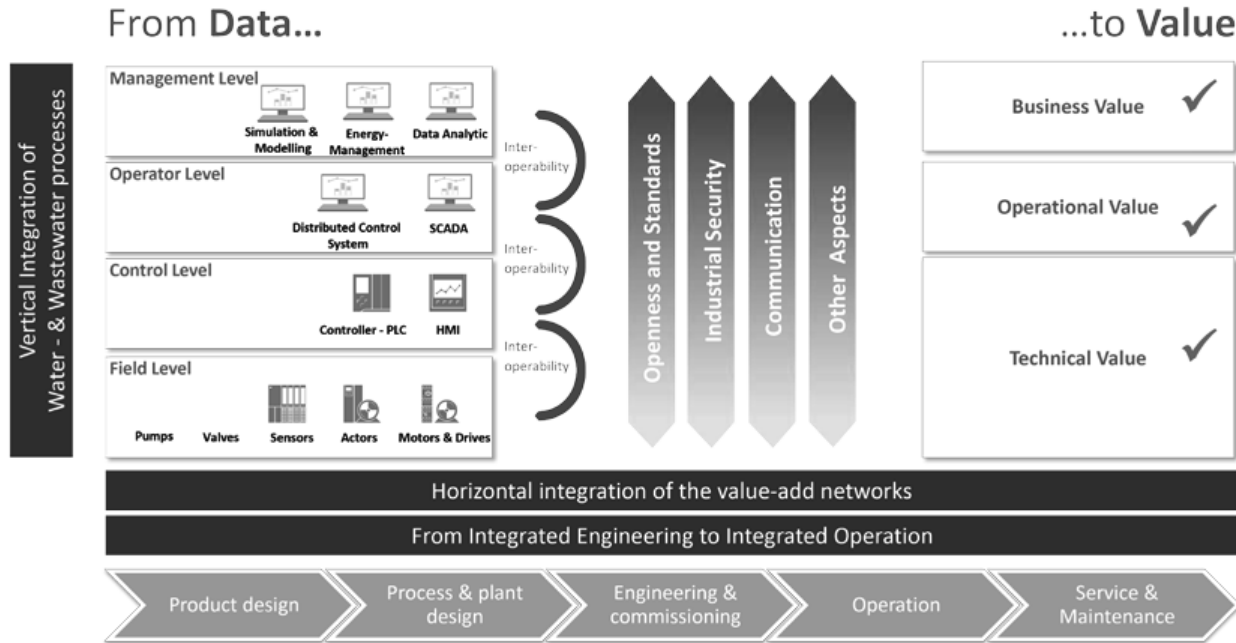


Figure 2.12: Digitization in water management (Bufler et al., 2019)

Regarding water revolutions, presented in Figure 2.13, local and *ad hoc* methods were used in the first water revolution. The second water revolution made extensive use of centralized water infrastructure. The third water revolution employed computers and control systems to connect site sensors with the information technology (IT) infrastructure. Finally, the fourth water revolution employs CPPS-enabled systems that integrate actual and virtual water systems (Omotayo et al., 2019).

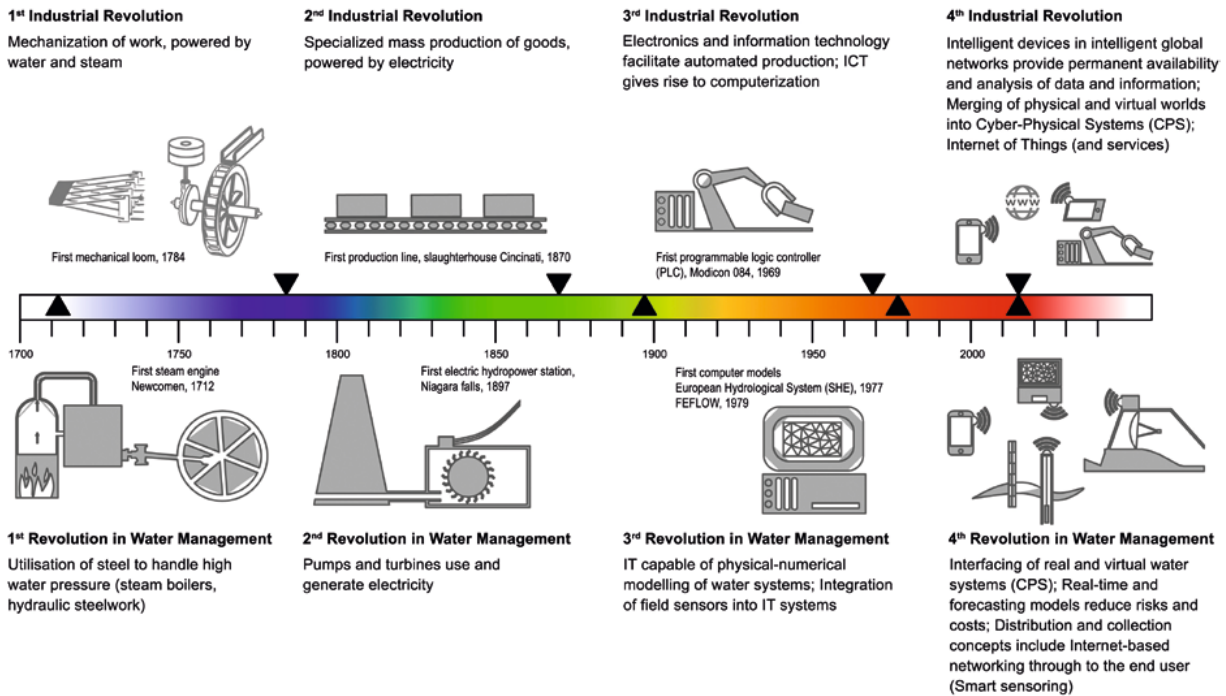


Figure 2.13: Comparison of the four industrial development stages (Bufler et al., 2019)

According to Omotayo et al. (2019), the most promising technologies of I4.0 than can be applied in W4.0 are the following:

- **Big Data and Analytics:** With the aid of analytics, the water industry will consequently become fully data-driven (Omotayo et al., 2019).
- **Cybersecurity:** With the digitalization of the water sector, the systems are more susceptible to suffering cyberattacks. Thus, it is of utter importance to decrease the security breach risk in these systems (Omotayo et al., 2019).
- **Cloud Platform:** Cloud technology is assisting the water business in storing, analyzing, and visualizing data acquired by sensors and pumps (Omotayo et al., 2019).
- **Internet of Things:** The Internet of Things is a potential technology for Water 4.0 since it allows water-related concerns to be remotely monitored in real-time (Omotayo et al., 2019).
- **Artificial Intelligence (AI) and Machine Learning (ML):** In the water business, AI and ML are being utilized to build leverage for quality water treatment by uncovering water-related difficulties and informing the service engineer in real-time (Omotayo et al., 2019).

2.6.1 Integrated Business Model

Omotayo et al. (2019) developed a system that enables the integration of W4.0 main characteristics in the industry - the *Integrated Business Model* (IBM). IBM has three main pillars: technologies, business stack of the systems and business strategy. These last two are known as "Water Balance".

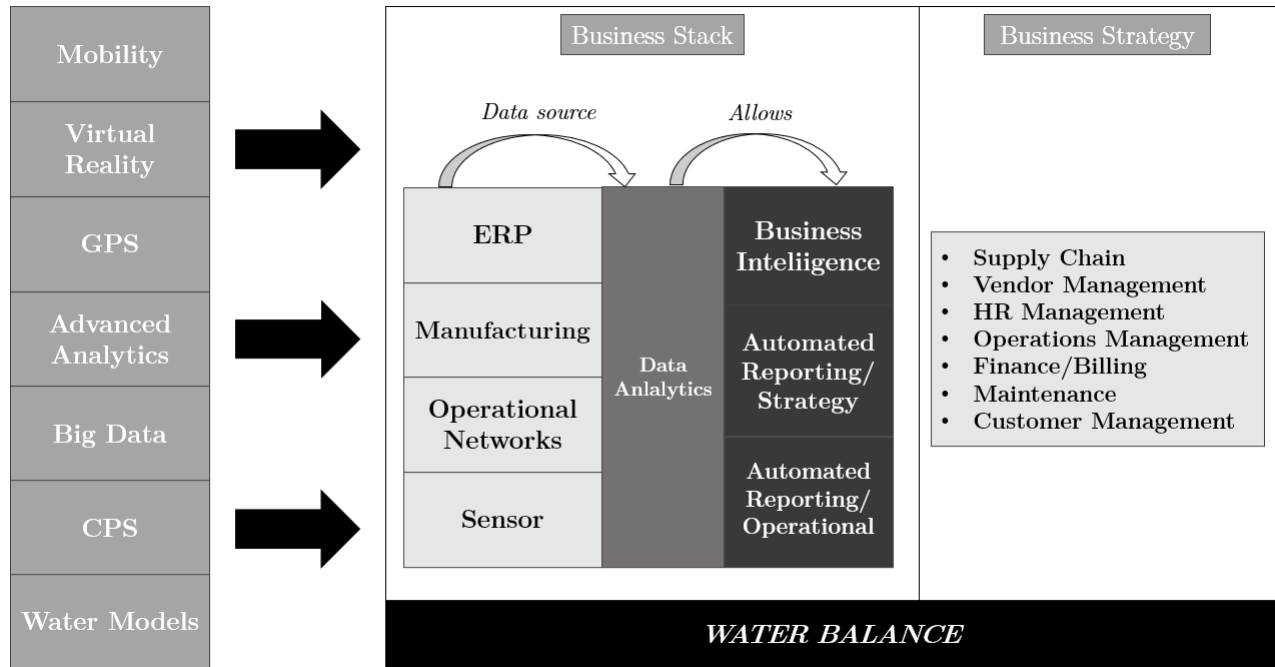


Figure 2.14: IBM functioning scheme (Adapted from Omotayo et al. (2019))

2.6.1.1 Technologies Component

Water businesses can apply these technologies to interact with the Business Stack of System components and evaluate data provided by new data analytic tools. They are not mandated to use all the technologies, but rather one or two that are effective for their work setting or system (Omotayo et al., 2019). A table with the main benefit of each technology is presented below.

Table 2.4: Main benefit of each technology for IBM (Omotayo et al., 2019)

Technologies Component	Benefits for IBM
Mobility	Wireless networks provide real-time communication and collaboration between machines and people, allowing data to be received and processed, as well as actions to be made.
Virtual Reality	Supply chain, pump station layout, asset management, and owner operators can interact in a virtual environment to discover problems and safety hazards prior to design, planning, implementation, and operation using virtual simulators.
GPS	GPS is a key technology in the water sector for precision monitoring and management, allowing access to water networks and spotting leak or pump problems.
Advanced Analytics	Water utilities may use advanced analytic techniques like ML and Deep Learning (DL) to handle water concerns including ageing infrastructure, climate change, and urbanization, as well as support the Water 4.0 paradigm.
Big Data	Big Data technology may assist the water business in gathering, monitoring, and analyzing data in real-time, assisting in cost savings, identifying concerns, and supporting accurate decisionmaking.
CPS	To perceive and communicate with the aquatic environment, sensors, actuators, and processors can be implanted.
Water Models	Using ML approaches, 4IR is assisting the water sector in developing realistic water models.

2.6.1.2 Business Stack of Systems Component

This area assists the water sector in the real-time management of their business system by combining the system and data created by the active systems into a more relevant insight. The "Business Stack of Systems" component includes data analytics, showing that data is critical to the Industry 4.0 revolution (Omotayo et al., 2019). Table 2.5 indicates the main benefits of each business stack of systems for IBM.

Table 2.5: Main benefit of each business stack of systems for IBM (Omotayo et al., 2019)

Business Stack Of Systems Component	Benefits for IBM
ERP	ERP systems are critical for handling structured business data in a CPS environment and decentralized system in the digital water sector.
Operational Networks	Monitor the operational management of water distribution networks and support the water business in generating, storing, and analyzing data in order to manage infrastructure, find problems, and avoid repeat occurrences.
Manufacturing System	Conduct actions and operations to provide high-quality water supply/services, as well as monitor and regulate data collected and aid in predictive maintenance.
Sensor Technology	BI applications are a technology-driven process that analyzes data and presents actionable information to help businesses make educated decisions, save money, expand their financial and market base, increase competitiveness, and support excellent business decisions.
Business Intelligence	BI applications help organizations save funds, expand their financial and market base, become more competitive, and make more informed decisions.
Automated Reporting/Strategic	Using automated data analytic tools, automated reporting/strategic assists the water sector in delivering appropriate business reports at precise moments and integrating data.
Automated Reporting/Operational	Allows computer systems, smart networks, and devices to do tasks without human involvement, reducing complexity, saving time and money, and improving accuracy.
Data Analytics	Data analysis is the foundation of a system's business stack, offering the capacity to analyze data to support the system's successful operation.

2.6.1.3 Business Strategy/Management Component

Business strategy and management are required for the Water 4.0 revolution to accomplish long-term goals. It is the process of developing, executing, and assessing cross-functional choices to accomplish desired outcomes or goals (Omotayo et al., 2019). The main benefits are presented in Table 2.6.

Table 2.6: Main benefit of each business strategy/management for IBM (Omotayo et al., 2019)

Business Strategy/ Management Component	Benefits for IBM
SCM	The digital supply chain management system will assist the water sector in becoming quicker, more effective, and precise while carrying out everyday operations by utilizing 4IR advancements.
Vendor Management	A digital platform for managing and optimizing connections with external suppliers for projects and business processes.
HR and Capacity Management	The digital HR system will assist the water sector in finding the ideal talents/graduates who are well-versed in Industry 4.0 technology. It would also assist in training new employees and establish a platform to retain talent.
Operations Management	Digitalized management system that helps to design and control water network operations, convert materials and labour into goods and services, and maximize profit for the organization.
Finance/Billing Management	Industry 4.0 technologies such as cloud, Big Data and Analytics, and Blockchain can help companies manage their financial resources to achieve their goals.
Maintenance Management	Sensor technologies could be connected to the CMMS to store and analyze data, and a predictive algorithm can be designed to handle the problem revealed by the sensor data.
CRM	CRM helps organizations manage relationships and communications with customers and potential customers through 4IR technologies such as cloud, IoT.

2.7 Predictive Forecasting Techniques

As seen in the previous section, data analytics is one of the key parameters in the water balance of the IBM model related to the business stack. Predictive forecasting techniques are extremely useful tools for analyzing past data and anticipating future trends and occurrences. These technologies employ statistical models, machine learning (ML) algorithms, and advanced data analysis to create accurate forecasts and projections. Predictive forecasting approaches seek to unearth hidden insights and identify future outcomes by evaluating data sets' patterns, correlations, and relationships. These methods include time series analysis, regression modelling, Artificial Neural Networks (ANN), and ensemble methods. By implementing these approaches, businesses and organizations may acquire significant insights into

customer behaviour, market trends, demand patterns, and financial performance. By applying predictive forecasting methodologies, decision-makers may make data-driven decisions, optimize resource allocation, manage risks, and grasp opportunities in a quickly changing and competitive world.

Makridakis et al. (2018) conducted a study comparing popular ML algorithms' performance for time series forecasting to classic statistical approaches. The paper compared the post-sample accuracy of eight classic statistical approaches with that of eight prominent ML algorithms using a large subset of 1045 monthly time series from the M3 Competition. The main accuracy measures that evaluated these methods were the following:

Symmetric Mean Absolute Scaled Error:

$$sMAPE = \frac{2}{k} \sum_{t=1}^k \frac{|Y_t - \hat{Y}_t|}{|Y_t| + |\hat{Y}_t|} \cdot 100\% \quad (2.3)$$

Mean Absolute Scaled Error:

$$MAPE = \frac{1}{k} \sum_{t=1}^k \frac{|Y_t - \hat{Y}_t|}{|Y_t| + |\hat{Y}_t|} \cdot 100\% \quad (2.4)$$

where,

k - forecasting horizon;

Y_t - actual observations;

$\hat{Y}_t$ - forecasts produced by the model at point t .

The study revealed, in accordance with Figure 2.15, that classical statistical approaches beat popular ML methods across all predicting horizons and accuracy criteria studied. Furthermore, the computing needs of ML approaches were far higher than those of statistical methods.

Several fundamental statistical definitions will be introduced briefly before presenting and discussing the basic statistical prediction models.

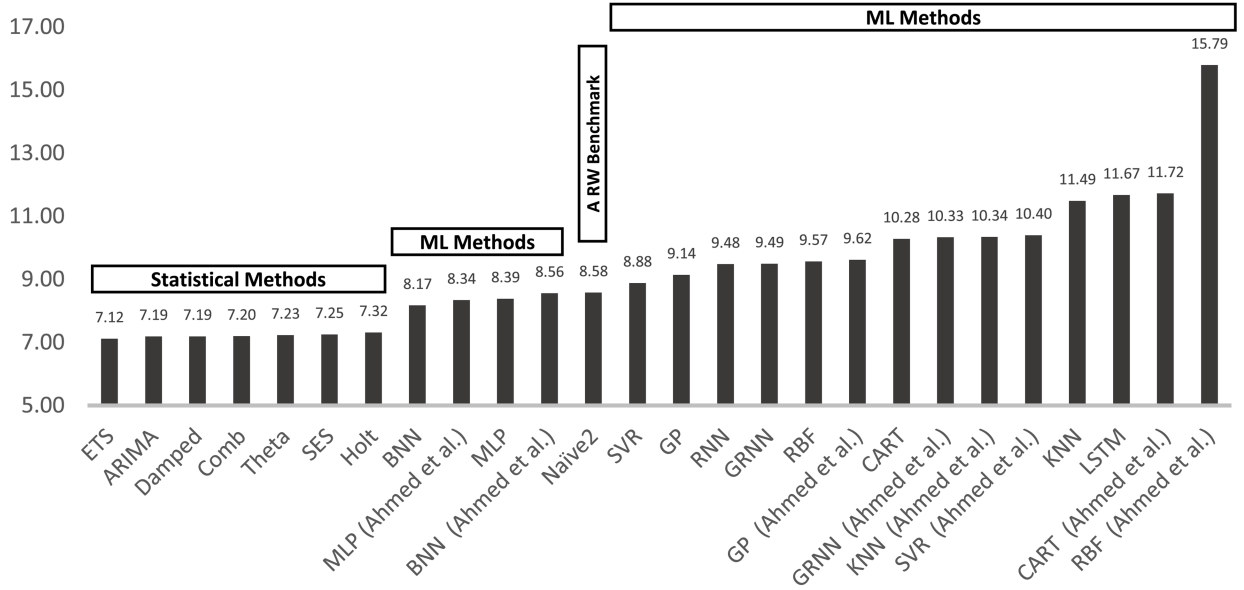


Figure 2.15: Forecasting performance (sMAPE) of the ML and statistical methods included in the study (Makridakis et al., 2018)

2.7.1 Statistical Measures For Univariate Datasets

Consider a dataset that consists of N data x_n , with $n = (1, 2, \dots, N)$, and a vector $(x_1^*, x_2^*, \dots, x_N^*)$ that orders the values of the dataset in ascending or descending order.

2.7.1.1 Measures of Location

There are three important measures of location in statistics: mean, median and mode. The three parameters can be calculated as follows: (Guimarães and Sarsfield Cabral, 2010):

Arithmetic Mean:

$$\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad (2.5)$$

Geometric Mean:

$$GM = \sqrt[N]{x_1 \cdot x_2 \cdot \dots \cdot x_N} \quad (2.6)$$

Harmonic Mean:

$$HM = \frac{N}{\sum_{n=1}^N \frac{1}{x_n}} \quad (2.7)$$

Median: Consider that the dataset is organized in ascending or descending order.

- If N is an odd number, then the median value is:

$$Med = x_{\frac{N+1}{2}}^* \quad (2.8)$$

- Else, the median value is calculated as:

$$Med = \frac{x_{\frac{N}{2}}^* + x_{\frac{N}{2}+1}^*}{2} \quad (2.9)$$

Mode: refers to the value or values that appear most frequently in the dataset.

2.7.1.2 Measures Of Dispersion

There are three important measures of dispersion in statistics: sample amplitude, mean absolute deviation and mean square deviation. The three parameters can be calculated as follows (Guimarães and Sarsfield Cabral, 2010):

Sample Amplitude:

$$A = (x_n)_{Max} - (x_n)_{min} \quad (2.10)$$

Mean Absolute Deviation:

$$MAD = \frac{\sum_{n=1}^N |x_n - \bar{x}|}{N} \quad (2.11)$$

Mean Square Deviation:

$$MSD = \frac{\sum_{n=1}^N (x_n - \bar{x})^2}{N} \quad (2.12)$$

2.7.2 Statistical Measures For Bivariate Datasets

Consider a dataset that consists of N data (x_n, y_n) , with $n = (1, 2, \dots, N)$.

Error associated with a point (x_n, y_n) - Least Squares Method:

$$e_n = y_n - (a' + b \cdot y_n) \quad (2.13)$$

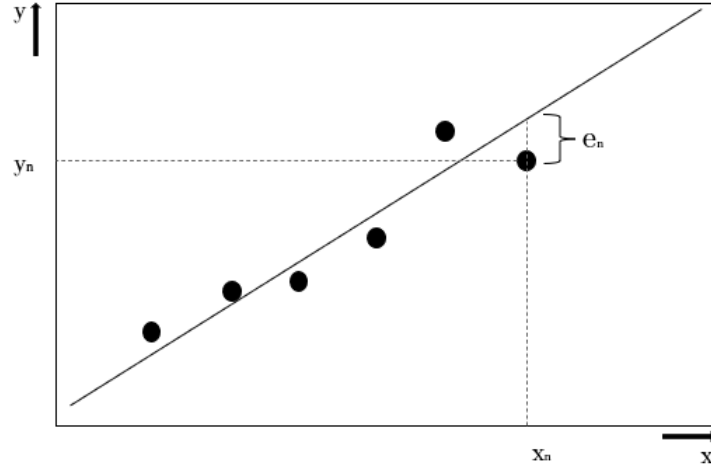


Figure 2.16: Error definition in the least square method (Adapted from Guimarães and Sarsfield Cabral (2010))

Coefficient of Determination:

$$r_{xy}^2 = 1 - \frac{\sum_{n=1}^N e_n^2}{s_{yy}} \quad (2.14)$$

Corrected Coefficient of Determination:

$$r_{xy(corrected)}^2 = 1 - \frac{\sum_{n=1}^N e_n^2 / (N - 2)}{s_{yy} / (N - 1)} \quad (2.15)$$

in which,

$$s_{yy} = \sqrt{\frac{1}{N - 1} \sum_{n=1}^N (y_n - \bar{y})^2} \quad (2.16)$$

is the variance associated with the yy values.

2.7.3 Normal Distribution

The normal distribution is one of the most famous continuous univariate distributions utilized in statistics due to being able to describe several phenomena related to continuous univariate variables (Guimarães and Sarsfield Cabral, 2010).

Consider that X is a random variable that results from a sum of several effects provoked by independent causes and that the effect of each cause can be neglected when related to the

sum of all other effects. Then, X can be said to follow a normal distribution (Guimarães and Sarsfield Cabral, 2010).

The normal distribution, represented in Figure 2.17, is characterized by the following parameters:

μ - expected value;

σ^2 - variance.

Its probability density function is given by the following equation:

$$f(x) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot \exp \left[-0.5 \left(\frac{X - \mu}{\sigma} \right)^2 \right] \quad (2.17)$$

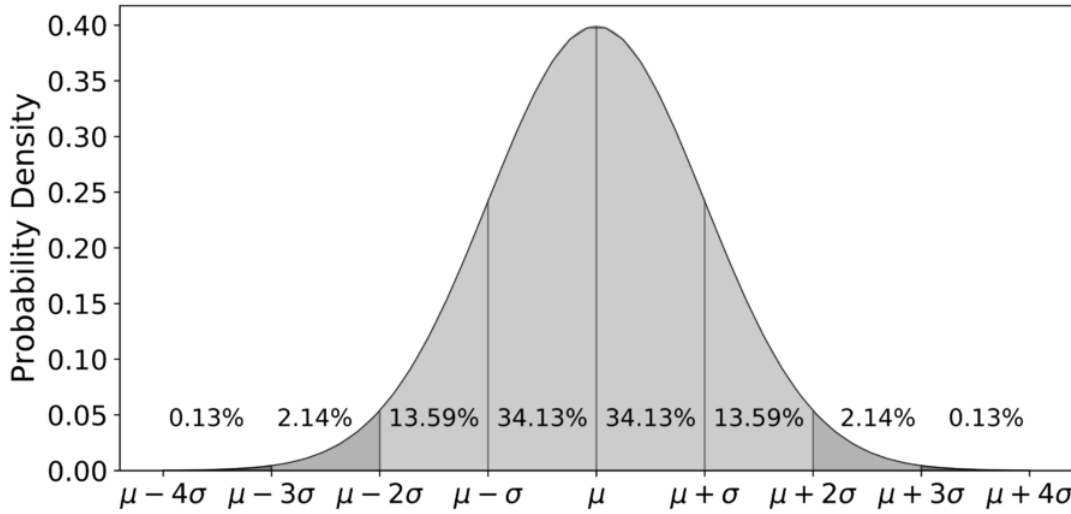


Figure 2.17: Normal Distribution (Adapted from Guimarães and Sarsfield Cabral (2010))

Consider the variable $V = a + b \cdot X$, obtained by the transformation of variable X . It can be proved that:

$$\mu_V = a + b \cdot \mu \quad (2.18)$$

and,

$$\sigma_V^2 = b^2 + \sigma^2 \quad (2.19)$$

The same can be proved that the variable Y also follows a normal distribution.

May Z be considered as a new variable that is calculated as follows:

$$Z = \frac{X - \mu}{\sigma} \quad (2.20)$$

Due to the fact that the transformation of the variable X is linear, the variable Z is also normal. Its parameters are calculated according to the following formulas:

$$\mu_Z = E\left(\frac{X - \mu}{\sigma}\right) = \frac{E(X) - \mu}{\sigma} = 0 \quad (2.21)$$

and,

$$\sigma_Z^2 = Var\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma^2} \cdot Var(X) = 1 \quad (2.22)$$

The distribution $N(0, 1)$ is designated as a standard normal distribution, and the variable Z as a standard normal variable. The probability density function is given by:

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot \exp -\frac{z^2}{2} \quad (2.23)$$

Let $z(\alpha/2)$ a value such that $P[Z > z(\alpha/2)] = \alpha/2$ and $-z(\alpha/2)$ the symmetrical value such that $P[Z < -z(\alpha/2)] = \alpha/2$, as seen in Figure 2.18.

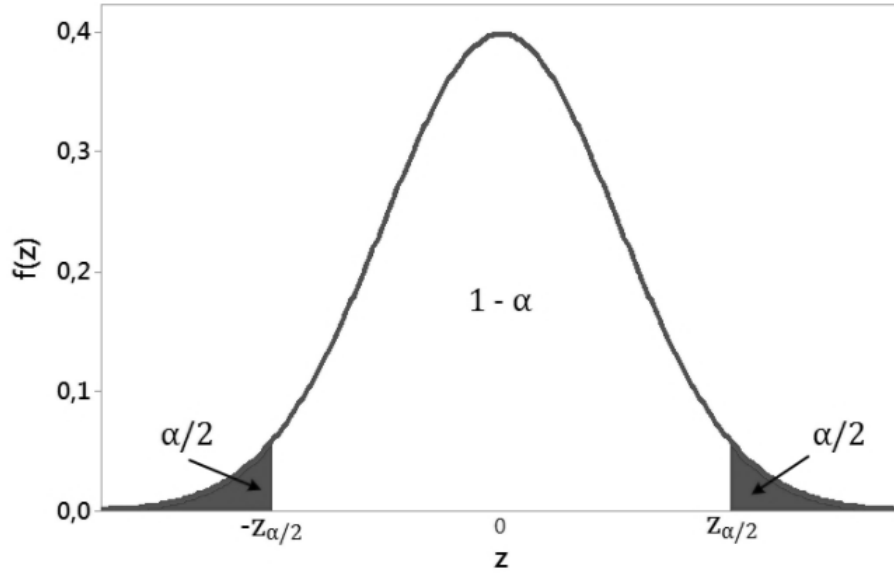


Figure 2.18: Symmetrical values of distribution $N(0,1)$ (Adapted from Guimarães and Sarsfield Cabral (2010))

This means that:

$$P[-z(\alpha/2) < Z < z(\alpha/2)] = 1 - \alpha \quad (2.24)$$

which can be re-written as,

$$P[-z(\alpha/2) < \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} < z(\alpha/2)] = 1 - \alpha \quad (2.25)$$

or even,

$$P\left[\bar{X} - z(\alpha/2) \cdot \frac{\sigma}{\sqrt{N}} < \mu < \bar{X} + z(\alpha/2) \cdot \frac{\sigma}{\sqrt{N}}\right] = 1 - \alpha \quad (2.26)$$

Such is called a confidence interval for μ with $(1 - \alpha) \cdot 100\%$.

2.7.4 Data Pre-Processing And Preparation

Zhang et al. (2003) define data preparation and pre-processing as the fundamental stages of data analysis. Data preparation techniques are those dealing with evaluating raw data in order to produce quality data, and they primarily include data collection, data integration, data transformation, data cleaning, data reduction, and data discretization.

2.7.4.1 Dealing With Missing Values

According to Mack et al. (2018), missing values can be divided into three main categories:

- **Missing Completely at Random - MCAR:** The premise of MCAR is that missing data is independent of seen and unobserved data. This means there are no systematic differences between instances with missing data and those with complete data, restricting the analyzable population and statistical power without introducing bias. The remaining data may be considered a simple random sample of the whole data set of interest (Mack et al., 2018).
- **Missing at Random - MAR:** For MAR, the likelihood of a missing value is proportional to the variable's value or other variables in the dataset. This indicates that not all data and variables are equally likely to be absent (datacamp, 2023).
- **Missing Not at Random - MNAR:** It is considered the most challenging of the three missing data scenarios. When neither MAR nor MCAR is relevant, it is employed. The risk of missing in this situation varies substantially depending on the value of the same variable, and the reasons for this are unknown (datacamp, 2023).

When it comes to dealing with these missing values, there are two ways to overcome this issue.

2.7.4.1.1 Deletion

- **Listwise:** This technique deletes all data for observation with one or more missing values, enabling the analysis to run solely on observations with all of their data. This, however, might result in skewed parameters and estimates, reducing the statistical power of the investigation (edX, 2023).
- **Pairwise:** The study assumes that data are missing totally at random, but all cases having data are utilized. This enables data scientists to use more data, but the outcomes may differ owing to differences in data sets. The results may be hard to reproduce with a whole set of data (edX, 2023).
- **Dropping Variables:** If data is missing for more than 60% of the observations, it may be prudent to delete it if the variable is negligible (edX, 2023).

2.7.4.1.2 Imputation

- **Mean, Median and Mode :** When coping with missing data, this is one of the most often used strategies for imputing values. When there are only a few missing observations, data scientists can compute the mean or median of the remaining observations. When there are numerous missing variables, however, mean or median values might lead to a loss of variety in the data (edX, 2023).
- **Last Observation Carried Forward (LOCF) And Next Observation Carried Backward (NOCB):** This technique is used to examine longitudinal repeated measures data that may be lacking follow-up observations, replacing each missing value with the most recently observed value. It is simple to grasp and apply but can lead to bias when data shows a clear trend (edX, 2023).
- **Linear Interpolation:** Linear interpolation is a formula for estimating the value of a function based on two known values at separate times. The weights are proportionate to the distance between the endpoints and the unknown point, with the closest point wielding more influence than the farthest point. This method is appropriate for time series with a trend line but not for seasonal data. If the data shows seasonality, seasonal adjustment with linear interpolation should be applied (edX, 2023).

2.7.4.2 Outliers Detection In Linear Regression Analysis

An outlier in a data collection is defined as an observation or group of observations that appear inconsistent with that set's rest of the data. Outliers may reflect tainted data or represent an accurate observation of a rare instance (Zakaria et al., 2014).

Begashaw and Yohannes (2020) refer that there are four main causes for the existence of these outliers:

- Incorrect recording data.
- Intentional or motivated misreporting.
- Sampling error.
- Legitimate cases sampled from the correct population.

Outliers can also be divided into three types depending on their composition and relationship to the rest of the data, as it is depicted in Figure 2.19 (Begashaw and Yohannes, 2020):

- **Point outliers:** Data points that are out of place with the rest of the data because their attribute values differ from the norm. The majority of outlier identification research is centred on this.
- **Context Outliers:** As opposed to point outliers, they are created by the presence of an individual data item in a specific environment. They are defined in relation to a context, which means they may not be an outlier in another context.
- **Collective outliers:** Collective outliers are groups of data points that are out of place in relation to the rest of the data set. They can only arise in data sets where data points are connected in some way since a subset of data instances is outlying in relation to the complete data set.

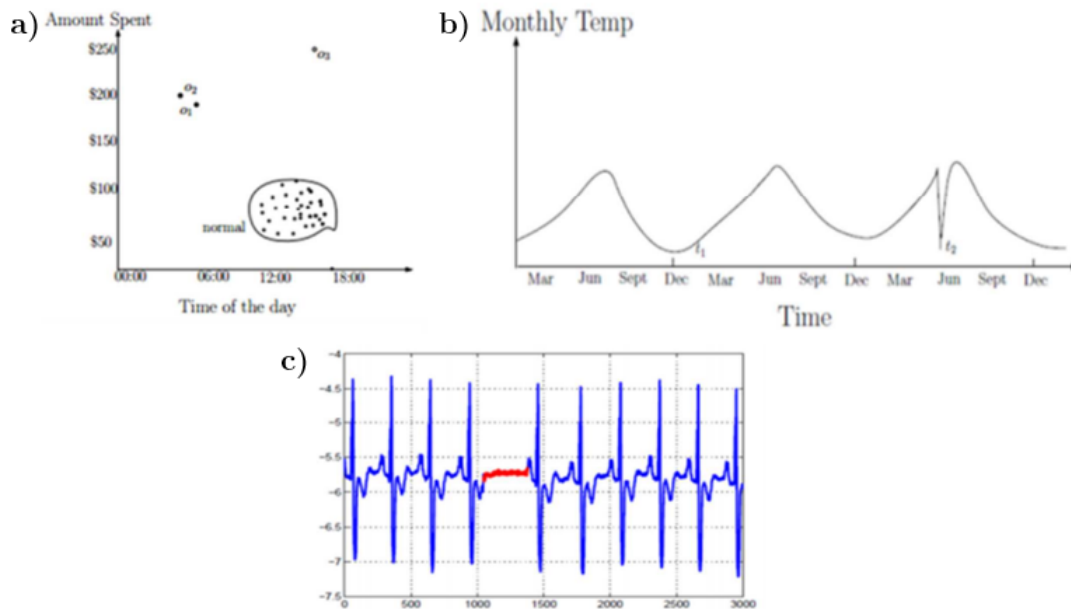


Figure 2.19: Types of outliers: (a) Point outliers, (b) Context outliers, (c) Collective outliers (Begashaw and Yohannes, 2020)

Several techniques can be employed to detect outliers in a dataset when using linear regression.

Leverage Values

Outliers with regard to their x values are identified using leverage values. This number represents the distance between the observation's x values and the data's centre.

The hat matrix H is a key player in the identification of possible outliers due to the fact that from it, the variance and covariance of the fitted value $\hat{y}$ and random error e can be determined (Borse, 2020):

$$H = X(X'X)^{-1}X' \quad (2.27)$$

$$Var(\hat{y}) = \sigma^2 H \quad (2.28)$$

$$Var(e) = \sigma^2(I - H) \quad (2.29)$$

The hat matrix's diagonal components (leverage values) are an effective indication (Zakaria et al., 2014). The leverage values are calculated as follows:

$$h_{ii} = x_i'(XX)^{-1}x_i \quad (2.30)$$

where,

$i = 1, 2, \dots, N$;

X - data matrix.

h_{ii} - leverage value (diagonal value of the hat matrix).

If a leverage value h_{ii} is more than twice as great as the mean leverage value $\hat{h}$, it is deemed high, and that instance may be considered as an outlier (Borse, 2020).

Standardized Residuals

This technique obliges the mean of the residuals to be zero and have approximately a unit variance (Borse, 2020). The following formula characterizes it:

$$d_i = \frac{\varepsilon_i}{\sqrt{MS_{Res}}} \quad (2.31)$$

where,

$i = 1, 2, \dots, N$;

$\varepsilon_i = y_i - \hat{y}_i$;

y_i - observed value;

$\hat{y}_i$ - fitted value.

An outlier is identified if $d_i > 3$.

Studentised Residuals

The studentized residual is given by the following equation (Borse, 2020):

$$t_i = \frac{\varepsilon_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}} \quad (2.32)$$

where,

$i = 1, 2, \dots, N$;

t_i - studentized residual;

ε_i - residuals;

$MSE_{(i)}$ - mean square error;

h_{ii} - hat matrix diagonal value.

According to Borse (2020), the studentized residuals are more effective for outliers detection than the standardized residuals. It can be considered an outlier if $t_i > 3$.

Cook's Distance

This parameter measures the influence that the removal of the i^{th} observation has from the dataset or sample (Borse, 2020). It can be calculated with the following equation:

$$D_i = \frac{r_i^2 h_{ij}}{p(1 - h_{ij})} \quad (2.33)$$

where,

$i = 1, 2, \dots, N$;

h_{ij} - diagonal element of the H matrix;

r_i - i^{th} column of matrix $R = (X'X)^{-1}X'$;

p - number of regression parameters.

If $D_i > 1$, then the i^{th} observation can be considered an outlier (Borse, 2020).

DFBETAS

DFBETAS stands for "difference in the estimates of the regression coefficients after excluding a particular observation". The method aims to indicate the change in the regression coefficient β_j , in standard deviation units (Borse, 2020). It is calculated as follows:

$$DFBETAS_{j,i} = \frac{r_{j,i}}{\sqrt{r'_j r_j}} \frac{t_i}{\sqrt{1 - h_{ii}}} \quad (2.34)$$

where,

$i = 1, 2, \dots, N$;

r'_j - j^{th} row of the matrix $R = (X'X)^{-1}X'$;

t_i - R-student residual;

h_{ii} - diagonal element of the H matrix.

The i_{th} observation is considered an outlier if $|DFBETAS_{j,i}| > \frac{2}{\sqrt{n}}$.

DFFITIS

DFFITIS stands for "difference in fit(s)". It investigates the influence of deleting the i_{th} on the predicted or fitted value:

$$DFFITIS_i = \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{0.5} t_i \quad (2.35)$$

where,

$i = 1, 2, \dots, N$.

h_{ii} - diagonal element of the H matrix.

t_i - R-student residual.

If $|DFFITIS_i| > 2\sqrt{\frac{p}{n}}$, then the i_{th} observation can be considered as an outlier.

2.7.4.3 Outliers detection using robust regression models

The Ordinary Least Squares (OLS) approach is the most popular regression method. However, certain requirements must be met, and outlier observations can breach these assumptions, which can significantly influence the regression findings. Researchers are aware that outlying observations in the data might violate model assumptions and significantly affect the outcomes of regression analysis (Begashaw and Yohannes, 2020).

Robust regression employs a weighting method to limit the influence of outliers on regression coefficient estimates, yielding coefficient values that differ from OLS (Begashaw and Yohannes, 2020). The models that rely on robust regression will be presented and further discussed.

Least Absolute Residuals

A reliable regression technique called the method of least absolute residuals minimizes the total absolute deviations of the Y data from their averages in order to estimate the regression coefficients (Begashaw and Yohannes, 2020). It is given by the following equation:

$$\min \sum_{i=1}^n |y_i - (\beta_0 + \beta_1 x_{i1} + \dots \beta_{p-1} x_{ip-1})| \quad (2.36)$$

where,

β - Regression coefficients.

y_i - i^{th} Observation.

Least Median of Squares

Reduces the median squared residuals, which allows up to half of the data to disagree with the model. It is a robust estimator of location because it replaces the sum of squared deviations in ordinary least squares with the median of squared deviations (Begashaw and Yohannes, 2020):

$$\min_{median} [y_i - (\beta_0 + \beta_1 x_{i1} + \dots \beta_{p-1} x_{ip-1})]^2 \quad (2.37)$$

where,

β - Regression coefficients.

y_i - i^{th} Observation.

M-Estimation

In the domain of location models, the M-estimation approach is an extension of maximum likelihood estimation. This is almost as effective as OLS. The M-estimation approach aims to minimize the residual function rather than the sum of squared errors (Almetwally and Almongy, 2018). The objective function of the M-estimate is:

$$\hat{\beta}_M = \min \sum_{i=1}^n \rho(y_i - \sum_{j=0}^k x'_{ij} \beta_j) \quad (2.38)$$

where,

$\hat{\beta}_M$ - Objective function.

ρ - Residual function.

x^{ij} - i^{th} Observation on the j^{th} independent variable.

β_j - Coefficient of the j^{th} variable.

S-Estimation

Based on the M-estimation method's residual scale, the S-estimator technique is a regression estimator related to M-scales. LMS and LTS are generalized to form S-estimators. Additionally, they handle 50% of the data's outliers and have the same asymptotic qualities as M-estimators (Almetwally and Almongy, 2018). The S-Estimator is given by:

$$\hat{\beta}_S = \min_{\beta} [\hat{\sigma}_S(\epsilon_1, \epsilon_2, \dots, \epsilon_n)] \quad (2.39)$$

with the determining minimum robust scale estimator $\hat{\sigma}_S$ and satisfying:

$$\min \sum_{i=1}^n \rho \left(\frac{y_i - \sum_{j=1}^n x_{ij} \beta_j}{\hat{\sigma}_S} \right) \quad (2.40)$$

where,

$$\hat{\sigma}_S = \begin{cases} \frac{\text{median}|\epsilon_i - \text{median}(\epsilon_i)|}{0.6745}, \text{iteration} = 1 \\ \sqrt{\frac{\sum_{i=1}^n w_i \epsilon_i^2}{nk}}, \text{iteration} > 1 \end{cases} \quad (2.41)$$

MM-Estimation

A specific kind of M-estimation approach is the MM-estimation method. The MM-estimator, which was the first estimation with a high breakdown point and high efficiency under normal error, is a combination of the high breakdown value estimation technique and the efficient estimation method (Almetwally and Almongy, 2018). The MM-Estimator is calculated as follows:

$$\sum_{i=1}^n \rho_1' \left(\frac{y_i - \sum_{j=0}^k x_{ij}' \beta_j}{s_{MM}} \right) x_{ij} = 0 \quad (2.42)$$

where,

s_{MM} - standard deviation obtained from the residual of the S-estimation method.

2.7.5 Stationarity, Trend And Seasonality Definitions In Time Series

Several aspects must be considered when defining the main technique for the model. Three of them are trend, seasonality and stationarity of the time series.

Hyndman and Athanasopoulos (2018) state that a stationary time series is one whose attributes are unaffected by the time at which it is examined, while a trend is when there is a long-term increase or decrease in the data, which does not have to be linear.

On the other hand, the same authors construe seasonality as when a time series is influenced by seasonal elements such as the season or day of the week. It has a definite and predictable periodicity.

Each parameter mentioned above will be further analysed and evaluated in the following subsections.

2.7.6 Stationarity Analysis

The main stationarity tests are going to be presented and reviewed in this section.

2.7.6.1 Augmented Dickey-Fuller Test

Dickey and Fuller (1979) created in 1979 one of the most notorious statistical tests for the stationarity analysis of a time series: the Dickey-Fuller test. In 1984 the same authors proposed an updated version of the same test that accommodated more complex models with unknown orders. It was named the Augmented Dickey-Fuller (ADF) test. ADF is used to detect if an autoregressive model contains a unit root.

A unit root refers to a stochastic trend in time series, having a random probability distribution or pattern that may be analysed statistically but not predicted precisely, or an event or prediction that derives from a random process or random probability distribution (MasterClass, 2022).

The ADF test follows the same method as the Dickey-Fuller test. Dickey and Fuller consider three differential-form autoregressive equations to determine the presence of a unit root given an observed time series $Y_1, Y_2, \dots, Y_N$ (RTC Lab, 2020), presented below:

$$\Delta Y_t = \gamma Y_{t-1} + \sum_{j=1}^n (\delta_j \Delta Y_{t-j}) + \varepsilon_t \quad (2.43)$$

$$\Delta Y_t = \alpha + \gamma Y_{t-1} + \sum_{j=1}^n (\delta_j \Delta Y_{t-j}) + \varepsilon_t \quad (2.44)$$

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{j=1}^n (\delta_j \Delta Y_{t-j}) + \varepsilon_t \quad (2.45)$$

where,

t - Time Index.

α - Intercept constant called a drift.

β - Coefficient on a time trend.

γ - Coefficient presenting process root.

p - Lag order of the autoregressive process.

ε_t - Independent identically distributing residual term. After this first step, the unit root test is carried out with the following hypotheses:

$H_0 - \gamma = 0$ (Time series is not stationary).

$H_A - \gamma < 0$ (Time series is stationary).

To conduct the unit root test, which is nothing more than a t-test for H_0 , the following formula is applied:

$$ADF = \frac{\hat{\gamma}}{SE(\hat{\gamma})} \quad (2.46)$$

where SE stands for Standard Error.

According to each formula, the following conclusions, presented in Table 5.8, can be drawn from the ADF test.

Table 2.7: Conclusions from the ADF Test according with each equation (RTC Lab, 2020)

ADF Equation	Test Result
3.1	H_0 : Y_t is a random walk or $\gamma = 0$. H_A : Y_t is stationary process or $\gamma < 0$.
3.2	H_0 : Y_t is random walk around a drift or $\gamma = 0, \alpha \neq 0$. H_A : Y_t is level stationary process or $\gamma < 0, \alpha \neq 0$.
3.3	H_0 : Y_t is random walk around a drift or $\gamma = 0, \beta \neq 0$. H_A : Y_t is trend stationary process or $\gamma < 0, \beta \neq 0$.

Note that a Random Walk is, according to Britannica (2023), a "process for determining the probable location of a point subject to random motions, given the probabilities (the same at each step) of moving some distance in some direction".

If the computed test statistic is less than the critical value, the null hypothesis is rejected, and there is no unit root. The ADF test guarantees that the null hypothesis is accepted unless there is a compelling reason to reject it in favour of the alternative stationarity hypothesis (RTC Lab, 2020).

2.7.6.2 Kwiatkowski–Phillips–Schmidt–Shin Test

Kwiatkowski et al. (1992) recommended that trend stationarity be treated as the null hypothesis, with the unit root as the alternative. The rejection of the null hypothesis of trend stationarity might subsequently be seen as compelling evidence in favour of a unit root (Kokoszka and Young, 2016).

According to Nielsen (2005), assuming that there is no trend and the point of departure is a data-generating process of the form:

$$Y_t = \xi_t + \epsilon_t \quad (2.47)$$

where ξ_t is a random walk and ϵ_t is stationary error.

For the case if $\sigma_y^2 = 0$, this is if the variance is zero, then $\xi_t = \xi_0$ for all t and Y_t is stationary.

By using a simple regression:

$$Y_t = \hat{\mu} + \hat{\epsilon}_t \quad (2.48)$$

we are able to find the estimated stochastic component. Under the null, $\hat{\epsilon}_t$ is stationary. Such observation can be used to design the following testing hypotheses:

$H_0 - \sigma_y^2 = 0$ (Time series is stationary).

$H_A - \sigma_y^2 < 0$ (Time series is not stationary).

The test statistic is given by:

$$KPSS = \frac{1}{T^2} \cdot \frac{\sum_{t=1}^T S_t^2}{\hat{\sigma}_\infty^2} \quad (2.49)$$

where:

S_t - sum of the residuals;

$\hat{\sigma}_\infty^2$ - Heteroskedasticity and autocorrelation consistent estimator of the variance of $\hat{\epsilon}_t$.

For further clarification, heteroskedasticity, according to Hayes (2022), in statistics, occurs when a variable's standard errors are not constant throughout a predetermined period of time. It is a violation of the assumptions for linear regression modelling.

2.7.7 Seasonality Analysis

When it comes to a seasonality analysis, there are three ways of achieving some conclusions:

- Plot a cycle plot.
- Decompose the time series.
- Plot the lagged autocorrelation.

2.7.7.1 Cycle Plot

Cycle plots were introduced by Cleveland et al. (1978) to analyze the behaviour of seasonal time series. They first split the time series into frequency components: trend, seasonality, oscillation, and the rest, and then investigated each component separately using various methodologies. The seasonal component was examined using cycle plots (Robbins, 2008). Figure 2.20 shows an example of a cycle plot.

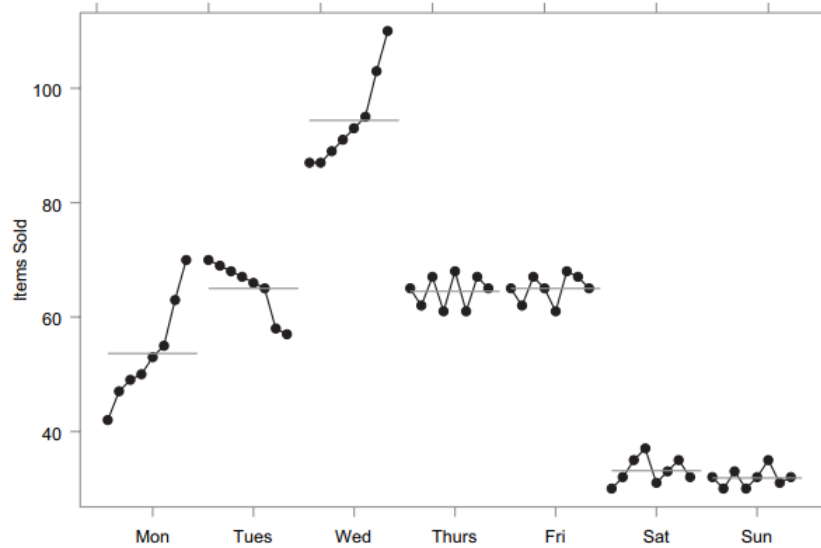


Figure 2.20: Cycle plot of item sales by weekday (Robbins, 2008)

2.7.7.2 Decomposition of the time series

A crucial method in retrospective time series analysis, deseasonalization, and other processes is the breakdown of a time series model into its component latent subseries. There are two types of decompositions: additive and multiplicative (West, 1997).

Consider Y_t as the time series, S_t as the seasonal component, T_t as the trend component and R_t as the random component.

The following equation gives an additive time series:

$$Y_t = S_t + T_t + R_t \quad (2.50)$$

If seasonal fluctuations do not change with the level of the time series, additive decomposition is the best method to use (Hyndman and Athanasopoulos, 2018).

On the other hand, the multiplicative is given by the following formula:

$$Y_t = S_t \cdot T_t \cdot R_t \quad (2.51)$$

When the variation in the seasonal pattern or the oscillation around the trend cycle seems to be proportionate to the time series level, a multiplicative decomposition is more appropriate (Hyndman and Athanasopoulos, 2018).

Howell (2022b) established an example by decomposing an airline passenger dataset. The dataset can be found in Figure 2.21, while the additive and multiplicative decompositions in Figures 2.22 and 2.23, respectively.

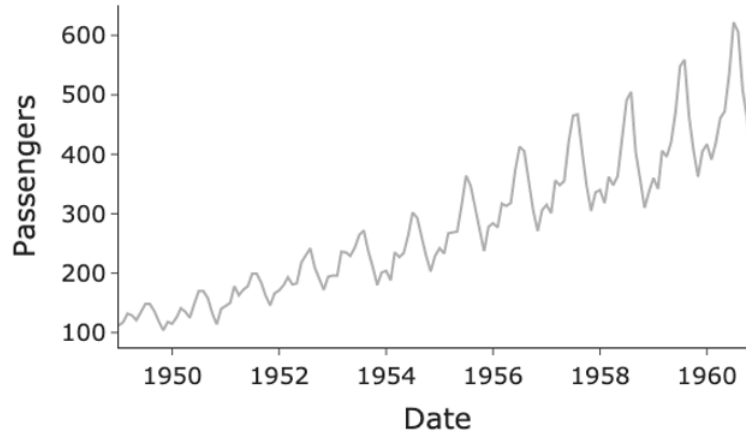


Figure 2.21: Airline dataset plotted (Howell, 2022b)

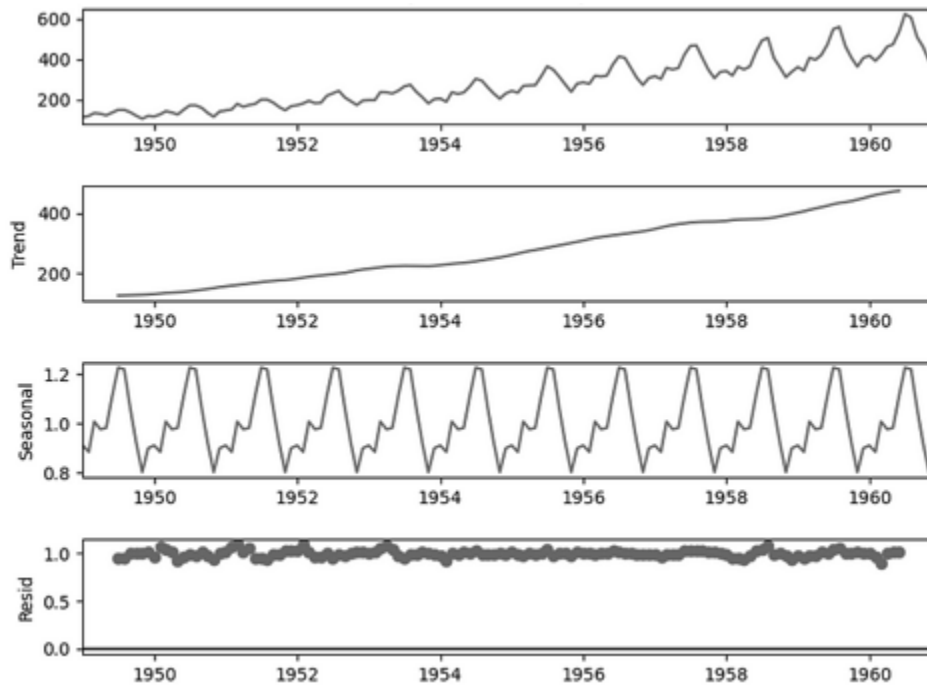


Figure 2.22: Multiplicative Decomposition (Howell, 2022b)

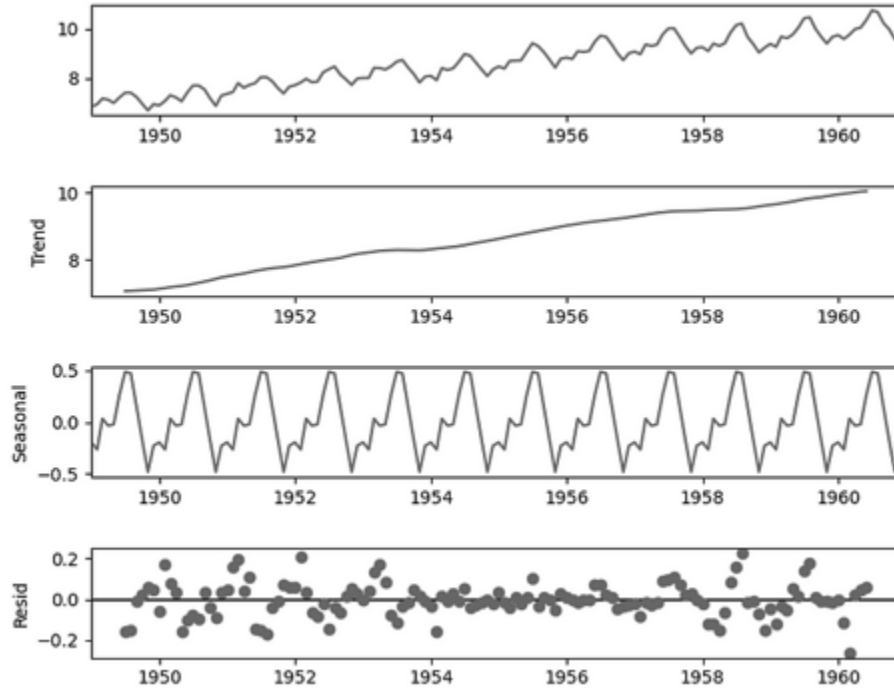


Figure 2.23: Additive Decomposition (Howell, 2022b)

The time series was also converted to an additive model by stabilising the variance using the Box-Cox Transform: a statistical technique to transform non-normal data into a normal distribution (Howell, 2022b). Thus, such indicates that the same conclusions can be achieved by using both models, even though, in the additive decomposition, the residuals have higher volatility in the earlier and later years.

One must note that, in this case, it was easy to perceive that this time series was multiplicative due to the size of the fluctuations increasing through time (Howell, 2022b).

However, it is usually very hard to understand if a time series is additive or multiplicative, as it could have both nuances in its data. Hyndman and Athanasopoulos (2018) also advise against the use of these classical decomposition methods due to the following drawbacks:

- The estimated trend cycle is not given for the initial few and last few observations.
- Rapid climbs and drops in the data tend to be overly smoothed by the trend-cycle estimate.
- Traditional decomposition techniques make the assumption that the seasonal component repeats itself year after year, which may not be true for longer series.
- When the time series values are highly unusual across short periods, the classical techniques are unreliable.

Hence, to overcome this issue, other techniques can be applied. One of them is the STL Decomposition. STL is a versatile and effective method for decomposing time series. STL stands for "Seasonal and Trend decomposition using Loess". Loess is a method for calculating nonlinear relationships (Hyndman and Athanasopoulos, 2018). This method has the following advantages:

- STL will handle any type of seasonality.
- Since it can withstand outliers, estimations of the trend-cycle and seasonal components won't be impacted by some unusual observations.
- The user has control over the seasonal component and may alter it over time, as well as the smoothness of the trend cycle.

2.7.7.3 Lagged Autocorrelation - ACF Test

According to Hyndman and Athanasopoulos (2018), the linear relation between lagged values of a time series is measured by autocorrelation. The autocorrelation coefficient for each lag, r_k , is calculated according to the following equation:

$$r_k = \frac{\sum_{t=k+1}^T (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2} \quad (2.52)$$

where T is the length of the time series.

2.7.7.4 Partial Lagged Autocorrelation - PACF

The partial correlation of a time series at two separate points in time is thus defined as partial autocorrelation. It is the correlation between two-time series with distinct lags, without considering the influence of any intermediate lags (Howell, 2022a).

Consider a general autoregression formula of the following type (Sachin Date, 2019):

$$T_i = \beta_0 + \beta_1 T_{i-1} + \beta_2 T_{i-2} + \dots + \beta_n T_1 \quad (2.53)$$

where,

T_i - parameter to be forecasted.

The formula for PACF at lag = k is given by:

$$PACF(T_i, k) = \frac{\text{Covariance}([T_i | T_{i-1}, T_{i-2}, \dots, T_{i-k+1}], [T_{i-k} | T_{i-1}, T_{i-2}, \dots, T_{i-k+1}])}{\sigma_{[T_i | T_{i-1}, T_{i-2}, \dots, T_{i-k+1}]} \cdot \sigma_{[T_{i-k} | T_{i-1}, T_{i-2}, \dots, T_{i-k+1}]}} \quad (2.54)$$

where,

$T_i|T_{i-1}, T_{i-2}, \dots, T_{i-k+1}$ - time series of residuals obtained from fitting a multivariate linear model to $T_{i-1}, T_{i-2} \dots T_{i-k+1}$ for predicting T_i . It represents the residual variance in T_i after stripping away the influence of $T_{i-1}, T_{i-2} \dots T_{i-k+1}$.

2.7.8 Univariate Time Series

Based on prior measurements and related error factors, UTS algorithms predict can predict, for example, future water consumption (Anele et al., 2017). The two most extensively used techniques for univariate time series forecasting are exponential smoothing and ARIMA models, which give complementary approaches to the problem. While exponential smoothing methods attempt to explain the data's trend and seasonality, ARIMA models strive to characterize the data's autocorrelations (Hyndman and Athanasopoulos, 2018). In the next subsections, the UTS models will be extensively presented.

2.7.8.1 Autoregressive Model

Researchers predict the variable of interest in an autoregression (AR) model by implementing a linear combination of the variable's historical values. The word autoregression refers to a regression of a variable against itself. An AR model of order p , or simply an AR(p) model, can be described by the following equation (Hyndman and Athanasopoulos, 2018):

$$y_t = \mu + \sum_{k=1}^p \phi_k y_{t-k} + \epsilon_t \quad (2.55)$$

where:

μ - Average value of the process.

ϕ - Autoregressive parameter.

ϵ_t - Forecast error at the time.

y_t - Observed value at the time t .

k - Number of historical periods.

2.7.8.2 Moving Average Model

A moving average (MA) model relies on past forecast errors in a regression rather than utilizing prior predicted values in a regression-like model. A moving average model of order q , or a MA(q) model, is given by the following equation (Hyndman and Athanasopoulos, 2018):

$$y_t = \mu + \sum_{k=1}^q \theta_k \varepsilon_{t-k} + \epsilon_t \quad (2.56)$$

where:

μ - average value of the process.

θ - moving average parameter.

ϵ_t - forecast error at the time.

y_t - observed value at the time t .

k - number of historical periods.

ε_{t-k} - observation at time $t - k$.

2.7.8.3 Autoregressive Integrated Moving Average Model

A non-seasonal autoregressive integrated moving average model (ARIMA) model may be obtained by combining differencing with autoregression and a moving average model. The full model can be written as (Hyndman and Athanasopoulos, 2018):

$$y'_t = \mu + \sum_{k=1}^p \phi_k y'_{t-k} + \sum_{k=1}^q \theta_k \varepsilon_{t-k} + \epsilon_t \quad (2.57)$$

where:

y'_t - Differenced series value.

μ - Average value of the process.

θ - Moving average parameter.

ϵ_t - Forecast error at the time.

k - Number of historical periods.

ε_{t-k} and y'_{t-k} - Observation at time $t - k$.

This is called an ARIMA(p, d, q) model, in which:

p - Order of the autoregressive part.

d - Degree of difference involved.

q - Order of the moving average part.

2.7.8.4 Seasonal Autoregressive Integrated Moving Average Model

The seasonal autoregressive integrated moving average model (SARIMA) is an extension of the ARIMA model with the implementation of seasonality. The general formula can be written as follows (Chang et al., 2013):

$$\phi_P(B^s)\varphi(B)\nabla_s^D\nabla^d x_t = \Theta_Q(B^s)\theta(B)w_t \quad (2.58)$$

where:

x_t - Time series.

$\phi_P(B^s)$ - Ordinary autoregressive seasonal component of order P .

$\Theta_Q(B^s)$ - Moving average seasonal component of order Q .

$\varphi(B)$ - Ordinary autoregressive component of order p .

$\theta(B)$ - Moving average component of order q .

∇^d - Ordinary difference component.

∇_s^D - Seasonal difference component.

B - Backshift operator.

w_t - Gaussian white noise.

Such is called a SARIMA $(p, d, q)(P, D, Q)$ model, in which:

p - Order of the autoregressive part.

d - Degree of first difference involved.

q - Order of the moving average part.

P - Order of the seasonal autoregressive part.

D - Degree of seasonal difference involved.

Q - Order of the seasonal moving average part.

2.7.8.5 Exponential Smoothing

General Exponential Smoothing Models

The first exponential smoothing methods trace back to the late 1950s when Holt and Winters developed a model that could estimate point predictions. The Holt-Winters seasonal technique consists of the forecast equation plus three smoothing equations: one for the level l_t , one for the trend b_t , and one for the seasonal component s_t . Each with its own smoothing parameters:

α - Smoothing factor for the level.

β^* - Smoothing factor for the trend.

γ - Smoothing factor for the seasonality.

Table 2.8 shows nine exponential smoothing approaches that may be used by considering trend and seasonal components fluctuations. Each approach is identified by a pair of letters (T, S) that indicate the kind of Trend and Seasonal components.

Table 2.8: Formulas for recursive calculations and point forecasts using exponential smoothing (Hyndman and Athanasopoulos, 2018)

Trend	Seasonal		
	N	A	M
N	$\hat{y}_{t+h t} = \ell_t$	$\hat{y}_{t+h t} = \ell_t + s_{t+h-m(k+1)}$	$\hat{y}_{t+h t} = \ell_t s_{t+h-m(k+1)}$
	$\ell_t = \alpha y_t + (1 - \alpha)\ell_{t-1}$	$\ell_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)\ell_{t-1}$	$\ell_t = \alpha(y_t/s_{t-m}) + (1 - \alpha)\ell_{t-1}$
		$s_t = \gamma(y_t - \ell_{t-1}) + (1 - \gamma)s_{t-m}$	$s_t = \gamma(y_t/\ell_{t-1}) + (1 - \gamma)s_{t-m}$
A	$\hat{y}_{t+h t} = \ell_t + hb_t$	$\hat{y}_{t+h t} = \ell_t + hb_t + s_{t+h-m(k+1)}$	$\hat{y}_{t+h t} = (\ell_t + hb_t)s_{t+h-m(k+1)}$
	$\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + b_{t-1})$	$\ell_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(\ell_{t-1} + b_{t-1})$	$\ell_t = \alpha(y_t/s_{t-m}) + (1 - \alpha)(\ell_{t-1} + b_{t-1})$
	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1}$
		$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m}$	$s_t = \gamma(y_t/(\ell_{t-1} + b_{t-1})) + (1 - \gamma)s_{t-m}$
Ad	$\hat{y}_{t+h t} = \ell_t + \phi_h b_t$	$\hat{y}_{t+h t} = \ell_t + \phi_h b_t + s_{t+h-m(k+1)}$	$\hat{y}_{t+h t} = (\ell_t + \phi_h b_t)s_{t+h-m(k+1)}$
	$\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1})$	$\ell_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1})$	$\ell_t = \alpha(y_t/s_{t-m}) + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1})$
	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)\phi b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)\phi b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)\phi b_{t-1}$
		$s_t = \gamma(y_t - \ell_{t-1} - \phi b_{t-1}) + (1 - \gamma)s_{t-m}$	$s_t = \gamma(y_t/(\ell_{t-1} + \phi b_{t-1})) + (1 - \gamma)s_{t-m}$

Note: N stands for None; A for Additive; M for Multiplicative and Ad for Additive Damped.

State Space Models For Exponential Smoothing Methods

In this case, each model comprises a measurement equation that explains the observed data and a set of state equations that describe how the unseen components or states - level, trend, seasonal - vary over time. As a result, these are known as state space models. These statistical models create point forecasts as well as prediction intervals. In order to distinguish between a model with additive errors and one with multiplicative errors and also to identify the models from the techniques, the models are labelled with ETS (Error, Trend, Seasonal) for each state space model (Hyndman and Athanasopoulos, 2018). Tables 2.9 and 2.10 present the equations that rule each model.

Table 2.9: Additive ETS Models (Hyndman and Athanasopoulos, 2018)

ADDITIVE ERROR MODELS			
Trend	N	Seasonal A	M
N	$y_t = \ell_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha \varepsilon_t$	$y_t = \ell_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha \varepsilon_t$ $s_t = s_{t-m} + \gamma \varepsilon_t$	$y_t = \ell_{t-1} s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha \varepsilon_t / s_{t-m}$ $s_t = s_{t-m} + \gamma \varepsilon_t / \ell_{t-1}$
A	$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha \varepsilon_t$ $b_t = b_{t-1} + \beta \varepsilon_t$	$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha \varepsilon_t$ $b_t = b_{t-1} + \beta \varepsilon_t$ $s_t = s_{t-m} + \gamma \varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1}) s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha \varepsilon_t / s_{t-m}$ $b_t = b_{t-1} + \beta \varepsilon_t / s_{t-m}$ $s_t = s_{t-m} + \gamma \varepsilon_t / (\ell_{t-1} + b_{t-1})$
Ad	$y_t = \ell_{t-1} + \phi b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha \varepsilon_t$ $b_t = \phi b_{t-1} + \beta \varepsilon_t$	$y_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha \varepsilon_t$ $b_t = \phi b_{t-1} + \beta \varepsilon_t$ $s_t = s_{t-m} + \gamma \varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1}) s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha \varepsilon_t / s_{t-m}$ $b_t = \phi b_{t-1} + \beta \varepsilon_t / s_{t-m}$ $s_t = s_{t-m} + \gamma \varepsilon_t / (\ell_{t-1} + \phi b_{t-1})$

Note: **N** stands for None; **A** for Additive; **M** for Multiplicative and **Ad** for Additive Damped.

Table 2.10: Multiplicative ETS Models (Hyndman and Athanasopoulos, 2018)

MULTIPLICATIVE ERROR MODELS			
Trend	N	Seasonal A	M
N	$y_t = \ell_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha \varepsilon_t)$	$y_t = (\ell_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1} s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha \varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma \varepsilon_t)$
A	$y_t = (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha \varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1}) s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha \varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma \varepsilon_t)$
Ad	$y_t = (\ell_{t-1} + \phi b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha \varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1}) s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha \varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma \varepsilon_t)$

Note: **N** stands for None; **A** for Additive; **M** for Multiplicative and **Ad** for Additive Damped.

2.7.9 Time Series Regression

According to Anele et al. (2017), unlike the UTS models, time series regression (TSR) forecasting models consider the effects of exogenous variables. This is because they generate forecasts based on the relationship between water demand and its determinants when it comes to predicting water consumption or demand patterns.

2.7.9.1 Multiple Linear Regression

Multiple Linear Regression (MLR) is used when there are two or more predictor variables. A multiple regression model takes the following general form:

$$y_t = \beta_0 + \sum_{i=1}^k \beta_i x_{i,t} + \varepsilon_t \quad (2.59)$$

where,

y - variable to forecast.

$x_1, \dots, x_k$ - k predictor variables.

$\beta_1, \dots, \beta_k$ - measure the marginal effects of the predictor variables.

2.7.9.2 Multiple Nonlinear Regression

Multiple Nonlinear Regression (MNLr) models establish the assumption that the mathematical relationship between the dependent variable, y_i , and the p vector of regressors x_i is nonlinear (Adamowski et al., 2012). The model is given by the following equation:

$$y_t = a + \beta_1 x_i + \beta_2 x_j + \beta_3 x_i^2 + \beta_4 x_j^2 + \dots + \beta_k x_i x_j \quad (2.60)$$

where,

a - intercept.

β - slope of the coefficient.

k - number of observations.

2.7.9.3 ARIMA with exogenous variables

The ARIMA with a single exogenous variable (ARIMAX) model is a type of ARIMA model that considers external factors or exogenous variables. For example, an ARIMAX model with a single exogenous variable can be described by the equation that follows (Donkor et al., 2014):

$$y_t = \mu + \sum_{k=1}^p \phi_k y_{t-k} + \epsilon_t + \sum_{k=1}^q \theta_{t-k} \epsilon_{t-k} + \sum_{k=0}^b \beta_k x_{t-k} \quad (2.61)$$

where b is the single exogenous variable considered in this ARIMAX model.

2.7.9.4 SARIMA with exogenous variables

SARIMAX, or SARIMA with exogenous variables, has the same model as ARIMAX. However, it considers as well the seasonality of the data. It is described as follows (Arunraj et al., 2016):

$$y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \dots + \beta_k X_{k,t} + w_t \quad (2.62)$$

where:

$$w_t = \frac{\theta_q(B)\Theta_Q(B^S)}{\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D} \cdot \varepsilon_t.$$

2.7.10 Artificial Neural Networks

Artificial Neural Networks (ANNs) are machine learning algorithms designed to mimic the structure and function of the human brain. ANNs are composed of linked processing components known as neurons that communicate with one another to accomplish a learning task based on a collection of observations. The neurons, whose layout is presented in Figure 2.24, process their inputs using a predefined activation function and produce output as a result. These are also known as Linear Threshold Units (LTUs). ANNs may be trained to detect data patterns, forecast outcomes, and perform classification tasks. The greater the number of hidden neurons in an ANN, the more complex functions it can learn (Alwosheel et al., 2018).

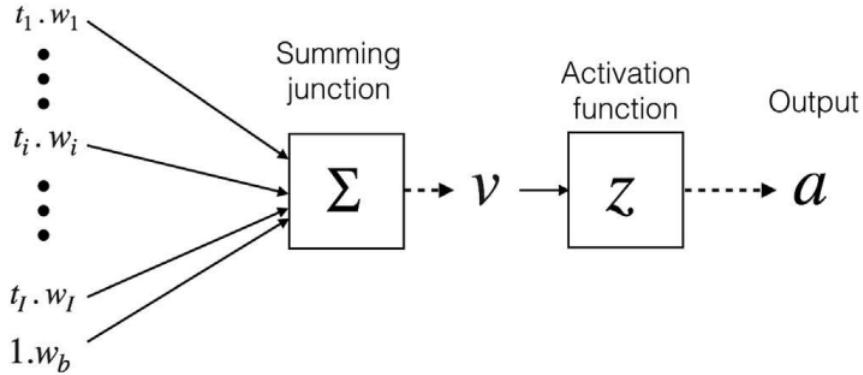


Figure 2.24: Neuron layout (Alwosheel et al., 2018)

An ANN neuron architecture comprises layers of interconnected processing components known as neurons. The input layer takes input signals t_i and forwards them to the following layer, known as the hidden layer. The hidden layer processes the inputs using a predefined activation function z , generating the output appropriately. The final output of the ANN is provided by the output layer. There may be one or more hidden layers between the input and

output layers, each consisting of several neurons that process their inputs separately. The connections between neurons are weighted according to w_i , which means that some inputs impact the neuron's output more than others. These weights are modified during training to optimize the ANNs' performance in a specific job. (Alwosheel et al., 2018). The output a can be calculated as follows:

$$a = z(\nu) = z \cdot \left(\sum_{i=1}^I w_i t_i + w_b \right) \quad (2.63)$$

where w_b is the weight associated with the bias.

An ANN parameter might change based on the purpose it is supposed to execute. However, some typical specs are as follows:

- **Number of Layers:** Between the input and output layers, ANNs can include one or more hidden layers (Alwosheel et al., 2018).
- **Number of Neurons:** The number of neurons in each layer varies based on the task's difficulty (Alwosheel et al., 2018).
- **Activation Function:** The activation function governs how the neuron receives input and produces output (Alwosheel et al., 2018). Table 2.11 presents the main functions used for the matter.

Table 2.11: Activation functions for ANNs (Alwosheel et al., 2018)

Activation Function Name	Function
Step Function	$z = \begin{cases} 0, & \text{if } \nu \leq 0 \\ 1, & \text{if } \nu > 0 \end{cases}$
Rectifier Linear Unit (ReLU) Function	$z = \max(0, \nu)$
Sigmoid Function	$z = [1 + \exp(-\nu)]^{-1}$
Tangent Sigmoidal Function	$z = \tanh(\nu)$

- **Loss function:** The loss function is used to change the weights during training by measuring how well the ANN performs on a particular task. The error function is defined as follows, assuming the data consists of N choice observations over K options (Alwosheel et al., 2018):

$$J(W) = - \sum_{n=1}^N \sum_{k=1}^K y_{nk} \cdot \ln(P_{nk}) \quad (2.64)$$

where,

$J(W)$ - Error function.

W - Vector with the estimated w parameters.

y_{nk} - An indicator that denotes whether alternative k is chosen in observation n .

P_{nk} - Choice probability.

- **Learning rate:** The amount by which the weights are altered during training is controlled by this option. The backward propagation procedure comprises taking the partial derivative of the error J with respect to the weights, known as the gradient vector g , to update the weights. In addition to a η learning rate value, W values are re-adjusted as follows (Alwosheel et al., 2018):

$$W_{p+1} = W_p + \eta_p g_p \quad (2.65)$$

where p is the step index.

- **Optimization algorithm:** Various optimization techniques, such as stochastic gradient descent (SGD) or Adam, can be employed to change the weights during training (Alwosheel et al., 2018).

The main ANNs models employed in this dissertation will be presented in the next sections.

2.7.10.1 Feed-Forward Neural Networks

Feed-forward Neural Networks (FNNs) is an ANN instance widely utilized in pattern recognition and classification, system modelling and identification, signal processing, image processing, control systems, and stock market forecasts. They are distinguished by their parallel structure and their capacity to learn and adapt to new information. FNNs are classified as "single layer" or "multi-layer" based on the number of layers (Sazli, 2006), as seen in Figures 2.25 and 2.26.

The back-propagation method is the most often used learning/training algorithm for FNNs. The network is regarded as an FNN if there is no "feedback" from the neurons' outputs to the inputs throughout the network (Sazli, 2006).

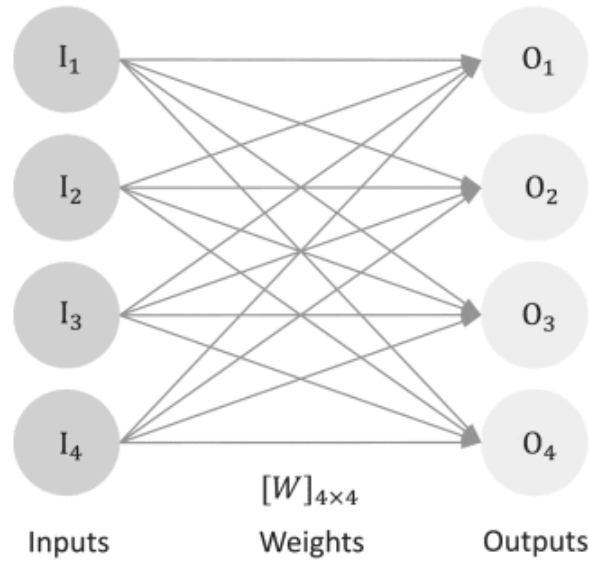


Figure 2.25: Single Layer FNN (Shokraneh et al., 2019)

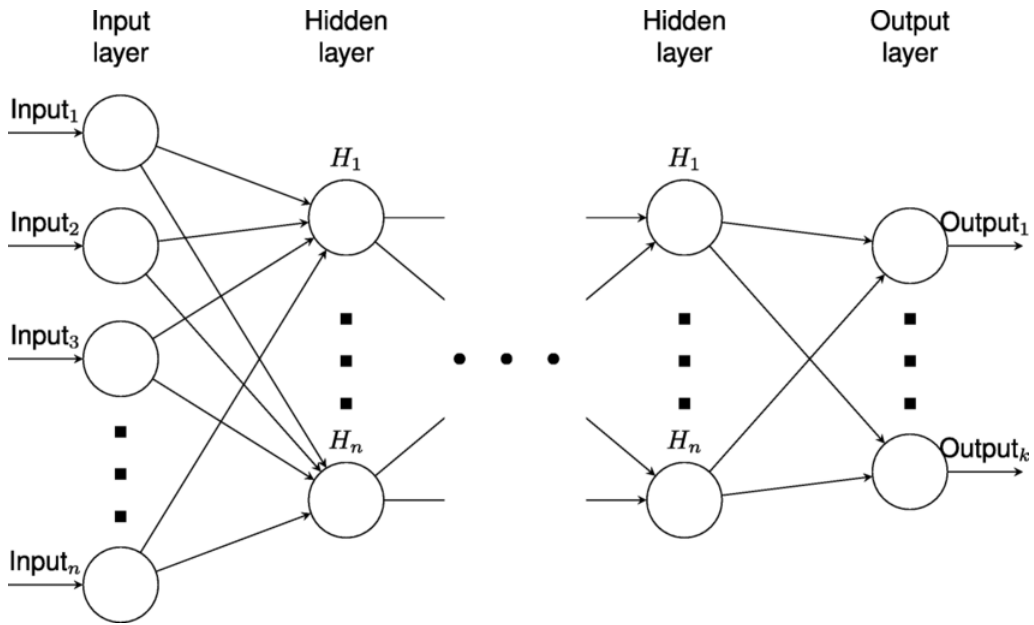


Figure 2.26: Multiple Layer FNN (Tong et al., 2019)

2.7.10.2 Extreme Learning Machines

ELM is an abbreviation for Extreme Learning Machine, a sort of machine learning method used for classification, clustering, and regression problems. It is a single-hidden layer

FNN that distributes input weights randomly and calculates output weights analytically to achieve rapid learning speed and high accuracy. Medical imaging, chemistry, transportation, economy, robotics, management, geography, and food safety are some areas where ELM has been used effectively (Wang et al., 2022).

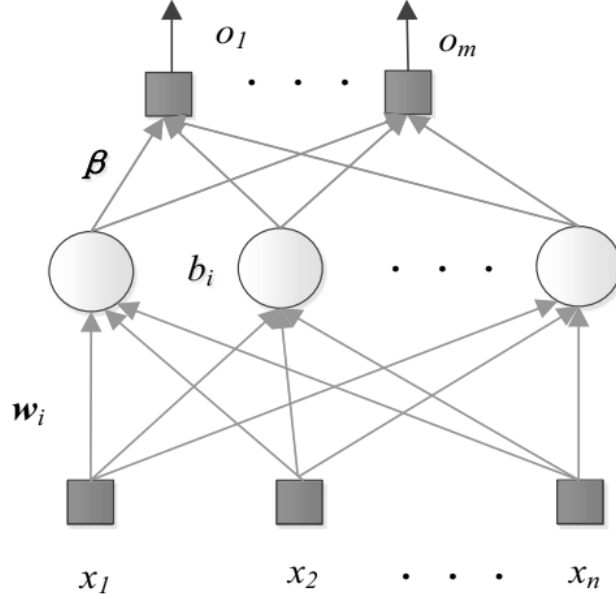


Figure 2.27: ELM structure (Wang et al., 2022)

ELM works by randomly assigning input weights and biases to hidden layer neurons and then calculating the output weights analytically using a pseudo-inverse matrix. During the training phase, ELM receives a set of input data and associated output labels and utilizes these to alter the hidden layer neurons' input weights and biases. Once the input weights are fixed, it computes the output weights using a pseudo-inverse matrix. ELM employs these fixed input and output weights throughout the testing phase to predict new output labels for unknown input data. Due to the random assignment of input weights, ELM can learn significantly quicker than typical Neural Networks (NNs) while maintaining good accuracy (Wang et al., 2022).

Consider that, for a given single hidden layer of an ELM, the output function of the i -th node is given by $h_i(x) = G(a_i, b_i, x)$, in which a_i and b_i are parameters of the i -th node. The output function, for L hidden nodes is given by:

$$f_L(x) = \sum_{i=1}^L \beta_i h_i(x) \quad (2.66)$$

where β_i is the output weight of the i -th hidden node.

Given N training samples and considering $h(x) = [h_1(x), \dots, h_L(x)]$ as the hidden layer output mapping, the hidden layer output matrix is given by:

$$H = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix} = \begin{bmatrix} G(a_1, b_1, x_1) & \dots & G(a_L, b_L, x_1) \\ \vdots & \vdots & \vdots \\ G(a_1, b_1, x_N) & \dots & G(a_L, b_L, x_N) \end{bmatrix} \quad (2.67)$$

and T , the training data target matrix is given as follows:

$$T = \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix} \quad (2.68)$$

The objective function that requires to be minimized is:

$$\|\beta\|_p^{\sigma_1} + C\|H\beta - T\|_q^{\sigma_2} \quad (2.69)$$

where $\sigma_1 > 0$, $\sigma_2 > 0$, $p, q = 0, 0.5, 1, 2, \dots, +\infty$.

2.7.10.3 Multi-Layer Perceptron

According to Daniel H. (2019), a Multi-Layer Perceptron (MLP) is made up of an input layer, at least one hidden linear threshold unit (LTU) layer, and an output LTU layer. A Deep Neural Network (DNN) is formed when an MLP contains two or more hidden layers.

The hidden layers comprise numerous neurons layered together, with a non-linear mapping between inputs and outputs. The MLP was created to address the perceptron's constraint of only being applicable to linear data (Carolina Bento, 2021).

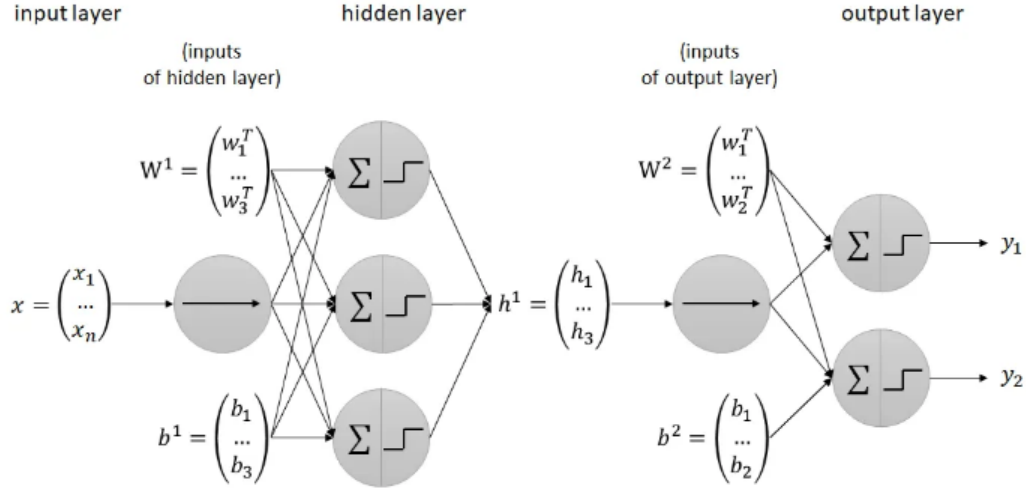


Figure 2.28: Example of an MLP with three LTUs in the hidden layer and two LTUs in the output layer (Daniel H., 2019)

Consider a perceptron as a basic ANN, built on a single layer of LTUs, with each LTU linked to all vector x inputs and a bias vector b , as seen in Figure 2.29.

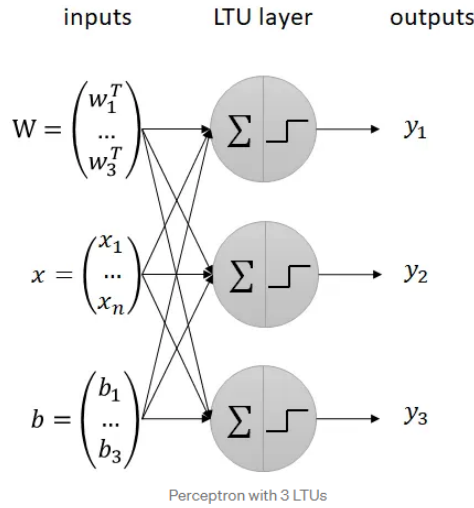


Figure 2.29: Perceptron with 3 LTUs (Daniel H., 2019)

Each output vector y can be calculated as follows:

$$y = \text{step}(W \cdot x + b) \quad (2.70)$$

in which,

step - refers to a step function.

$$W = \begin{bmatrix} w_1^T \\ \vdots \\ w_u^T \end{bmatrix}, \text{ matrix of shape } (u, n).$$

u - number of LTUs.

n - number of input values.

The mathematical formulation of the MLP presented in Figure 2.28 follows the same formulation that the perceptron one, and the variables h and y are calculated in the following manner:

$$h^1 = \text{step}(z^1) = \text{step}(W^1 \cdot x + b^1) \quad (2.71)$$

$$y = \text{step}(z^2) = \text{step}(W^2 \cdot h^1 + b^2) \quad (2.72)$$

If a general case is considered, in which, when training, k instances are retrieved from the m available instances, then:

$$y = \text{step}(Z) = \text{step}(X \cdot W^T + b) \quad (2.73)$$

where,

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_k^T \end{bmatrix}, \text{ is a matrix of shape } (k, n).$$

The Industrial Water Supply System

This chapter will provide an overview of the present water supply system and explain its operation and main problems.

3.1 The Water Supply System

The purpose of an industrial water supply system is to fulfil the demands of individuals who depend on it and industrial activities by delivering a reliable and regular supply of high-quality water. Water source, intake structure, treatment structure, distribution system, storage tanks, and monitoring and control systems are some of the common parts of the system. A factory top view is shown in Figure 3.1 while some of these infrastructures' locations are presented in Figures 3.2 and 3.3.



Figure 3.1: Location of the Toyota Caetano Ovar plant - satellite top view



Figure 3.2: Location of the boreholes and ETAR

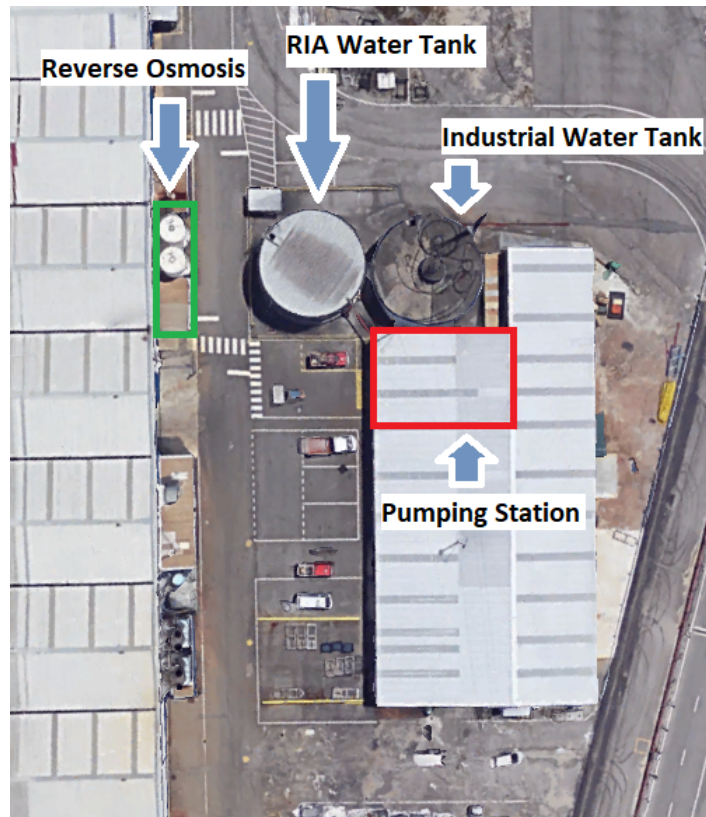


Figure 3.3: Location of the main equipment of the water supply system

3.1.1 Pumping Station

Factory 1's general water supply system is equipped with an EFAFLU AQUA Premium CFD system, presented in Figure 3.4. It consists of two vertical electric pumps in parallel with an independent frequency variation and has an electrical panel and corresponding flow limit deposit. A technical drawing of this pumping station and fire protection network (that will be presented in the next section) is presented in Annex C.

Its operation can be merged, individually or together. Prepared to work 24 hours a day always to maintain the same pressure in the circuit. The maximum pressure of the pumps is 11 bar, and it has its efficiency in practice for a flow of $18 \text{ m}^3/\text{h}$.

This system's pumps are BMV16-80 vertical multistage pumps powered by normal electric motors. The electric motors installed are 7.5 kW and 2900 rpm , with the Y2 132S2-2 model. The available power is 400V at 50 Hz. This pump has a nominal flow of $16 \text{ m}^3/\text{h}$ and can work in a range between 8 and $22 \text{ m}^3/\text{h}$. The maximum working pressure is 22 bar, and the maximum efficiency is 66%.



Figure 3.4: EFAFLU AQUA Premium CFD system

3.1.2 The Fire Protection System - RIA

One of the most critical components of the water supply system is the Fire Protection System (in Portuguese, Rede de Incêndios Armada - RIA), which is also located in the Pumping Station. Its primary purpose is to give water to the fire brigade and factory in the case of a fire. The water supply is enabled by two pumps, which run individually.

In the case of an emergency, the first pump to work is the electrical pump, presented in Figure 3.5. It is an Eco-line ECO315S-2 that, when working at 50 Hz, provides 110 kW and 2980 rpm.

In the event of failure of the electric pump, the diesel pump, seen in Figure 3.6, starts working automatically. It is a GOULDS 3408 8x6x12 F-M, with a diesel engine of 123 kW working at 2000 rpm. The capacity of supplying a maximum of 315 m³/h of water.



Figure 3.5: Eco-line ECO315S-2 electrical pump



Figure 3.6: GOULDS 3408 8x6x12 F-M diesel pump

If the network pressure decreases to a value below 8.5 bar, the jockey pump starts working and increases the pressure until it achieves 10 bar. The pump, presented in Figure 3.7, is a WBJP 5-16 and has an available rating of 2.2 kW. Capacity to supply $4m^3/h$ of water at 10 bar.

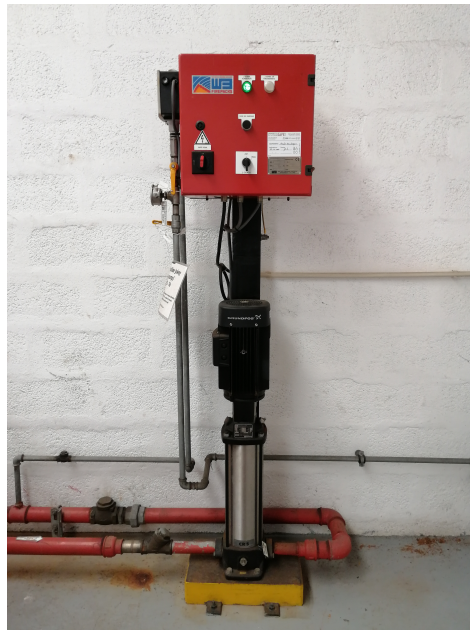


Figure 3.7: WBJP 5-16 jockey pump

3.1.3 Reverse Osmosis - Demineralized Water

Reverse Osmosis (RO) is a pressure-driven process that removes dissolved components from feed water owing to size, charge, and physical-chemical interactions between solute, solvent, and membrane (Malaeb and Ayoub, 2011).

The process, depicted in Figure 3.8, has the following steps:

- **Pre-Treatment:** Pretreatment of feed water removes suspended particles, organic debris, and other impurities that may foul or harm the RO membrane (Malaeb and Ayoub, 2011).
- **Pressurization:** To force the water molecules through the semi-permeable membrane and overcome the osmotic pressure of the dissolved components, the pre-treated feed water is pressured (Malaeb and Ayoub, 2011).
- **Separation:** The semi-permeable membrane divides the input water into the concentrate (wastewater) and permeate (treated water) streams (Malaeb and Ayoub, 2011).
- **Post-treatment:** Before being put to use or discharged, the permeate may go through post-treatment to change its pH, mineral content, or other qualities (Malaeb and Ayoub, 2011).

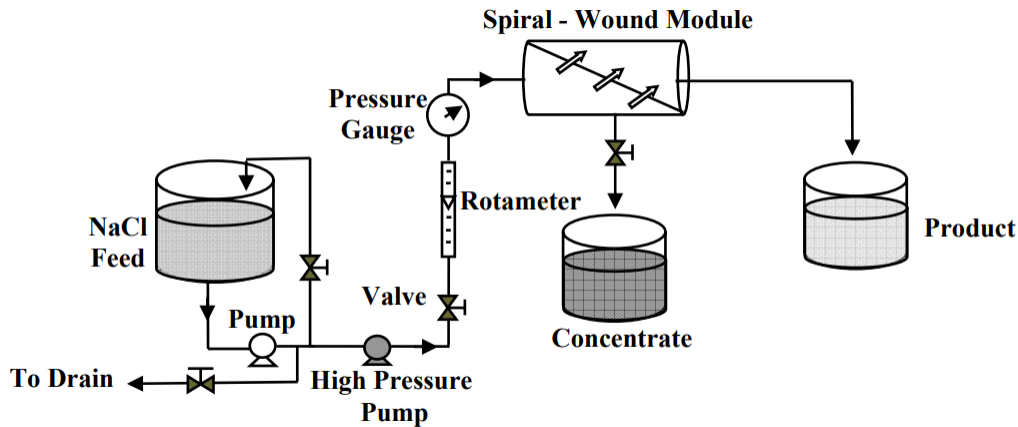


Figure 3.8: Reverse osmosis process scheme (Rassoul et al., 2012)

Operating factors, including pressure, flow rate, recovery ratio, and membrane and feed water characteristics, affect how effective the RO process is (Malaeb and Ayoub, 2011). The factory process is depicted in Figure 3.9.



Figure 3.9: Reverse osmosis process in the plant

3.2 System Analysis

This section will review the water supply process utilizing an *AS IS* - *TO BE* approach.

3.2.1 *AS-IS*

The first step in the *AS-IS* approach is to have an overview of the water supply system. For that, a meeting in the department was organized to investigate and evaluate the main problems regarding the system.

From the meeting, several problems related to the water supply process were exposed:

- There are errors in recording water consumption data when inserting it into the database in Excel.
- Method of recording water consumption is archaic - manual method.
- There is no careful and critical analysis of recorded consumption.
- Lack of predictive consumption analysis methods.

- Lack of maintenance to the pipes of the water supply system.
- Missing PokeYoke system to warn of possible leaks in the water supply system.
- Inexistence of dashboards for process and consumption visualization in real-time.

The final analysis is presented in an Ishikawa diagram, as seen in Figure 3.10, that was completed during the meeting.

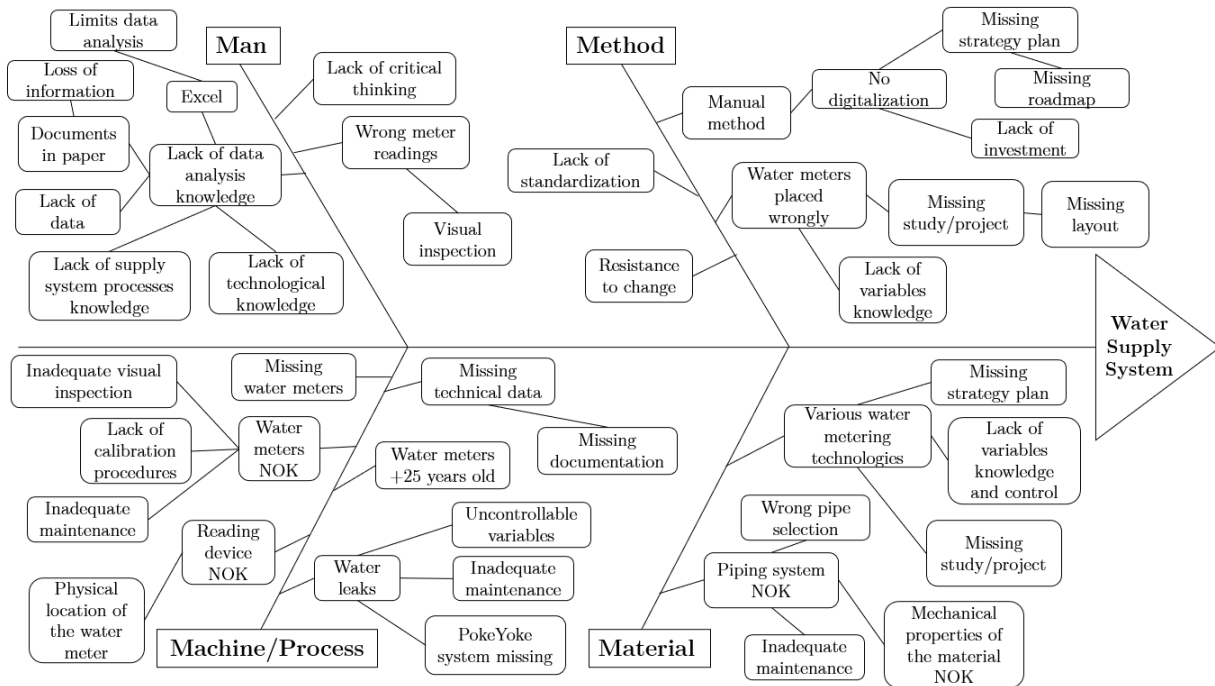


Figure 3.10: Water supply system Ishikawa diagram

For a company like TCAP, very strict water consumption management is expected, considering the metrics planned for the Toyota Environmental Challenge 2050 and external audits. This justifies a new management process and optimization of water consumption based on the principles of digitalization and Industry 4.0.

Chapter 6 will provide the *TO-BE* and *TO-DO* analysis, the presentation of improvement proposals for optimization and management of the supply system and the digitalization plan for water management.

Industrial Water Dashboard

One of the main problems regarding factory management was tracing and monitoring each sector's water consumption effectively. At the time of writing, this was done by one person, using an *Excel* spreadsheet that registered the values of each water meter daily or monthly. Thus, creating a quick and easy way to visualize the water consumption data was paramount. The solution was the creation of an industrial water dashboard.

According to Few (2006), dashboards can be defined as "a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance." The same author also determined what are the key aspects that a dashboard must have:

- Very well-organized.
- Condensed, with emphasis on summaries and exceptions.
- Specific and tailored to the dashboard's audience and goals.
- Displayed utilizing simple and frequently small media that provides the data and message most cleanly and directly attainable.

However, one question arises before even developing the dashboard: *What metrics need to be calculated and displayed for a broad and effective understanding of the factory's water consumption?* One approach that effectively answers this question is the Goal Question and Metric approach, which will be further explained in the next section.

4.1 Goal Question And Metric Approach

The GQM model was created by Victor Basili and his colleagues in 1984. It is a top-down technique that proposes that measurement goals should guide software development measurements. As the name suggests, it is composed of three important phases/stages, present in Figure 4.1:

1. **Goal - Conceptual Level:** A goal is defined for an item for a number of reasons, with respect to various quality models, from diverse points of view, in relation to a certain context (Basili et al., 1994).
2. **Question - Operational Level:** Based on a defining model, a series of questions is employed to establish how the assessment/achievement of a given objective will be executed (Basili et al., 1994).
3. **Metric - Quantitative Level:** A data collection is attached with each question to provide a quantitative answer (Basili et al., 1994).

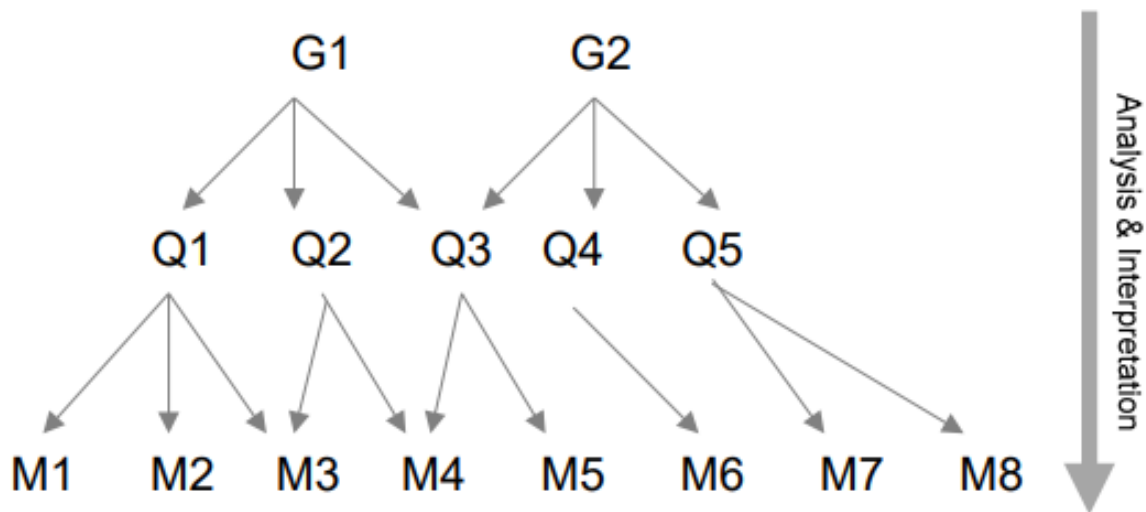


Figure 4.1: GQM notation in three levels - G: Goal, Q: Question, M: Metric (Asghari, 2012)

Although this approach is most commonly used in software development, it may be utilized in this situation to clarify and justify which metrics should be included in the dashboard layout.

4.2 GQM Applied To The Industrial Water Dashboard

After presenting how the GQM model works, the focus is now on how this model can be implemented to find which metrics are the most useful. The main question that the factory board of directors wants to be answered is *"How to evaluate the factory's water consumption in an effective way?"*.

According to the annual environmental report that the factory's environment department presents at the end of each year, three key goals were needed to achieve a full understanding of the water consumption:

- Water consumption by sector.
- Average water consumption by sector.
- Key Performance Indicator (KPI) of water consumption by car produced.

For the water consumption and average water consumption goal, it became essential to obtain the water meter records, both the day and night ones. Albeit the average consumption is a consequence of a series of calculations from the day and night records, these were considered independent metrics according to established goals that were defined previously.

On the other hand, in relation to the water KPI, and to establish a fair comparison ground, four metrics were necessary: the target car production and target KPI (established by the factory's engineering and production board), the effective car production and the Factory 1 total water consumption.

In the end, the proposed GQM model was constructed and is presented in Figure 4.2.

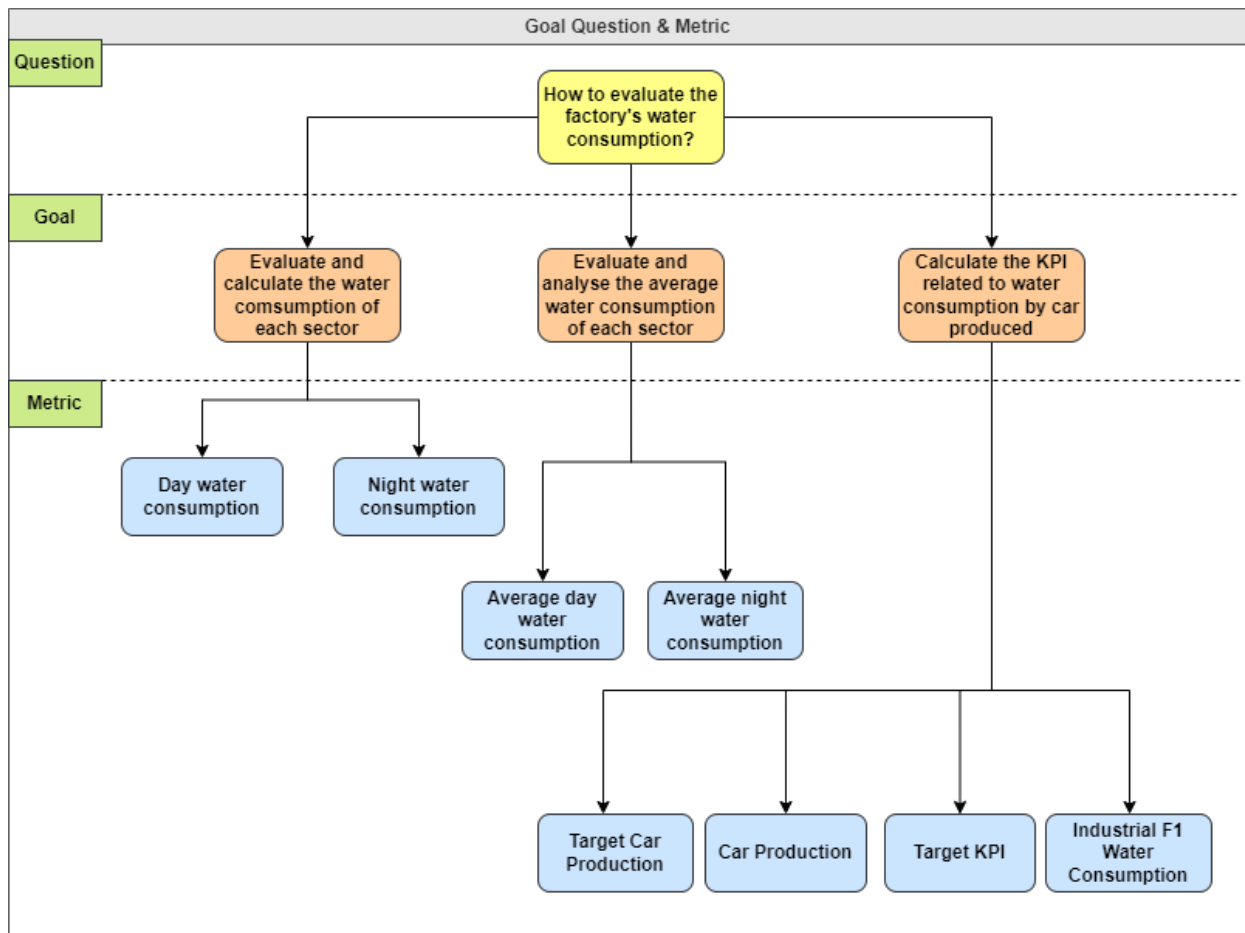


Figure 4.2: GQM model to obtain the metrics for the industrial water dashboard

4.3 The Industrial Water Dashboard

The previous section established the main metrics needed for the dashboard implementation with the GQM approach. The next step is the development of the industrial water dashboard. However, it is still important to establish the following guidelines:

- System requirements and Unified Modeling Language (UML) diagram.
- Data pre-processing.
- Database creation and maintenance.
- Connection between database and *PowerBI*.
- Dashboard UI development.

4.3.1 System Requirements and UML Diagram

Prior to the dashboard development, the requirements that the system must fulfil need to be identified. The following items are the information system's primary requirements:

- The dashboard will be utilized mainly by technicians and engineers.
- Each water meter must have its own dashboard page that includes the following information: meter location, meter picture (if available), total month consumption, day and night average month consumption, and the corresponding month consumption bar graph.
- The user, on the water meter page, must provide a set of filters for selecting the year, month, and day of the week. The page must be immediately updated with the selections made.
- The type of day filters the data according to the factory's production status or season:
 1. In production.
 2. Stoppage due to equipment breakdown.
 3. Vacations.
 4. National/Regional holidays.
 5. Planned yearly maintenance.
 6. Planned production stop.

- The goal vehicle production, car production, Industrial F1 water consumption, target water KPI, and water KPI must all be included on the Water KPI dashboard page, together with a yearly KPI bar chart.
- The Water KPI dashboard must have the same year and month filters as the water meter dashboard page.

Having characterized the main user requirements, the following stage is to create the associated UML diagram. UML stood for Unified Modeling Language and was created in 1996 by Grady Booch, Ivar Jacobson, and James Rumbaugh (Booch et al., 2005). The UML is a set of established best engineering practices for modelling large and complex systems. It employs graphical notations to explain the architecture of software projects, supporting project teams in communicating, exploring new designs, and verifying the program’s architectural design of the program (Visual Paradigm, 2022).

Several iterations of the UML diagram were created, the first one (handwritten) presented in Annex D and the final in Figure 4.3. Each table of the UML diagram is going to be further clarified next:

- **Water Meter ID:** This table aims to introduce other water meters easily if needed. Each water meter has its own ID number and given name.
- **Water Meter Daily/Monthly X:** The X in this table denotes the ID of the water meter. According to the periodicity of the records, the water meter will be subdivided into monthly or daily. These tables will have all the records, daily or monthly, as exposed before.
- **Water KPIs:** In this table, all the information regarding the monthly water KPIs and monthly target values of car production and industrial water consumption.
- **Factory Calendar:** The factory calendar is introduced in the system with the aid of this table.
- **Calendar Stoppages ID:** Each stoppage will have its own ID number to ease its introduction into the system.
- **Type of Stoppage ID:** The type of stoppage is the same as the ones that were previously defined in the user requirements of the system. Each one will have its unique ID number.
- **Car Prod:** This table will have the daily records of the car production of the factory.

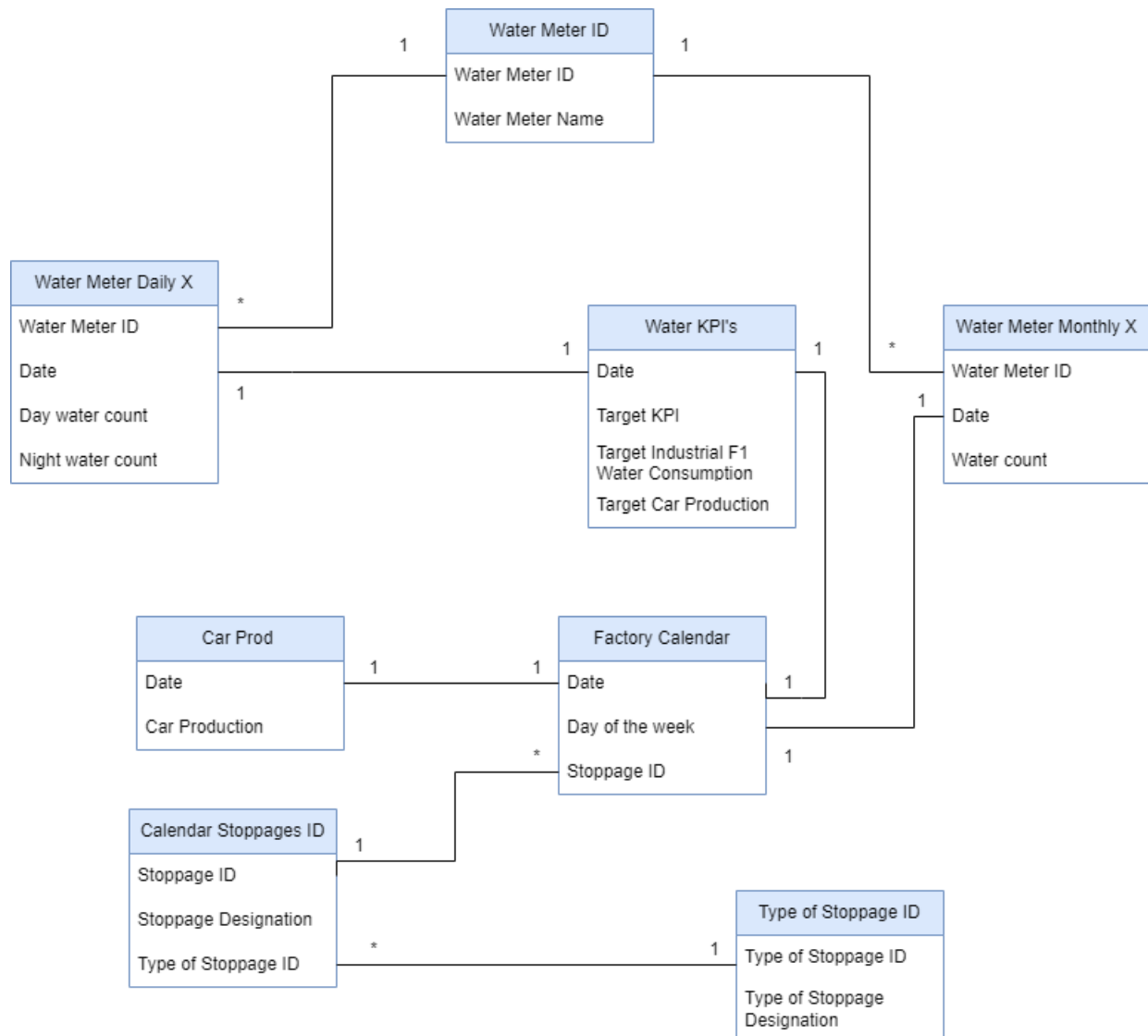


Figure 4.3: Final UML Diagram for the industrial water dashboard

4.3.2 Data Pre-Processing And Preparation

The most difficult aspect in the design and creation of this dashboard was undoubtedly data pre-processing. Data from each water meter was stored in monthly Excel files from 2018 to 2020 and annual Excel files from 2021 to 2022. As a result, it was challenging to enter all of the information into a single Excel file - it took about a month to get all the data. This case highlights the relevance and beneficial influence that I4.0 may have on the company. All data and the resulting KPI reports would have been freely accessible if its pillars were applied.

4.3.3 *MySQL* Database

After retrieving the data from 2018 to 2022, the next step was the data introduced in the *MySQL* database. The main goal of creating this database was not to rely solely on Excel files for water records. The database was then connected with *PowerBI* using the existent *MySQL* connector supplied by the software.

4.3.4 *PowerBI* Dashboard

The final step was the creation of the dashboard, which took approximately one month and a half to be created and tested. The metrics present in the dashboard were created using the DAX (Data Analytics Expressions), a function and operator library that can be used to generate formulae and expressions in *PowerBI*.

4.3.4.1 Main Menu

The main menu, presented in Figure 4.4, comprises two buttons: one for the Consumption Records dashboard page and another for the Industrial Water KPIs.

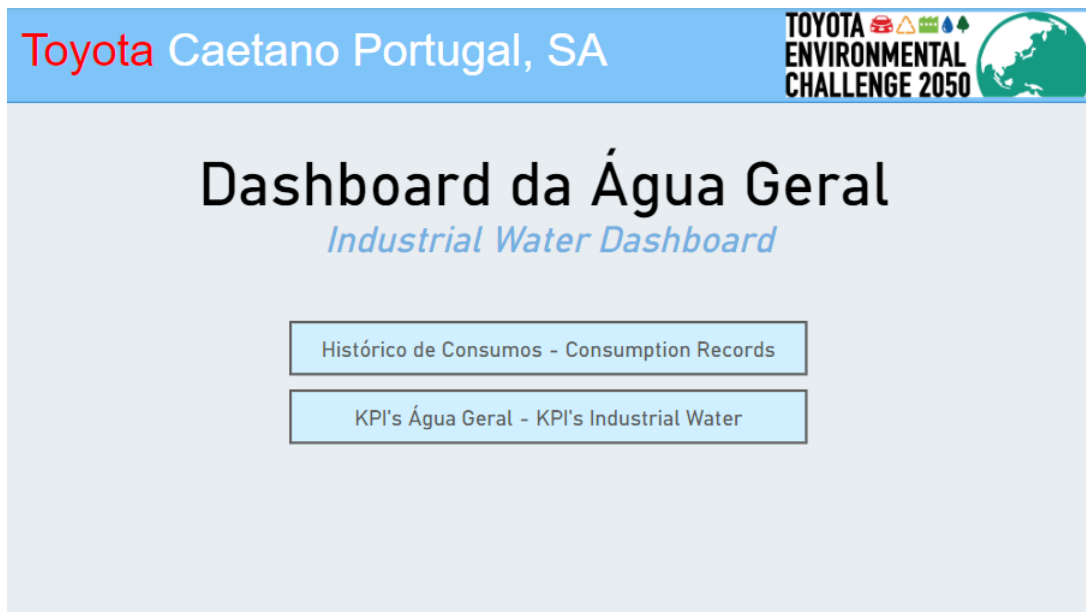


Figure 4.4: Dashboard main page

4.3.4.2 Water Consumption Records Menu

The user can select the desired water meter using the buttons on this menu, as seen in Figure 4.5.

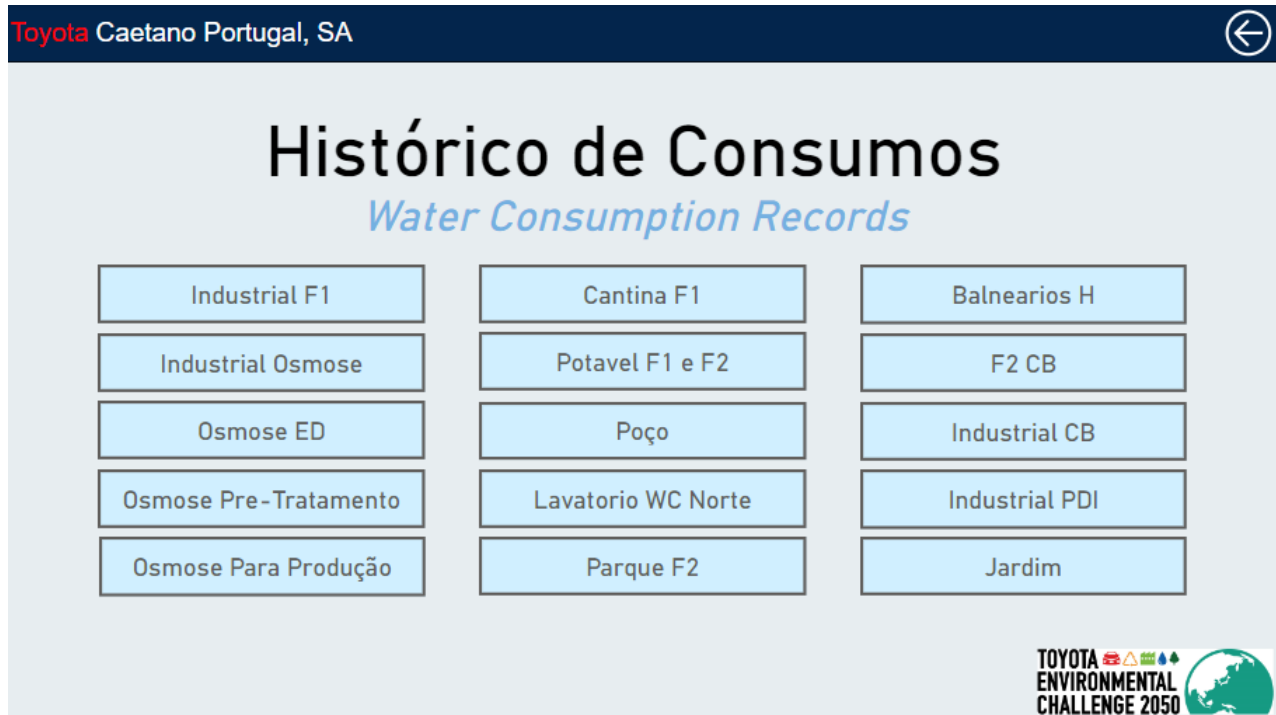


Figure 4.5: Consumption records menu

4.3.4.3 Water Consumption Records Dashboard

There are two types of consumption record pages, depending if the water meter record is made daily, like in Figure 4.6, or monthly, present in Figure 4.7.

The following specifications are present in the daily record dashboard:

- Water meter picture (if available) and location.
- Filters to choose the type of day, month and year.
- Information about the total, average daily day and average daily night consumption.
- Bar chart indicating the daily water consumption according to the month that was chosen.

The monthly record dashboards follow the same approach; however only provide the average consumption per day, total monthly consumption, and a bar graph representing the monthly consumption in the selected year.

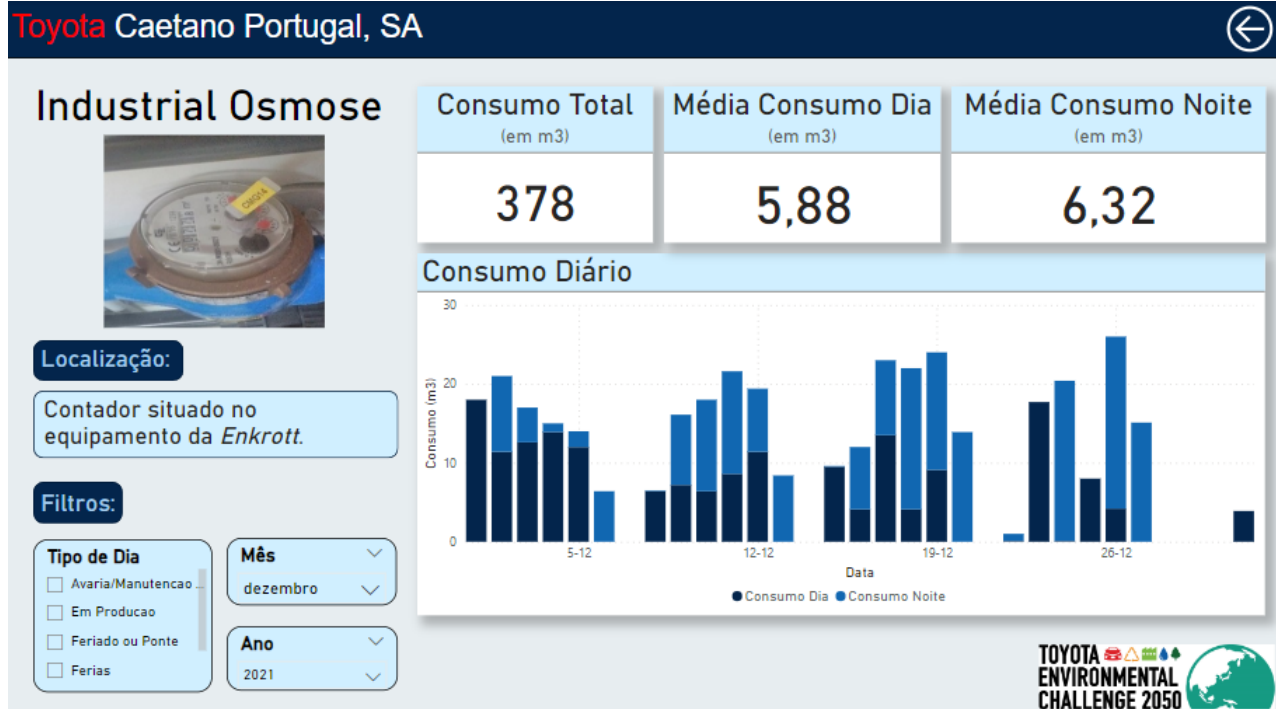


Figure 4.6: Daily consumption records type menu

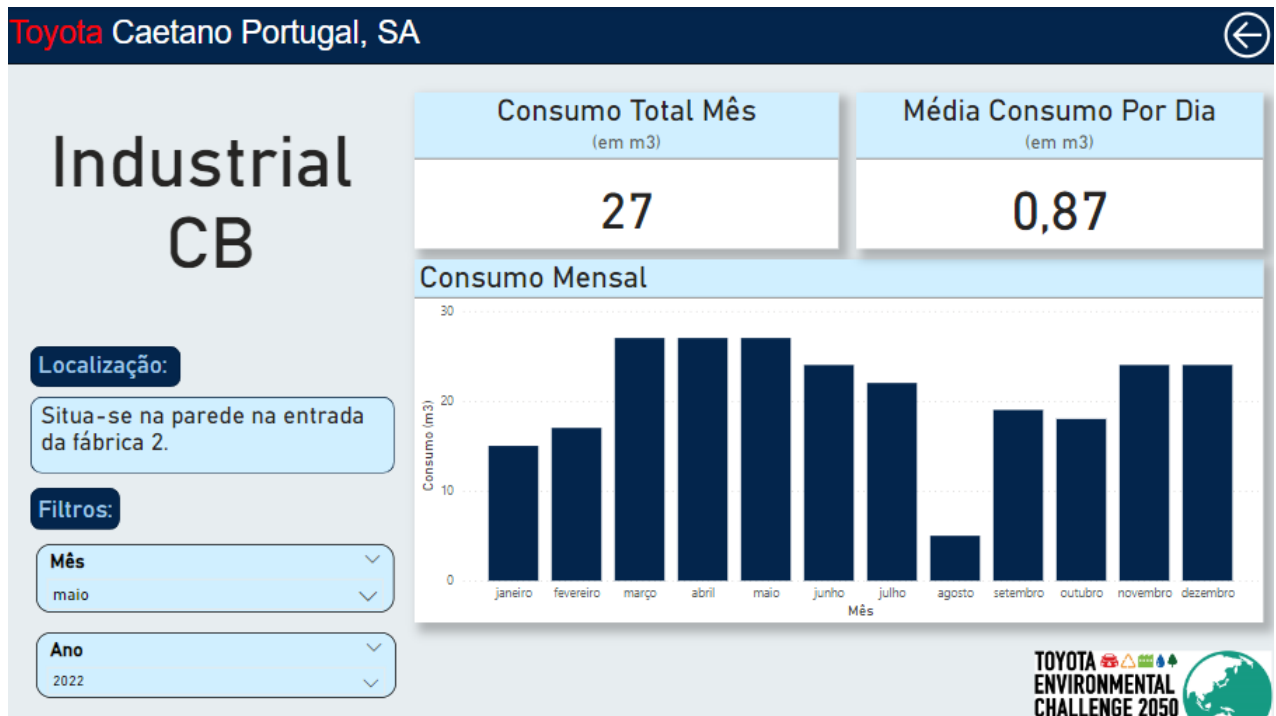


Figure 4.7: Monthly consumption records type menu

- The KPI related to water consumption per vehicle produced. The red line in the gauge meter indicates the target KPI relative to the month and year chosen.
- Target vehicle production and the vehicle production for the month and year in question.
- A bar chart indicates the chosen month's daily vehicle production.
- Water KPI evolution throughout the chosen year. The line in red indicates the target KPI for each month of the year.



Water Consumption Analysis and Prediction

Data analysis and prediction models, which may be used to forecast demand, predict equipment malfunctions, and enhance production processes, are critical components of I4.0. These models can assist firms in lowering costs, improving efficiency, and gaining a competitive advantage.

One of the main concerns regarding the plant's water supply system was the overall water consumption of the factory. Thus, the presented challenges were creating a prediction model for water consumption and analysing and evaluating water consumption related to vehicle production.

Kotu and Deshpande (2018) defined a generic process for a data analysis study, which is present in Figure 5.1. Such a model will be the main methodology for this study.

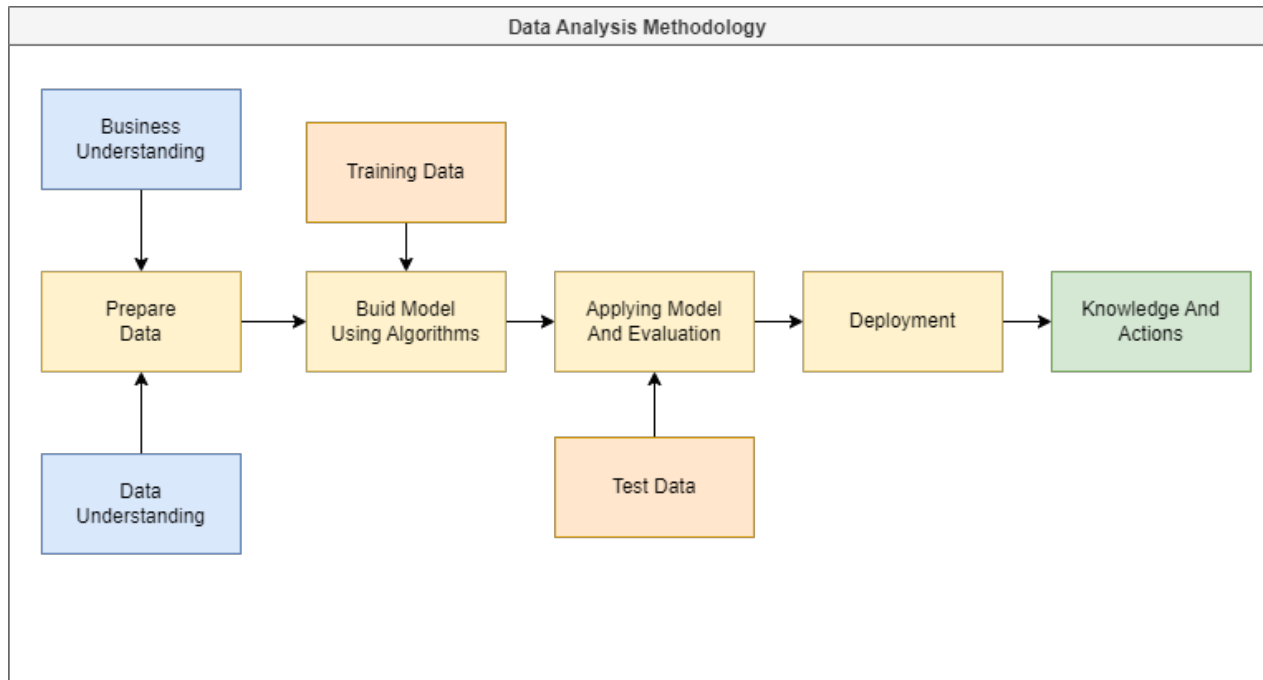


Figure 5.1: Data science process (Adapted from Kotu and Deshpande (2018))

5.1 Water Consumption Analysis In Car Production

The initial analysis was performed on the quantity of water utilized by each car manufactured. The information originated from the Industrial Water database, which compiles monthly automobile manufacturing numbers and industrial and demineralized water use from 2018 to 2022. Throughout this data pre-processing phase, there were no missing values. The code created in R for this section is presented in Annex E.

5.1.1 Industrial Water Consumption Analysis

After performing the data pre-processing, one must evaluate the existence or not of outliers in the dataset. An algorithm was created in the *R* language, using the *ols_plot_resid_lev* function to evaluate what was needed. The results are presented in Figure 5.2. This made it possible to detect which values were considered outliers or leverage.

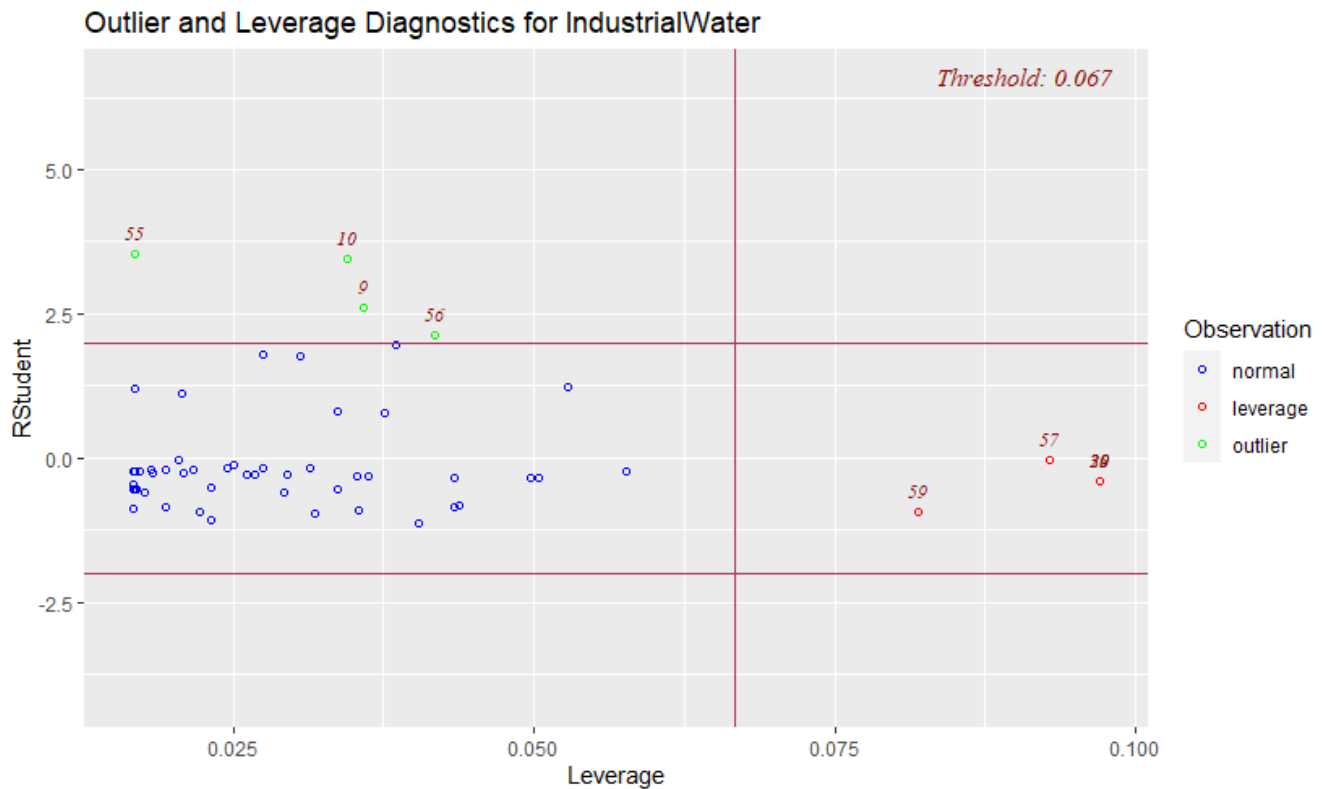


Figure 5.2: Outlier detection in the industrial water dataset

The next stage in the research was to run a robust regression to predict the factory's industrial water usage for each car produced. In order to have a comparison ground, an

ordinary least square regression was also employed.

The robust regression was carried out using the *R* programming language and the *lmrob* function, which allows the creation of this sort of regression using the MM-Estimators parameter. The OLS regression was performed using the *lm* function. Results are presented in Table 5.1 and Figure 5.3, where IWC stands for *Industrial Water Consumption* and MVP for *Monthly Vehicle Production*.

Table 5.1: Results from the linear regression analysis of the industrial water consumption

Regression Type	Equation	r^2
OLS	$IWC = 5.0898 \cdot MVP + 184.3661$	0.446
Robust	$IWC = 4.7087 \cdot MVP + 50.6295$	0.876

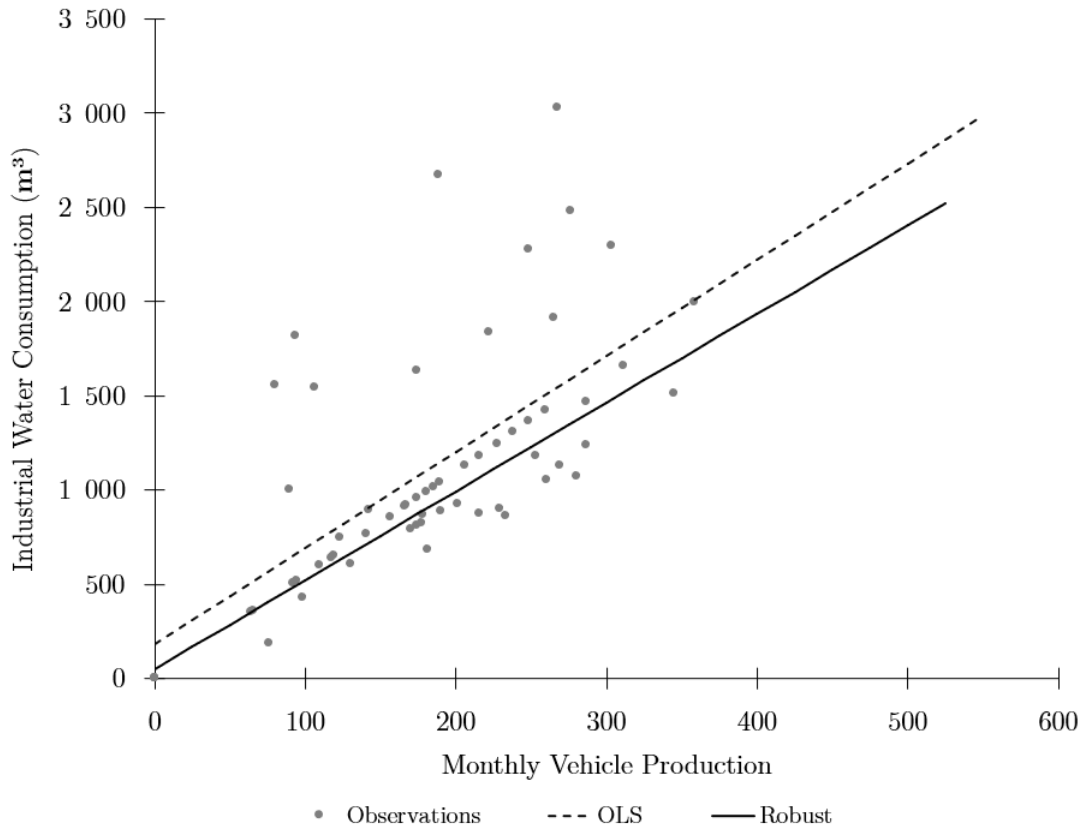


Figure 5.3: Industrial Water Consumption Analysis

The r^2 value clearly shows that the robust regression model fitted the presented dataset better. This could be explained by the fact that, as stated in Chapter 2, this sort of regression decreases the degree of importance that the outliers have in the data. Such removes the

negative influence that those have on the analysis. The robust regression also suggests that the KPI for industrial water usage per car manufactured is $4.71 \text{ m}^3/\text{vehicle}$.

5.1.2 Demineralized Water Consumption Analysis

The same analysis was carried out for demineralized water consumption. Figure 5.4 presents the outlier detection test for the demineralized water dataset, and Table 5.2, together with Figure 5.5, shows the final results of the analysis. Note that DWC stands for *Demineralized Water Consumption* and MVP for *Monthly Vehicle Production*.

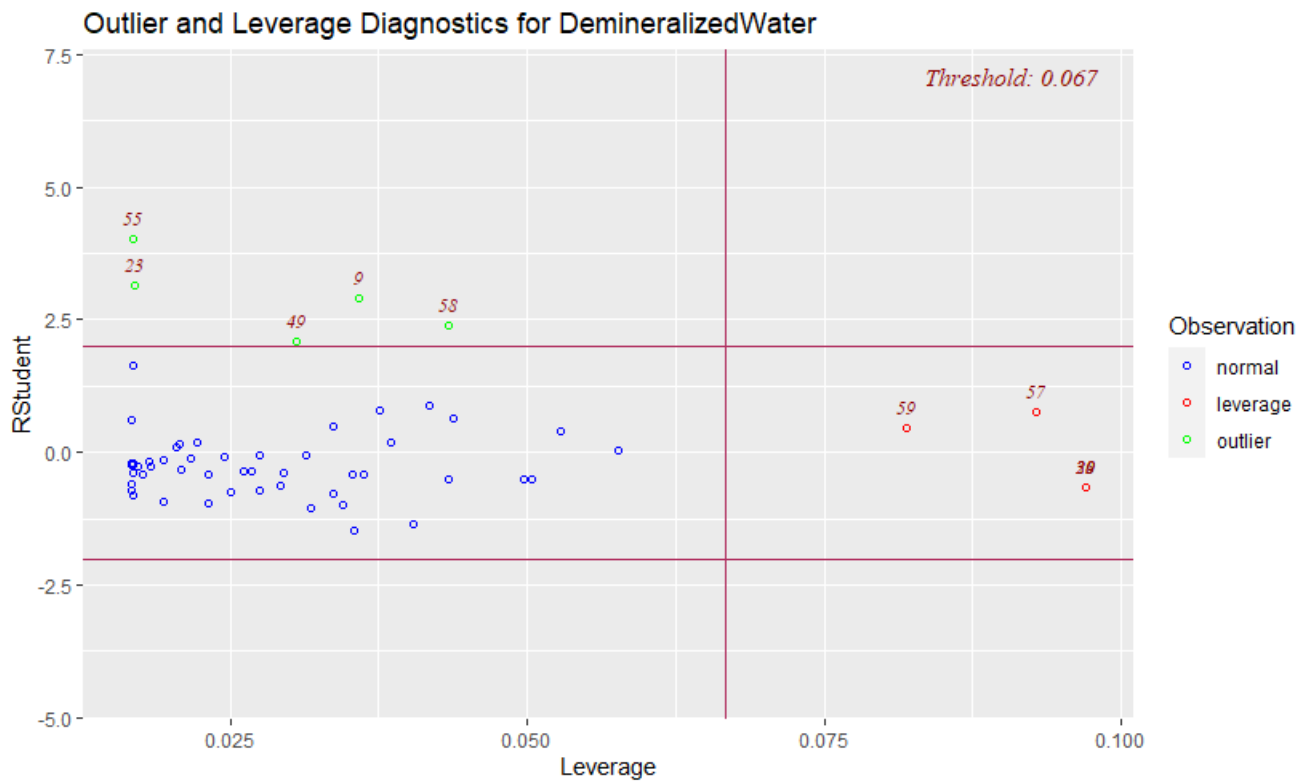


Figure 5.4: Outlier detection in the demineralized water dataset

Table 5.2: Results from the linear regression analysis of the demineralized water consumption

Regression Type	Equation	r^2
OLS	$\text{DWC} = 1.3136 \cdot \text{MVP} + 51.3700$	0.649
Robust	$\text{DWC} = 1.4512 \cdot \text{MVP} + 3.0635$	0.925

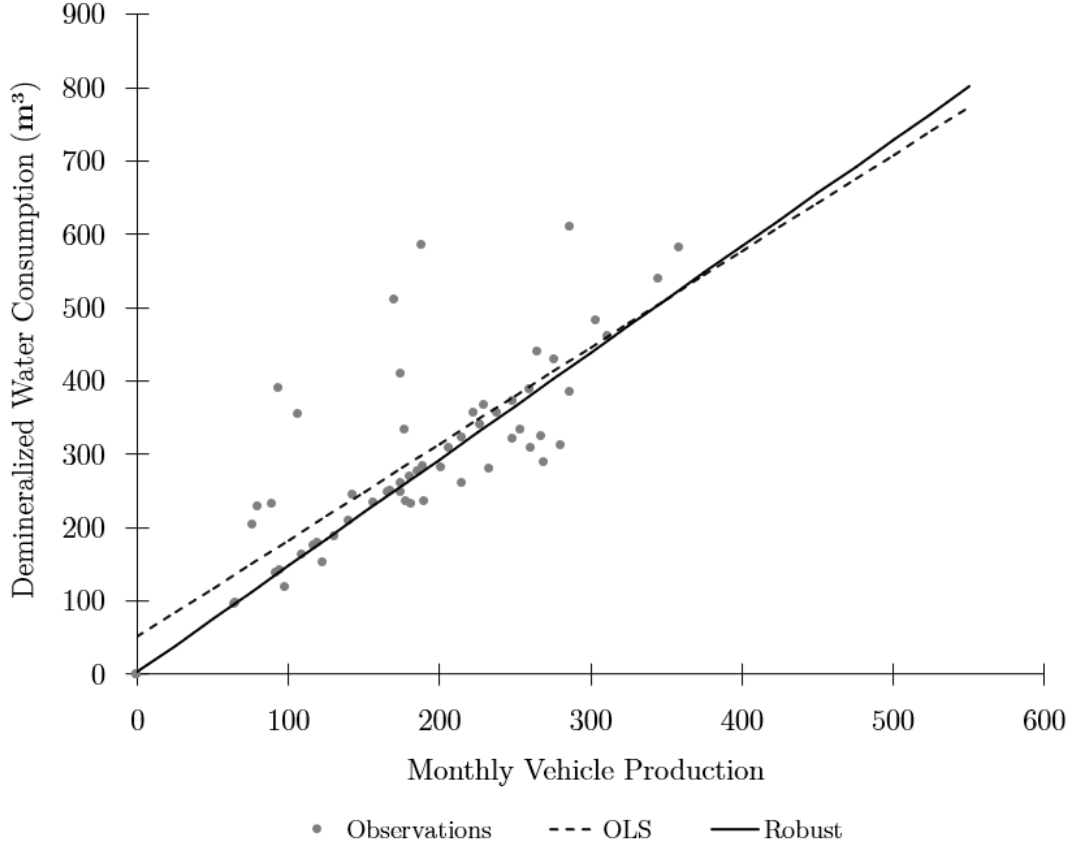


Figure 5.5: Demineralized Water Consumption Analysis

From the results, one can state that the robust regression outperformed, in a substantial way, the OLS regression, achieving a r^2 of almost 1. This also indicates that the robust regression is close to perfectly capturing the behaviour of the demineralized water consumption. Thus, it is suggested that the KPI for demineralized water consumption per car manufactured is $1.45 \text{ m}^3/\text{vehicle}$.

5.1.3 Analysis Conclusions

The robust regression model, using the MM-Estimator, proved to be the best for estimating both the industrial water consumption - $r_{IWC}^2 = 0.876$ - and the demineralized water consumption - $r_{DWC}^2 = 0.925$. Table 5.3 presents the estimates for water consumption according to daily car production.

Table 5.3: Predicted value for water consumption according to the daily production

Daily Car Production	Industrial Water Predicted Consumption (m^3)	Demineralized Water Predicted Consumption (m^3)	Total Predicted Daily Consumption (m^3)
10	97.72	17.58	115.29
11	102.43	19.03	121.45
12	107.13	20.48	127.61
13	111.84	21.93	133.77
14	116.55	23.38	139.93
15	121.26	24.83	146.09
16	125.97	26.28	152.25
17	130.68	27.73	158.41
18	135.39	29.19	164.57

5.2 Water Consumption Analysis In The Paintshop

Another interesting analysis that was made was the water consumption in the two most important paint shop line processes: the Pre-Treatment section (PT) and the Electrophoretic Deposition section (KTL). The same methodology that was employed in the last section was used in this analysis. The data was retrieved from the industrial water company's database, tracing the time period between 2018 and 2022. The code created in R for this section is presented in Annex F.

5.2.1 Industrial Water Consumption Analysis in Pre-Treatment

The results are presented in Figures 5.6 and 5.7 and Table 5.4. The outlier analysis revealed a significant number of outliers as well as leveraged values. As a result, the OLS regression fared so poorly ($r^2 = 0.119$) compared to the robust regression ($r^2 = 0.997$). It should be observed that the robust regression effectively captured the linear behaviour of water consumption. From the analysis, one could state that a KPI for the industrial water used in the Pre-Treatment for each vehicle produced is $0.513 \text{ m}^3/\text{vehicle}$.

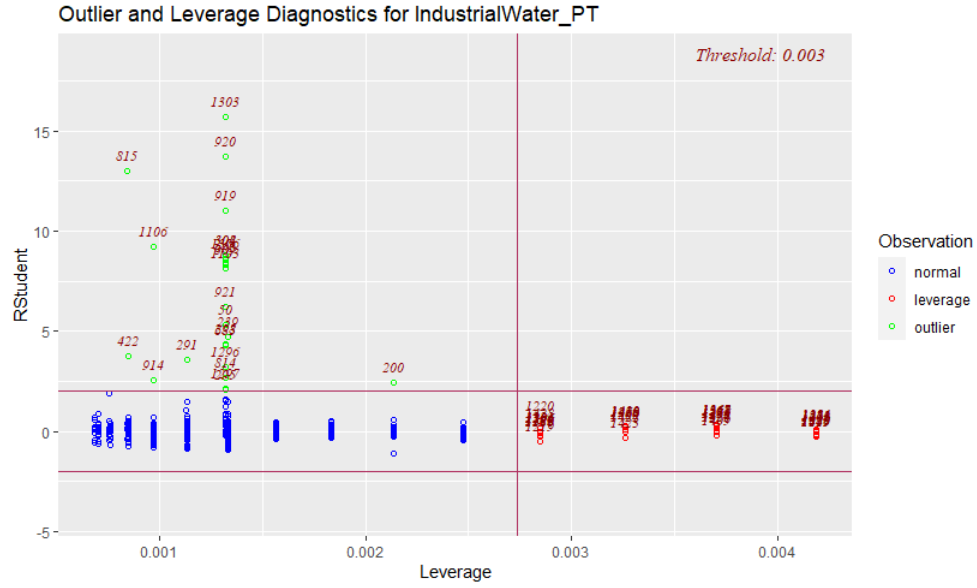


Figure 5.6: Outlier detection in the industrial water dataset - PT

Table 5.4: Results from the linear regression analysis of the industrial water consumption

Regression Type	Equation	r^2
OLS	$IWC = 0.36355 \cdot DVP + 1.29786$	0.119
Robust	$IWC = 0.513435 \cdot DVP + 0.003468$	0.997

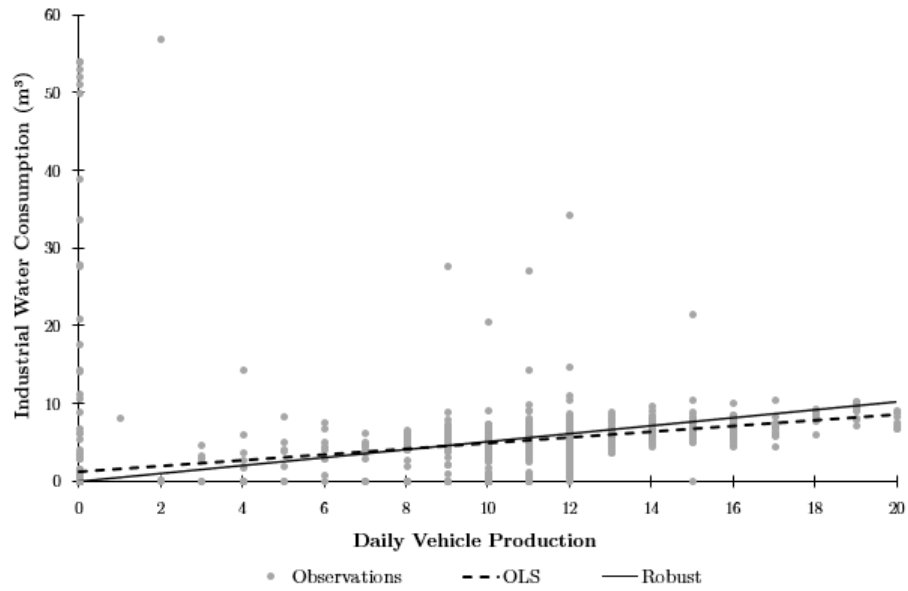


Figure 5.7: Industrial Water Consumption Analysis - PT section

5.2.2 Demineralized Water Consumption Analysis in Pre-Treatment

Figures 5.8 and 5.9, as well as Table 5.5, illustrate the results. The outlier analysis discovered a large number of outliers as well as leveraged values. Consequently, the OLS regression performed much worse ($r^2 = 0.201$) than the robust regression ($r^2 = 0.973$).

The robust regression caught the linear tendency of water use rather well. Based on the investigation, a KPI for the demineralized water utilized in the Pre-Treatment for each car manufactured is $0.683 \text{ m}^3/\text{vehicle}$.

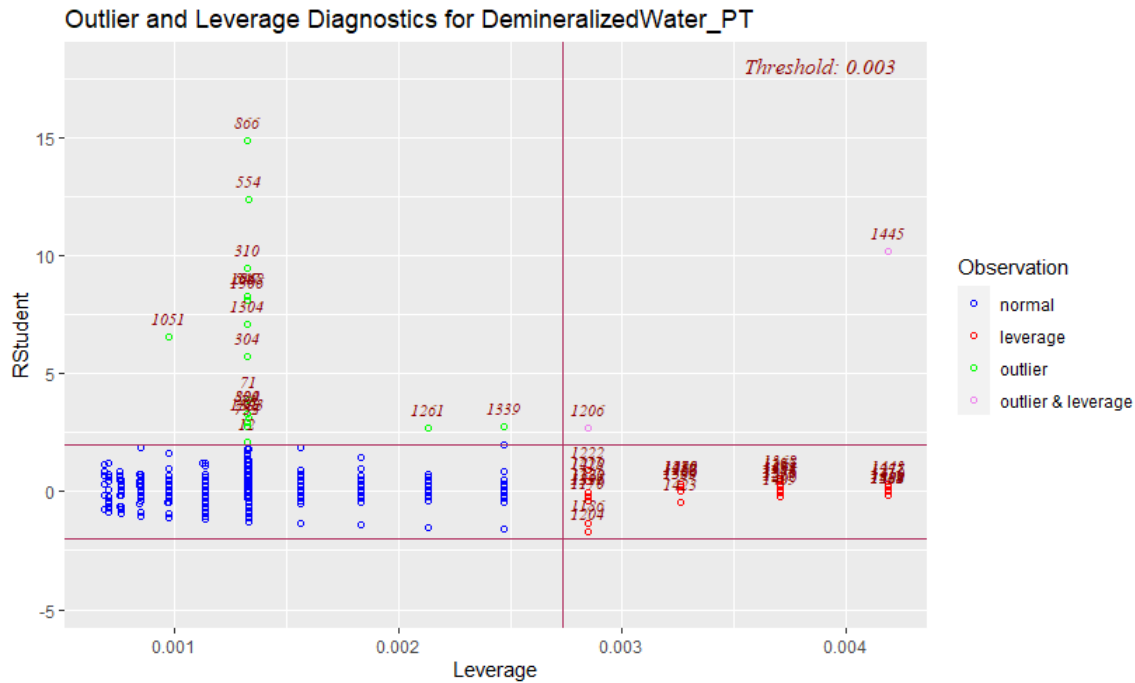


Figure 5.8: Outlier detection in the demineralized water dataset - PT

Table 5.5: Results from the linear regression analysis of the demineralized water consumption - PT

Regression Type	Equation	r^2
OLS	$DWC = 0.50258 \cdot DVP + 1.80044$	0.201
Robust	$DWC = 0.683433 \cdot DVP + 0.025828$	0.973

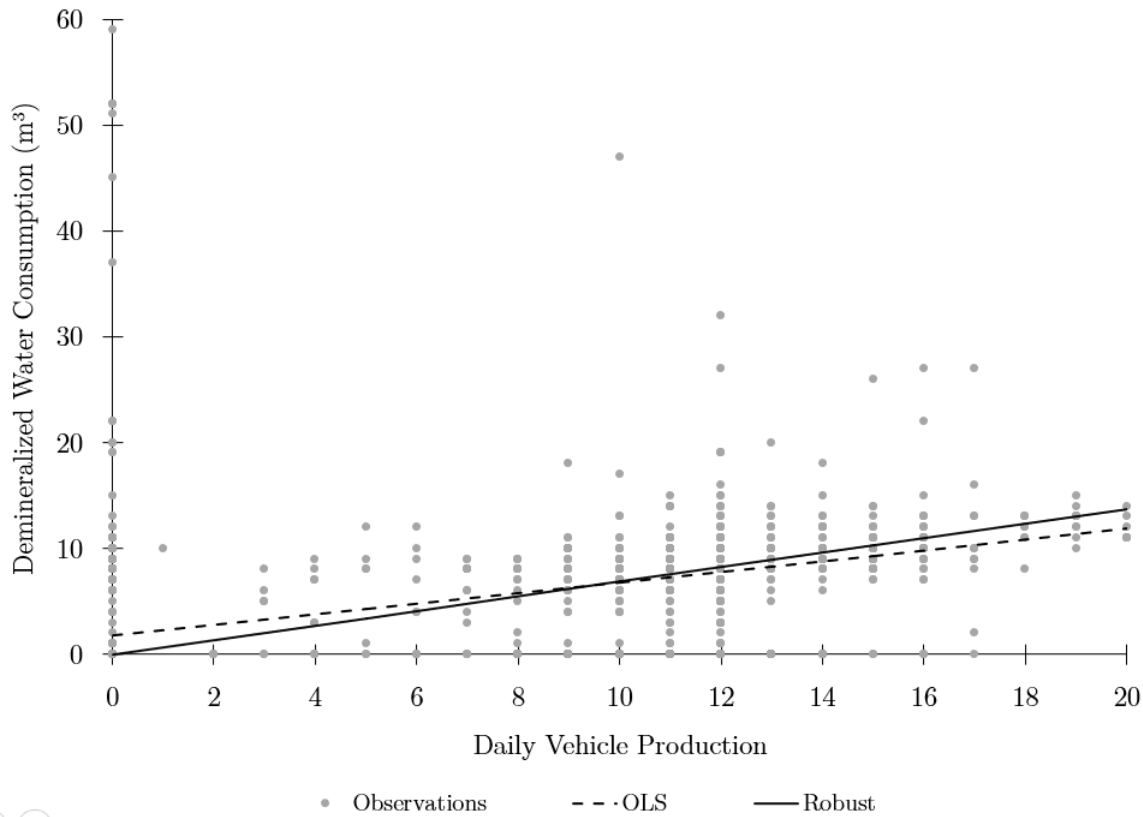


Figure 5.9: Demineralized Water Consumption Analysis - PT

5.2.3 Demineralized Water Consumption Analysis in KTL

The findings are depicted in Figures 5.10 and 5.11, as well as Table 5.6. The outlier analysis identified several outliers as well as leveraged values.

This analysis concluded that the behaviour of the consumption of demineralized water in the KTL section could not be modelled as linear. Both the results of the OLS ($r^2 = 0.008$) and robust regression ($r^2 = 0.449$) support that conclusion. The results could possibly be attributable to the KTL process relying on a highly stringent chemical balance; therefore, the water consumption pattern will not be as regular as originally predicted.

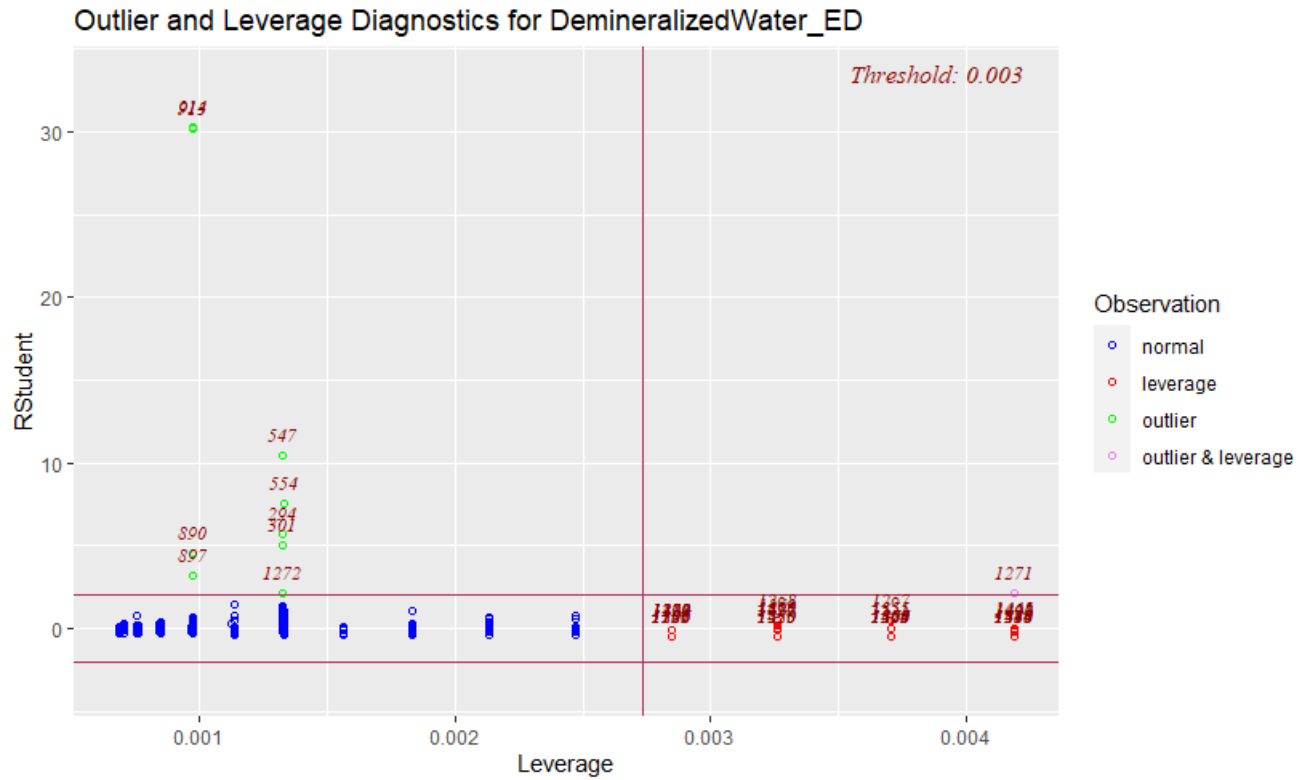


Figure 5.10: Outlier detection in the demineralized water dataset - KTL

Table 5.6: Results from the linear regression analysis of the industrial water consumption

Regression Type	Equation	r^2
OLS	$DWC = 0.24197 \cdot DVP + 3.00147$	0.008
Robust	$DWC = 0.35599 \cdot DVP + 0.77957$	0.449

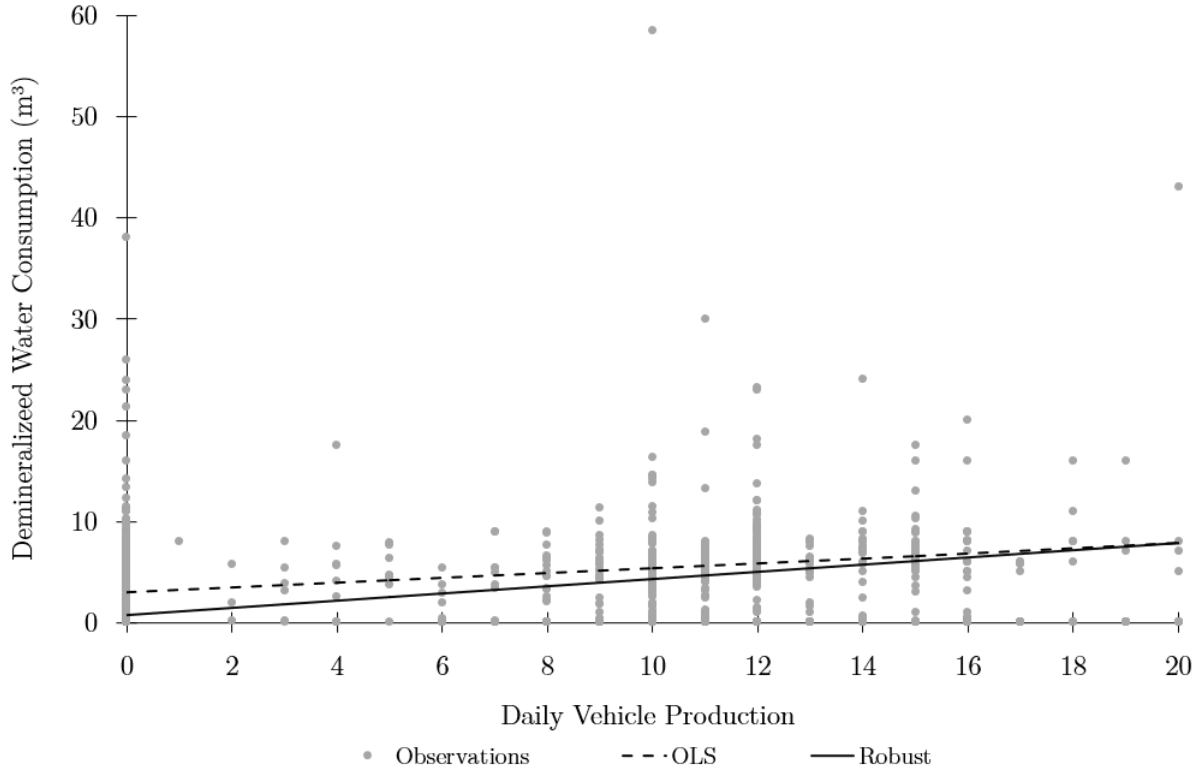


Figure 5.11: Demineralized Water Consumption Analysis - KTL

5.2.4 Analysis Conclusions

The robust regression model with the MM-Estimator was found to be the best for estimating both industrial ($r_{IWC}^2 = 0.997$) and demineralized water ($r_{DWC}^2 = 0.973$) used in the PT section. However, no conclusion could be achieved from the robust analysis regarding the demineralized water consumption in the KTL section. Table 5.7 shows the estimated water use based on daily automobile manufacturing.

Table 5.7: Predicted value for water consumption according to the daily production - PT and KTL

Daily Car Production	Industrial Water PT Predicted Consumption (m^3)	Demineralized Water PT Predicted Consumption (m^3)	Demineralized Water KTL Predicted Consumption (m^3)
10	5.130	6.830	No conclusion achieved.
11	5.643	7.513	
12	6.156	8.196	
13	6.669	8.879	
14	7.182	9.562	
15	7.695	10.245	
16	8.208	10.928	
17	8.721	11.611	
18	9.234	12.294	

5.3 Water Consumption Prediction Model

5.3.1 Data Pre-Processing

The data was collected from the company's water consumption database, built-in *Microsoft Excel*. This database contained data from 2018 to 2022. Hence, the water consumption in each plant sector was collected each month of the year. The data was then treated, resulting in Figures 5.12, 5.13, and 5.14.

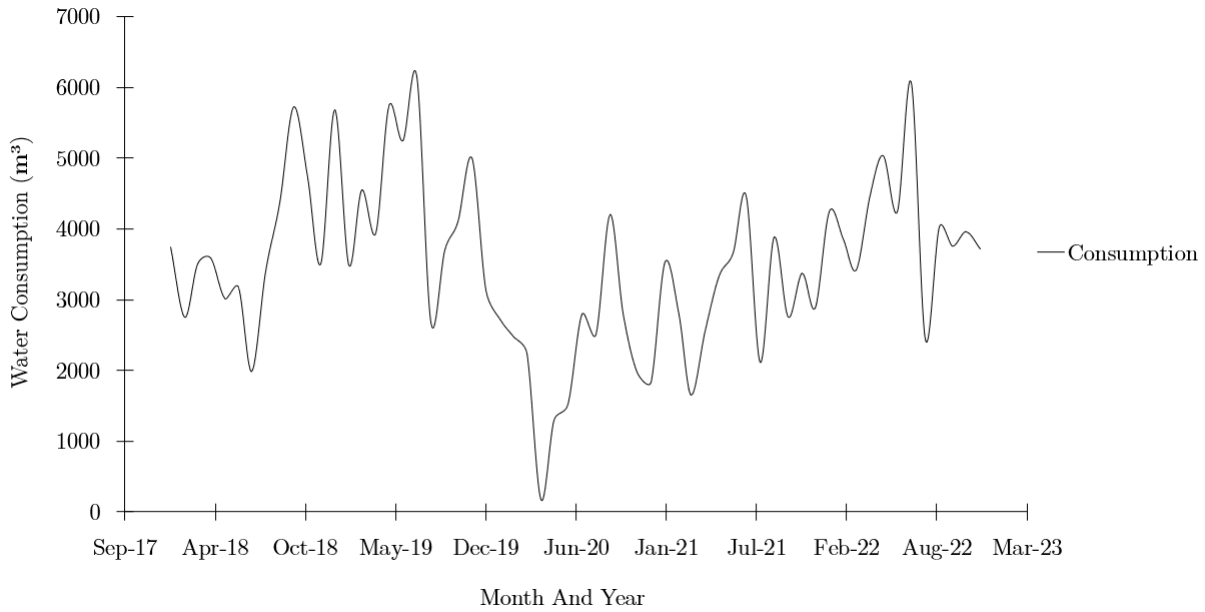


Figure 5.12: Total water consumption - 2018 to 2022

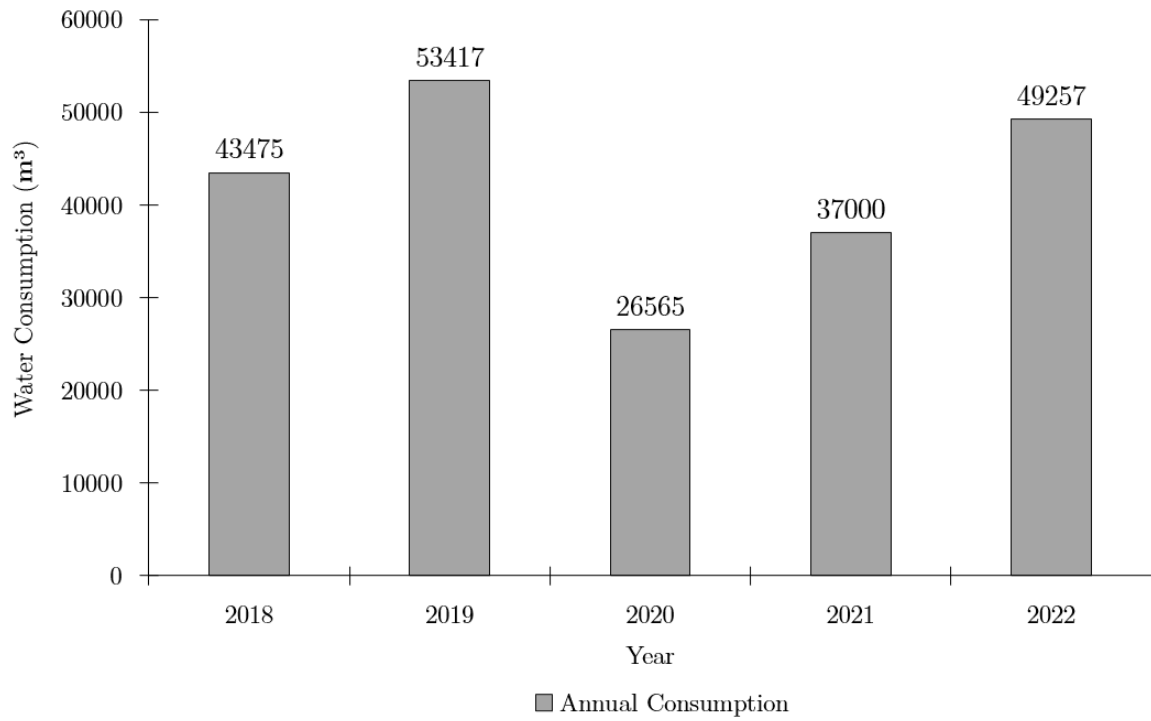


Figure 5.13: Annual water consumption - 2018 to 2022

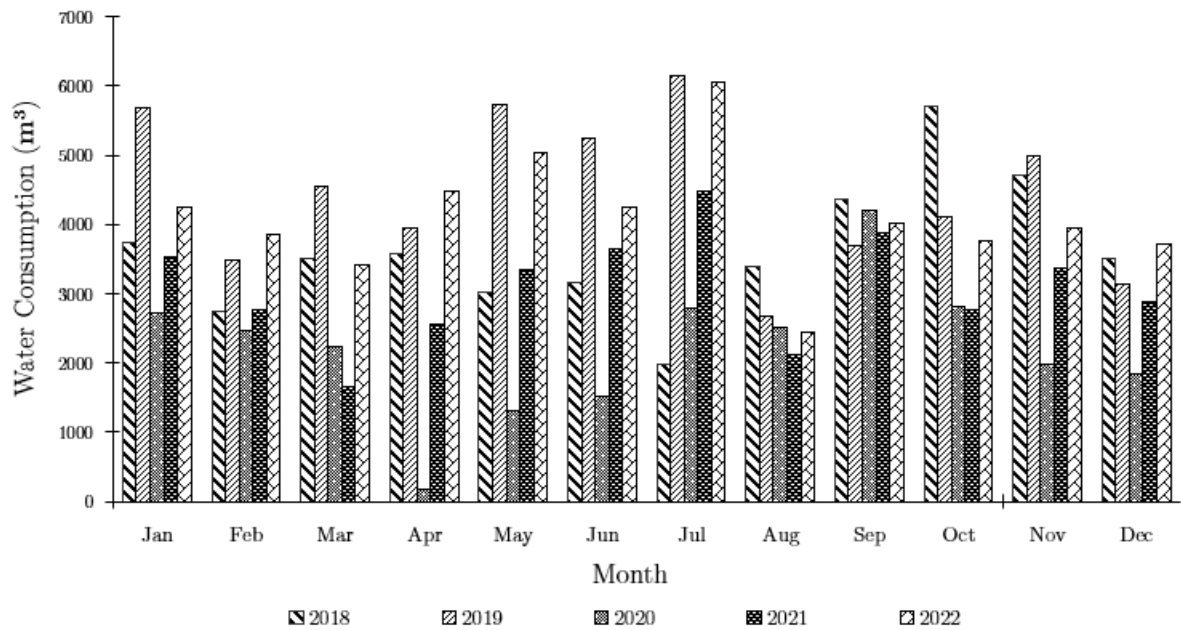


Figure 5.14: Monthly water consumption - per year

Figure 5.12 demonstrates a substantial drop in water use in 2020 compared to the prior year. This phenomenon is explained by the Portuguese government's first COVID-19 lockdown. As of 2018, the data does not reflect the same trends as 2019, 2021 and 2022, as seen in Figure 5.14. Therefore, the data from 2018 and 2020 was not included in the final dataset required for model implementation.

5.3.2 Outlier Detection With Hampel Filtering

A robust outlier identification analysis employing the Hampel filter was performed to assess the existence of outliers in the dataset and reduce their impact on the analysis. The Hampel filter uses the Median and Mean Absolute Deviation (MAD) to determine whether a particular time-series point is an outlier in a particular moving window (MathWorks, 2019).

Given a sequence $x_1, x_2, x_3, \dots, x_n$ and a sliding window of length k , one can define:

$$m_i = \text{median}(x_{i-k}, x_{i-k+1}, x_{i-k+2}, \dots, x_i, \dots, x_{i+k-2}, x_{i+k-1}, x_{i+k}) \quad (5.1)$$

$$\sigma_i = \kappa \cdot \text{median}(|x_{i-k} - m_i|, \dots, |x_{i+k} - m_i|) \quad (5.2)$$

$$\kappa = \frac{1}{\sqrt{2} \cdot (\text{erf})^{-1}(1/2)} \approx 1.4826 \quad (5.3)$$

where erf is the associated error function.

If a sample x_i is such that $|x_i - m_i| > n_\sigma \cdot \sigma_i$, for a given threshold n_σ , then the Hampel filter identifies x_i as an outlier and it replaces with m_i (MathWorks, 2019).

The function truncates the window used to compute m_i and σ_i near the sequence endpoints:

If $i < k + 1$:

$$m_i = \text{median}(x_1, x_2, \dots, x_i, \dots, x_{i+k-2}, x_{i+k-1}, x_{i+k}) \quad (5.4)$$

$$\sigma_i = \kappa \cdot \text{median}(|x_1 - m_1|, \dots, |x_{i+k} - m_i|) \quad (5.5)$$

If $i > n - k$:

$$m_i = \text{median}(x_{i-k}, x_{i-k+1}, x_{i-k+2}, \dots, x_i, \dots, x_{n-2}, x_{n-1}, x_n) \quad (5.6)$$

$$\sigma_i = \kappa \cdot \text{median}(|x_{i-k} - m_i|, \dots, |x_n - m_n|) \quad (5.7)$$

Figures 5.15 and 5.16 present the industrial water dataset before and after the Hampel filter implementation. It can be clearly noted that the 800 m^3 instance in the original dataset was detected by the Hampel filter, and replaced with a lower value, according to the algorithm previously explained. Thus, it can be concluded that the Hampel filter positively affected the removal of outliers in the data set.

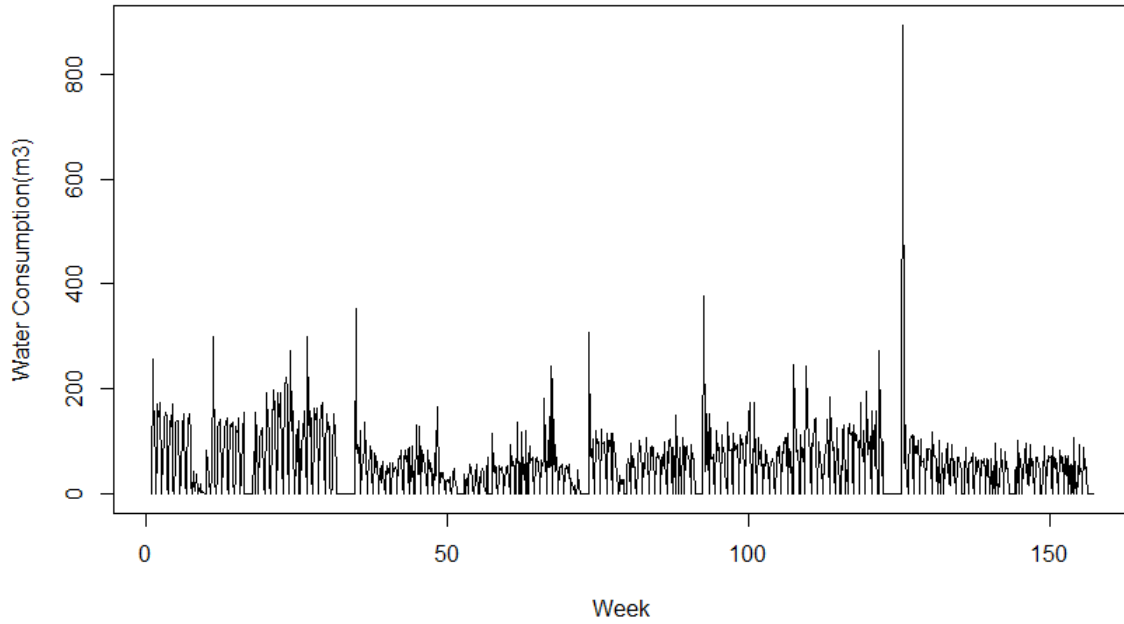


Figure 5.15: Original industrial water dataset

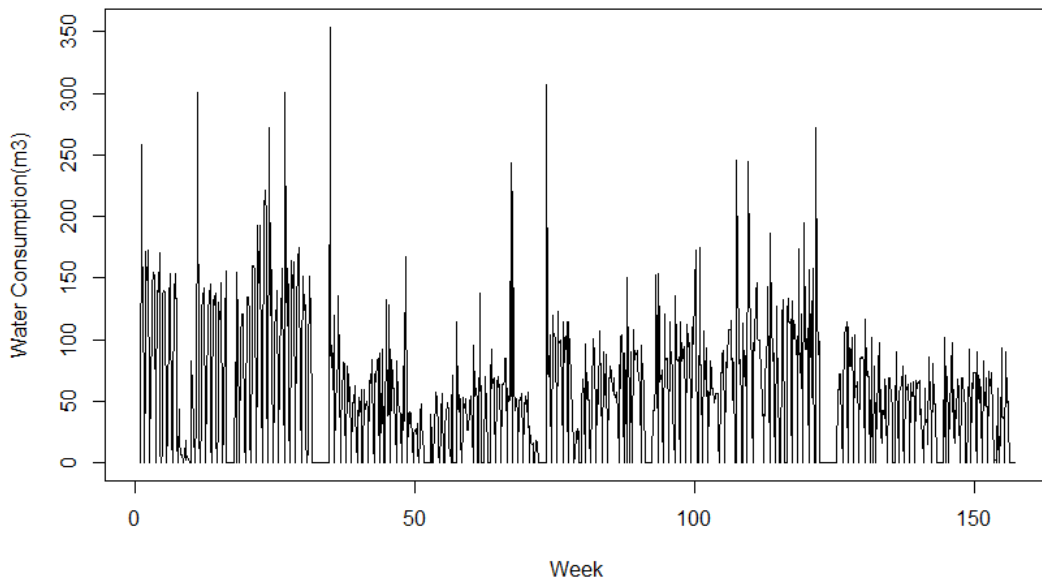


Figure 5.16: Industrial water dataset after applying the Hampel filtering

5.3.3 Seasonality Analysis

In this section, the dataset will be subjected to the seasonality tests.

5.3.3.1 Decomposition of the time series

With the use of the *stl* function, the STL decomposition results were acquired, and each component was plotted, as seen in Figure 5.17.

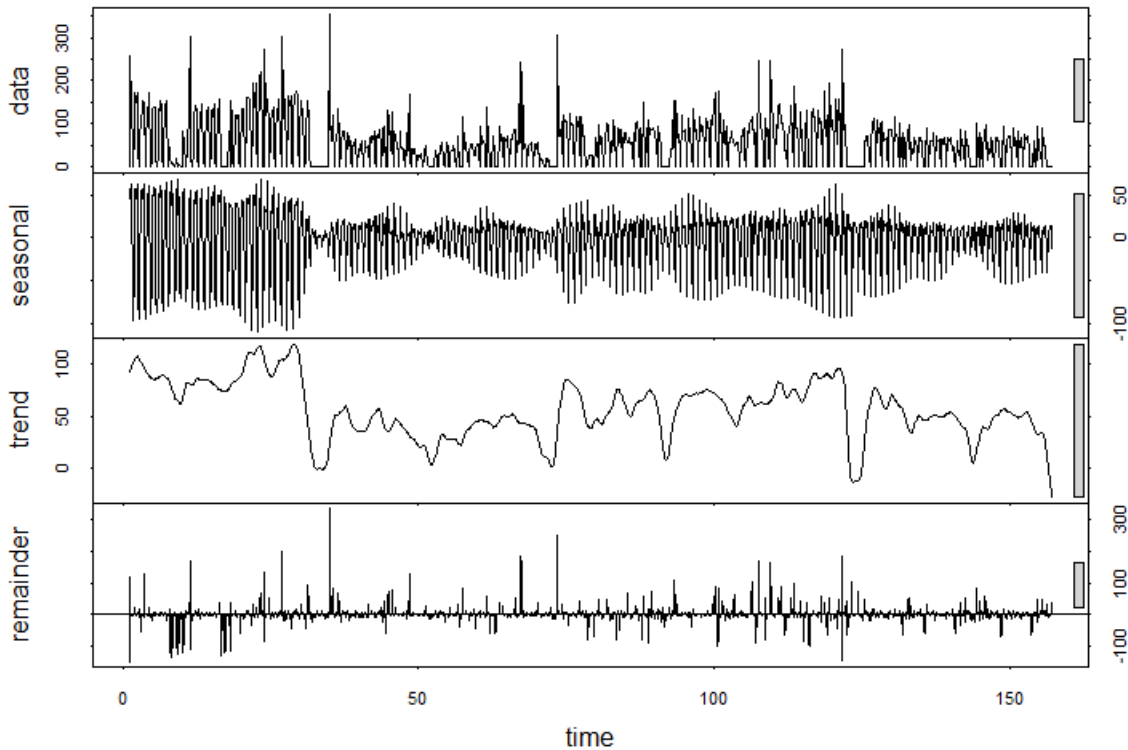


Figure 5.17: STL decomposition of the industrial water consumption time series

Observing the figures and the results related to the STL decomposition, the following comments can be made:

- In the seasonal component of the decomposition, it can be clearly seen a seasonality in the water consumption between the end of the first quarter and the mid of the third quarter of each year - from March to August.
- A decreasing trend can be perceived in the 2019 data and until mid-2021 as well. On the other hand, there was a slight increase in the trend value until the end of the year 2022.

5.3.3.2 ACF Test

The ACF test results were obtained using the *acf.test* function. Figure 5.18 present the final results of the test. It can be concluded that there is a certain correlation at the beginning of each week (Lag 1, 2, 3, 4, ...).

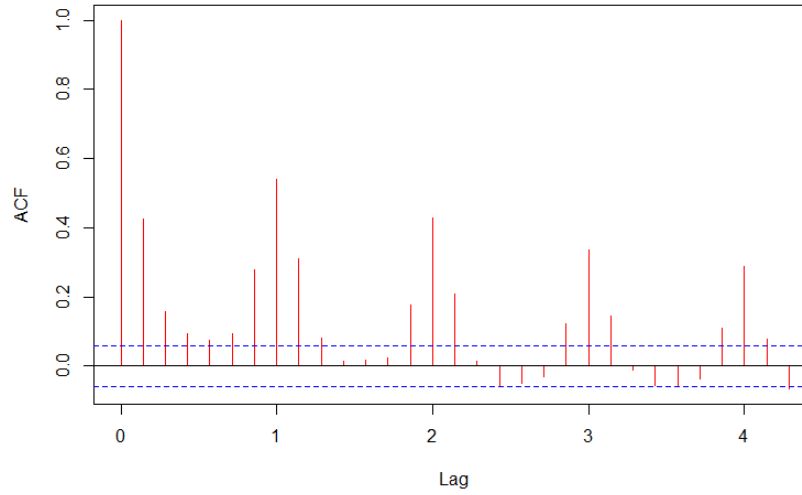
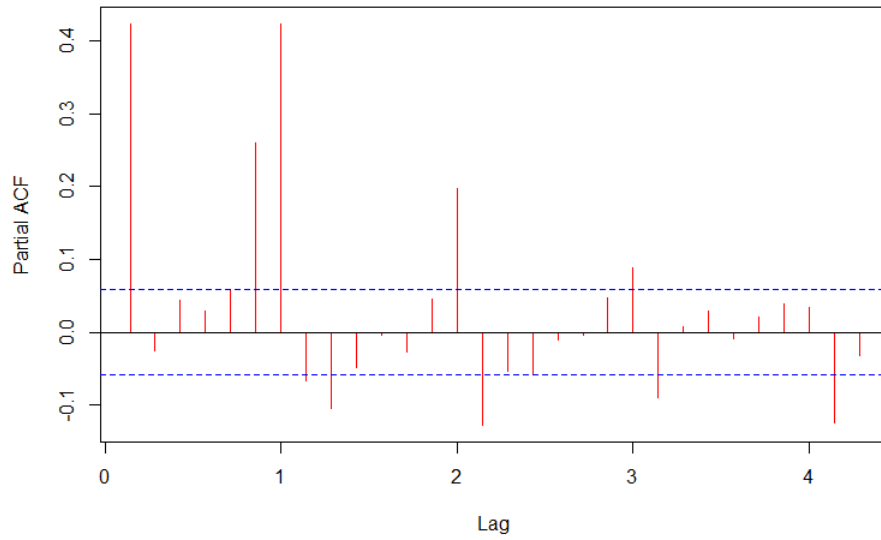


Figure 5.18: ACF Test

5.3.3.3 PACF Test

The PACF test results were obtained using the *pacf.test* function. Figure 5.19 presents the final results of the test. From direct observation, one can state the same that was previously stated for the ACF Test: there is some correlation with regard to the beginning of each week.

**Figure 5.19:** PACF Test

5.3.4 Stationarity Analysis

The corresponding stationarity tests will be performed and analyzed in this section.

5.3.4.1 Augmented Dickey-Fuller Test

The *adf.test* function enables the conduction of the ADF test in R, with 5% significance. Table 5.8 presents the test results.

Table 5.8: ADF test results

Parameters	Result
ADF	-6.8071
Lag Order	10
p-value	0.01

Conclusion: The p-value of the test is below the 5% significance level ($p = 0.05$). Thus, H_0 is rejected, and the series is stationary.

5.3.4.2 Kwiatkowski–Phillips–Schmidt–Shin Test

The *kpss.test* function enables the conduction of the KPSS test in R, with 5% significance. Table 5.9 presents the test results.

Table 5.9: KPSS test results

Parameters	Result
KPSS Test	0.69053
Truncation Lag Parameter	7
p-value	0.01441

Conclusion: The p-value of the test is below the 5% significance level ($p = 0.05$). Thus, H_0 is rejected, and the series is not stationary.

5.3.4.3 Data Differencing

When employing, for example, the ARIMA model, the non-stationarity of a time series could affect its accuracy due to the fact that such a model is only employed with stationary data. From the stationarity analysis, the test results were ambiguous: ADF concludes that the series is, in fact, stationary, while KPSS does not.

The *ndiffs* function was used to ensure which analysis was correct. This function calculates whether or not the given time series requires to be differenced. The output result was 1, meaning that the series, for the ARIMA model, required to be previously first order differenced, $d = 1$. Figure 5.20 presents the differenced time series.

The difference operator of order d can be defined as:

$$\nabla^d x_t = (1 - B)^d x_t \quad (5.8)$$

where, B is the backshift operator - $B^k x_t = x_{t-k}$, for $k \geq 1$.

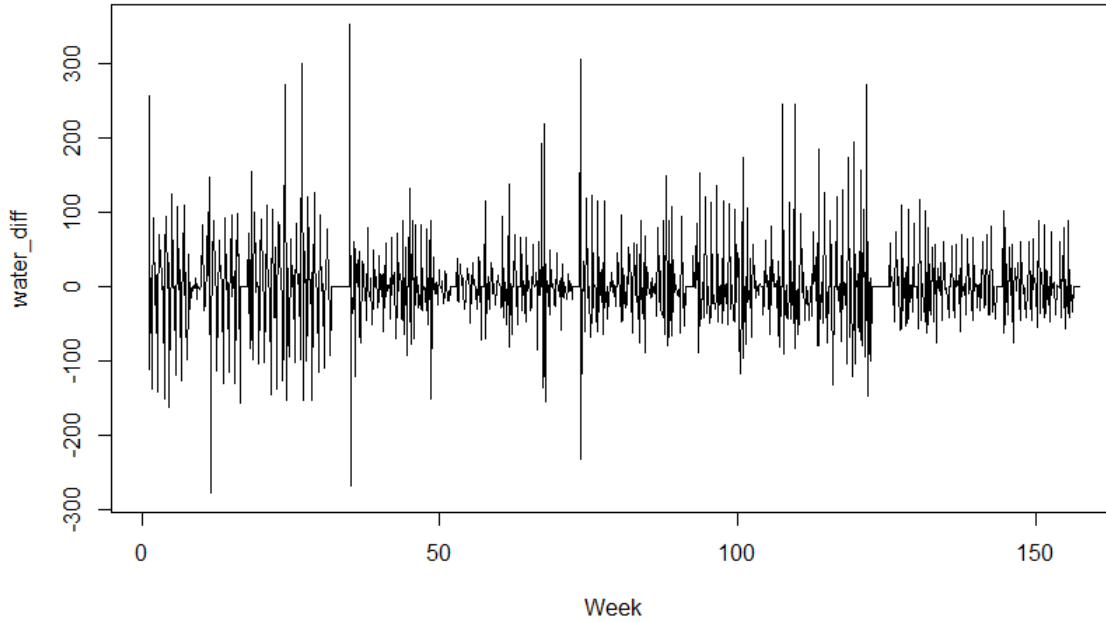


Figure 5.20: Differenced industrial time series

5.3.5 Model Training and Evaluation

After processing the data, the models are trained and tested using the same data. The dataset was divided as follows: 70% was utilized for training and 30% for testing. This split is one of the most well-known in data analysis. Gholamy et al. (2018) explain in their work why the 70/30 or 80/20 divisions are the best to employ in these analyses. SARIMA and SARIMAX, Exponential Smoothing and NNETAR models were used to train and evaluate the models. Each section will present the results as well as further remarks on them.

5.3.5.1 Exogenous variables dataset

To broaden the examination of the various models, a search for a dataset containing the highest temperature, humidity, and precipitation in Ovar was conducted. However, such information could not be found. Instead, weather information from Porto was obtained from the following website: <https://www.visualcrossing.com/weather-history/Porto%2CPortugal>. The data relating to 2019, 2021 and 2022 were collected.

5.3.5.2 Metrics for model evaluation

To determine whether or not the model tests were successful, evaluators who characterize the quality of the model and its projections are required. The following metrics were considered for the model evaluation:

- Mean Error - *ME*.
- Root mean squared error - *RMSE*.

$$RMSE(\%) = \left(\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n} \right)^{0.5} \cdot 100 \quad (5.9)$$

- Mean Average Error or Deviation - *MAE* or *MAD*.

5.3.5.3 SARIMA and SARIMAX models

The SARIMA and SARIMAX models were employed in R with the aid of the *auto.arima* function that automatically calculates the model that best fits the data according to the highest log-likelihood value and lowest Akaike Information Criterion (AIC). These two last parameters are calculated as follows (Taboga, 2021):

$$\hat{\theta} = \operatorname{argmax}[l(\theta; \xi)] \quad (5.10)$$

where:

$$l(\theta; \xi) = -0.5n \cdot \ln(2\pi) - 0.5n \cdot \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

ξ - vector whose entries x_i are draws from a normal distribution.

$$\theta = [\mu, \sigma^2].$$

$$AIC = 2K - 2\ln(L) \quad (5.11)$$

where:

K - number of independent variables used.

L - the log-likelihood estimate.

Annex G contains the R code that was developed for this section. The results, presented in Table 5.10 and 5.11, demonstrate that the best model was the SARIMAX (0,1,2)(2,0,0)[7] with the temperature and humidity as exogenous factors. This last model's training and test graphics are present in Figures 5.21 and 5.22.

Table 5.10: Training and testing values for SARIMA and SARIMAX with single exogenous variable

	SARIMA		SARIMAX (Temperature)		SARIMAX (Humidity)		SARIMAX (Precipitation)	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
ME	-0.357	-24.306	-0.181	-22.235	-0.349	-22.850	-0.364	-24.169
RMSE (%)	44.866	48.354	42.840	47.255	42.817	47.531	42.839	48.332
MAE (%)	28.712	39.136	28.724	38.121	28.745	38.266	28.665	39.156
Model Parameters	(0,1,2)(2,0,0)[7]							

Table 5.11: Training and testing values for SARIMAX with three and two exogenous variables

	SARIMAX (With 3 variables)		SARIMAX (Temperature + Humidity)	
	Training	Testing	Training	Testing
ME	-0.376	-23.217	-0.166	-20.806
RMSE (%)	42.801	47.818	42.815	46.407
MAE (%)	28.711	38.626	28.727	37.120
Model Parameters	(0,1,2)(2,0,0)[7]			

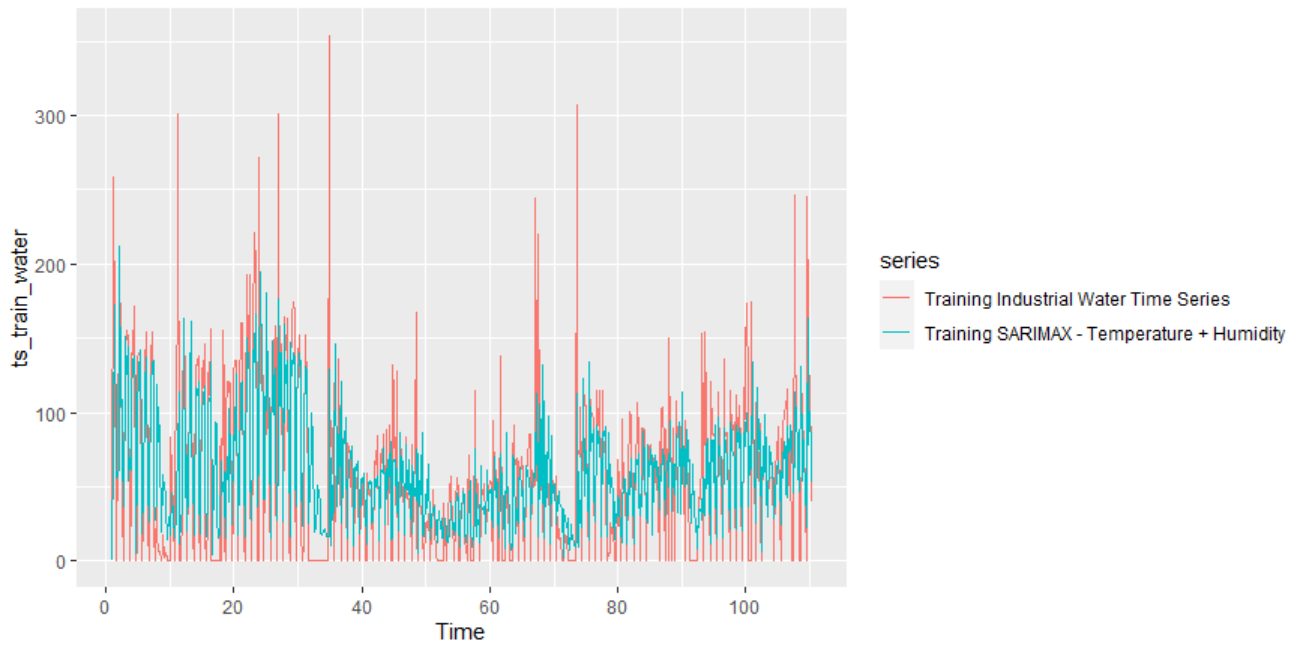


Figure 5.21: Training of the SARIMAX - Temperature + Humidity model

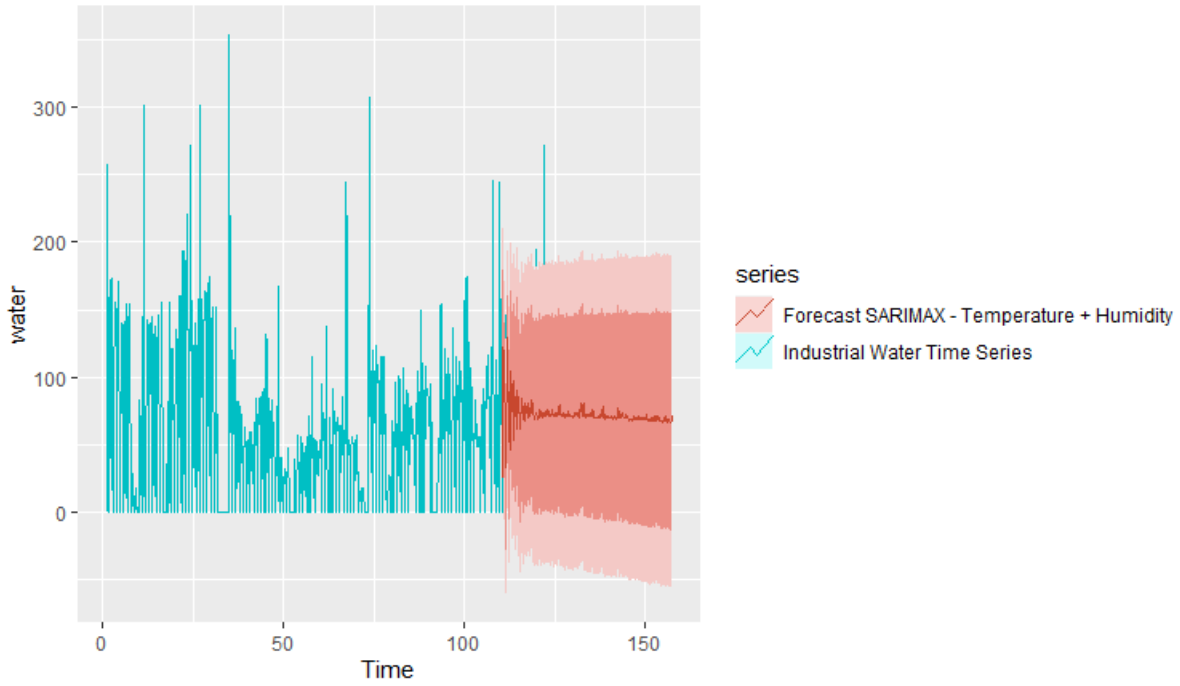


Figure 5.22: Forecast of the SARIMAX - Temperature + Humidity model

5.3.5.4 Exponential Smoothing models

Two exponential smoothing methods were used: the Holt-Winters and ETS, with the aid of the *HoltWinters* and *ets* functions, present in the R *forecast* package. It must be noted that in these models, the exogenous variables were not taken into account. Annex H contains the R code that was developed for this section.

From the results presented in Table 5.12, one can state that the best model is the ETS with the following parameters: $\alpha = 0.2879$ and $\gamma = 1e - 04$. The graphics from testing and training are depicted in Figures 5.23 and 5.24.

Table 5.12: Training and testing values for exponential smoothing models

	Holt - Winters		ETS	
	Training	Testing	Training	Testing
ME	0.163	-83.777	-0.005	-45.002
RMSE (%)	42.476	98.436	42.563	60.158
MAE (%)	28.335	87.314	28.468	50.349
Model Parameters	$\alpha = 0.1743$ $\beta = 2e-04$ $\gamma = 0.1461$		$\alpha = 0.2879$ $\gamma = 1e-04$	

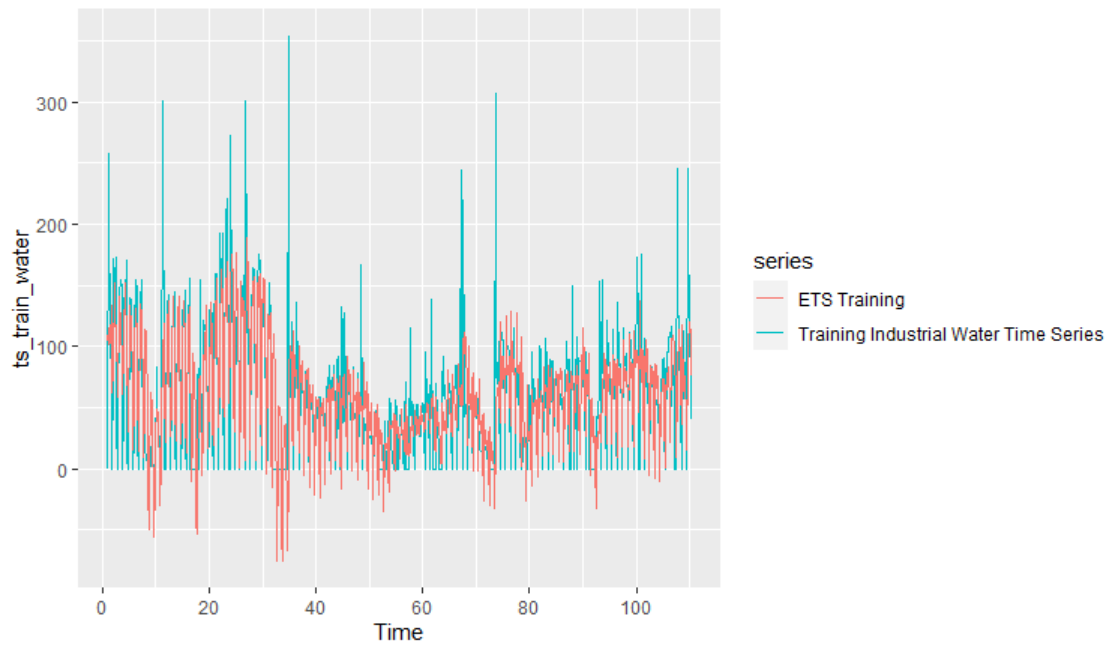


Figure 5.23: Training of the ETS model

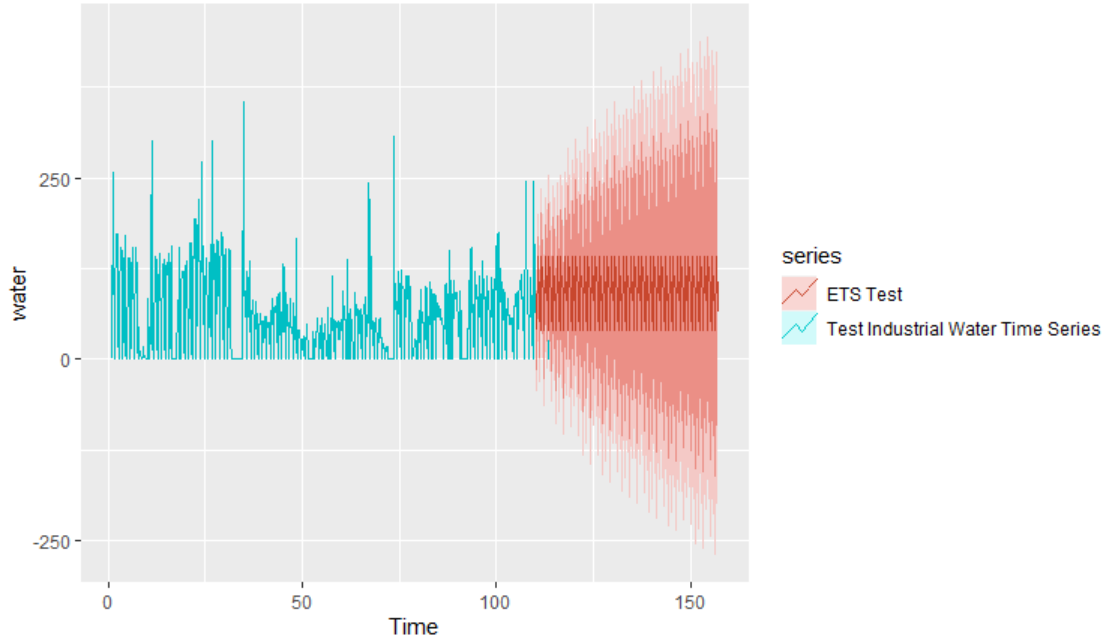


Figure 5.24: Forecast of the ETS model

5.3.5.5 NNETAR models

The NN model was trained and tested using the *nnetar* function, included in the *forecast* package. Annex I contains the R code that was developed for this section. The results, as shown in Figures 5.13 and 5.14, suggest that the model that best fitted the dataset was the NNETAR (28,1,16)[7] model with exogenous variables such as temperature and humidity. Figures 5.25 and 5.26 depicts the training and test graphics.

Table 5.13: Training and testing values for NNETAR and NNETAR with single exogenous variable

	NNETAR		NNETAR (Temperature)		NNETAR (Humidity)		NNETAR (Precipitation)	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
ME	-0.120	-45.837	-0.091	-13.322	-0.062	-32.053	-0.162	-57.023
RMSE (%)	10.695	79.672	9.605	50.404	10.270	74.334	9.546	92.883
MAE (%)	5.605	63.576	4.552	38.401	4.865	57.436	4.692	73.197
Model Parameters	(28,1,15)[7]							

Table 5.14: Training and testing values for NNETAR with three and two variables

	NNETAR (With 3 variables)		NNETAR (Temperature + Humidity)	
	Training	Testing	Training	Testing
ME	-0.120	-15.512	-0.099	5.253
RMSE (%)	7.315	58.466	7.810	47.997
MAE (%)	3.569	42.802	3.681	38.560
Model Parameters	(28,1,16)[7]			

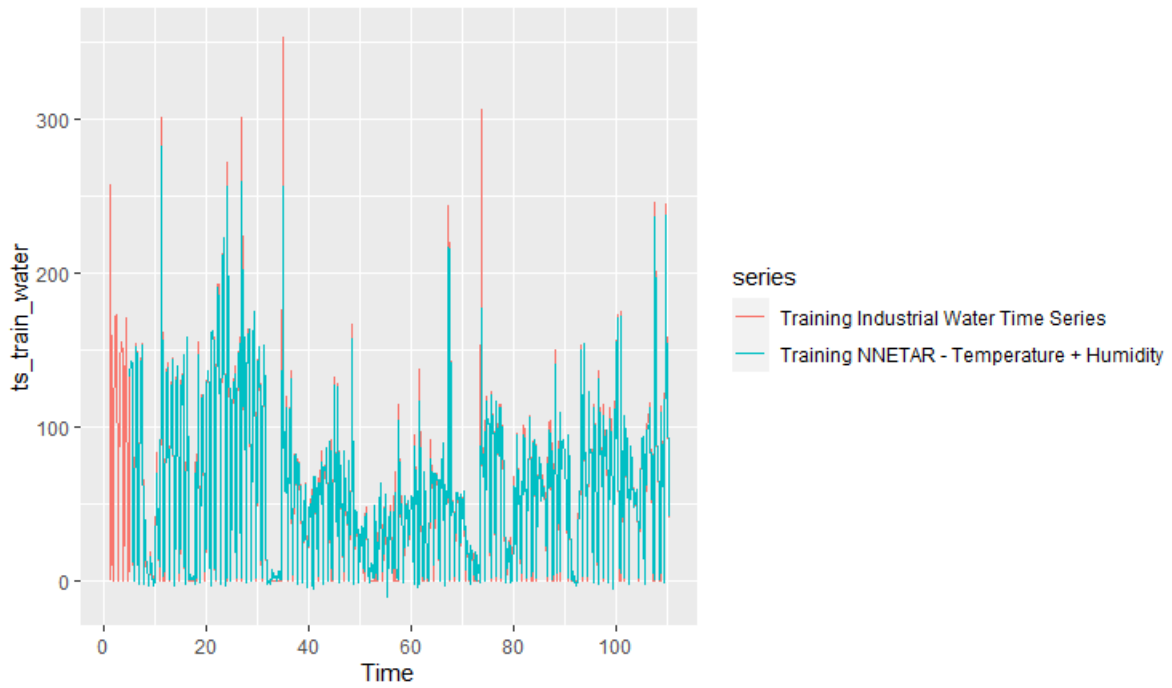


Figure 5.25: Training of the NNETAR - Temperature + Humidity model

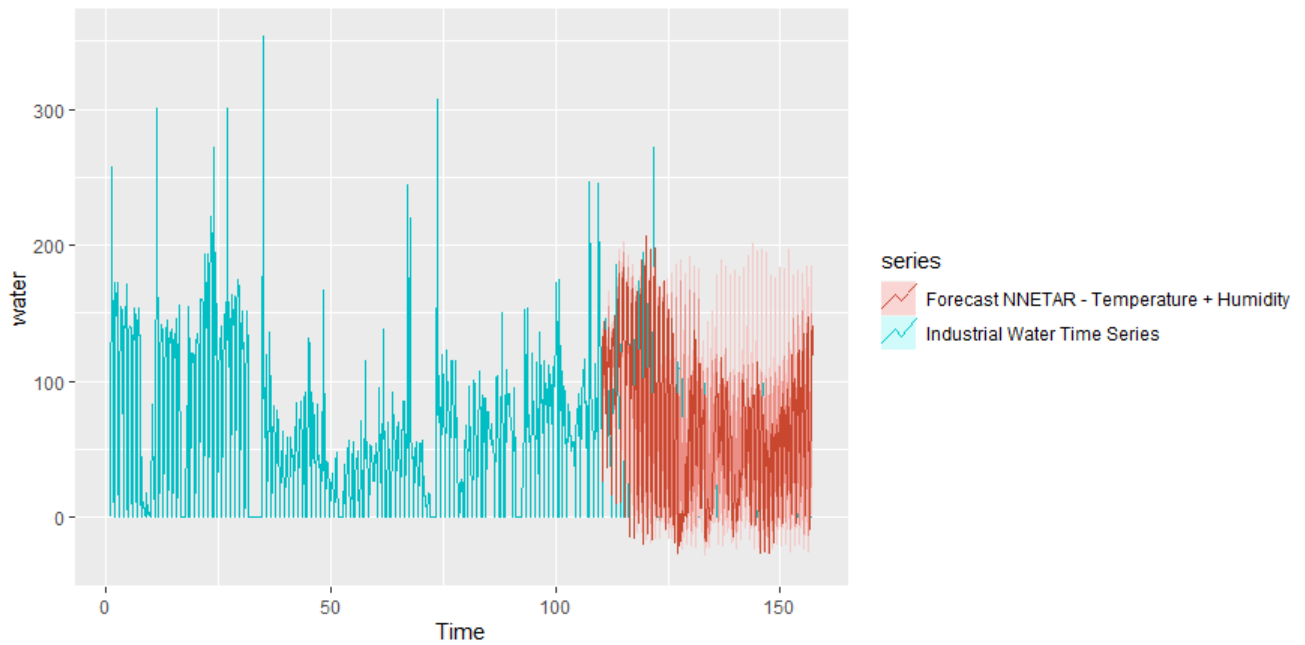


Figure 5.26: Forecasting of the NNETAR - Temperature + Humidity model

5.3.5.6 ELM and MLP models

The ELM and MLP models were both trained and tested, with the aid of the *nnfor* package. However the related test RMSE was above 1000%. As a result, the findings were excluded from this study.

5.3.5.7 Conclusions

From the three subsets of tests that were done, three models rose to prominence:

- SARIMAX (0,1,2)(2,0,0)[7] - Temperature + Humidity.
- ETS ($\alpha = 0.2879$, $\gamma = 1e-04$).
- NNETAR (28,1,16)[7] - Temperature + Humidity.

The ETS fared the worst of all three models, whereas the NNETAR and SARIMAX had similar metrics regarding the testing part. However, the computational simulation time of the NNETAR model was significantly longer than that of the SARIMA: NNETAR took around 8 minutes to forecast, yet the SARIMAX model took only seconds. This reinforces one of the disadvantages of NN models: their longer computing time.

Also, in general, introducing the exogenous variables positively affected the model's evaluation criteria, diminishing the associated errors.

Conclusion: As a result of this analysis, it is possible to infer that:

- SARIMAX (0,1,2)(2,0,0)[7] - Temperature + Humidity is the best model for estimating the factory's industrial water usage if computational simulation time is a critical requirement.
- If detecting seasonality in the model is more essential than the computing time required, the NNETAR (28,1,16) [7] - Temperature + Humidity model is the best for predicting the factory's industrial water consumption.

Digitalization of the Water Supply System

Digitalization of water supply systems is critical in increasing the efficiency and sustainability of water treatment procedures. As technology advances, the water sector is progressing toward full digitization in order to improve system sustainability. Digitalization has the ability to remove uncertainty from the water treatment process by increasing data accessibility and allowing users to monitor and manage water supply systems remotely. As a result, digitization may play an important role in assisting water treatment firms in improving operations without the need for complex processes (ABB Measurement & Analytics, 2023) (doxee, 2023).

This chapter will give a proposal for system digitization, with as *AS IS - TO BE* approach, whose analysis was previously started in Chapter 3.

6.1 Improvement Proposals

In this section, the water supply process is going to be reviewed utilizing an *AS IS - TO BE* approach, commonly used in business process management (BPM). It is usually employed to evaluate how a certain organization is in regard to future process improvement. In this context, it will be applied to evaluate the needed process transformations to achieve full water digitalization. The final document that will aggregate all of this section is the A3 Solving of the digitalization of the water supply system, presented in the Annex J. of this master dissertation. From now on, the analysis will focus on improvements that could be made in each section of the water supply system.

6.1.1 Pumping Station And RIA

In the pumping station, there is a missing water meter at the exit of the industrial water tank. Such a water meter would enable the readings of that tank. Also, to calculate the efficiency of the process related to the water retrieving from the paintshop well, it is crucial the installation of a water meter in the highlighted section presented in Figure 6.1. With the following equation, one would have the capability to calculate the water that is wasted:

$$Waste_{Well} = Reading_{NewMeter} - Reading_{Well} \quad (6.1)$$

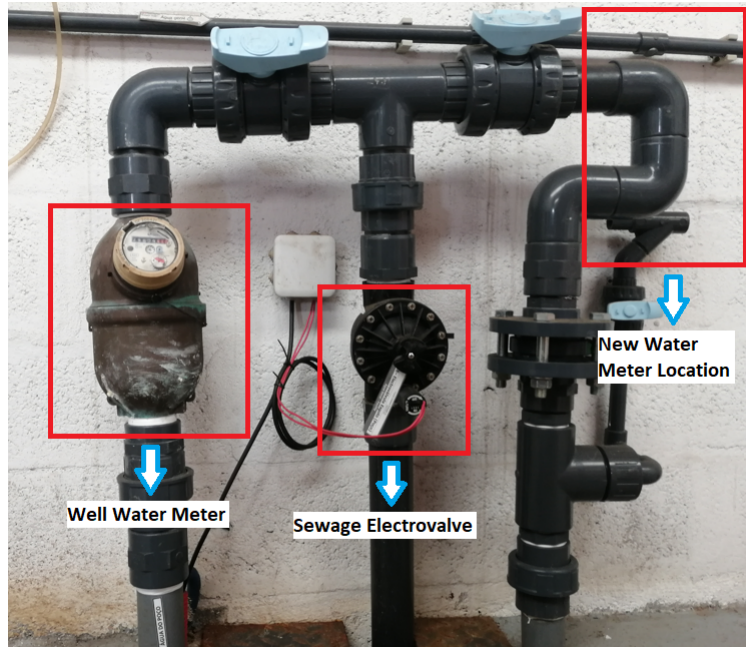


Figure 6.1: New water meter location in the pump station

6.1.2 Factory Entrance And Surroundings

The electrovalve placed next to the factory entrance is wrongly positioned: it does not allow the independent cut-off of water to each plant (TCAP and CaetanoBus). At this moment, the activation of this electrovalve cuts off the water supply to both. The solution could be the placement of another electrovalve, which would allow the independent cut-off of water to each plant.



Figure 6.2: Electrovalve placed next to the factory's entrance

6.1.3 Water Proof Test And *Decapagem*

A water meter is needed at the *Decapagem* tank to retrieve the water readings of it.

6.1.4 Canteen and Men's Bathroom

A meter is needed to record industrial water consumption in the men's bathroom. The three meters placed on the wall next to the canteen access stairs (Canteen F1, Women's Bathroom, Men's Locker Room) are in good condition but require contact with the supplier to confirm possible digitization of the process.

6.1.5 Reverse Osmosis

The Osmosis Pre-Treatment meter, represented in Figure 6.3, is in poor condition and needs replacement.



Figure 6.3: Reverse Osmosis PT meter

6.1.6 Laundry Room

The two water meters are recent (from 2022) and are fully capable of receiving the pulse or telemetry device.

6.1.7 Paint Shop

Cuba ED (Demineralized Water) meter is in poor condition and needs to be replaced with a new one. PT Tank 3 and 4 meters are in good condition, but no information was found about the possible adaptation for digitalization of consumption.

6.2 *TO-DO And TO-BE*

The next step in the analysis is to define what are the main goals to achieve at the end of the digitalization process. Some of these goals were defined in the meeting as follows:

- Monitor the entire water supply process in real time.
- Creation of a PokeYoke system to detect system anomalies - excess consumption and leaks, among others.
- Direct visualization of the process parameters in live dashboards.
- Implementation of predictive analysis methods for water consumption.

- Creation of a database for directly recording water data from each water meter.

Chapter 4 already presented some work developed to improve data visualization and data recording by creating a database for water meter recordings and a dashboard in *PowerBI* for water data visualization and monitoring. The next section will present the investigation regarding communication between the water meters and the database.

6.2.1 Integrated Water Metering

There are two ways to integrate water metering records automatically, using the following systems: Automatic Meter Reading (AMR) or Advanced Metering Infrastructure (AMI).

AMR is a measurement technique used to acquire billing data. AMR converts mechanical dial motions into digital signals, which are transmitted to the utility company via cables, telephone lines, electricity lines, and so on, (Nimbargi et al., 2016). Figure 6.4 shows how the data flows in an AMR system.

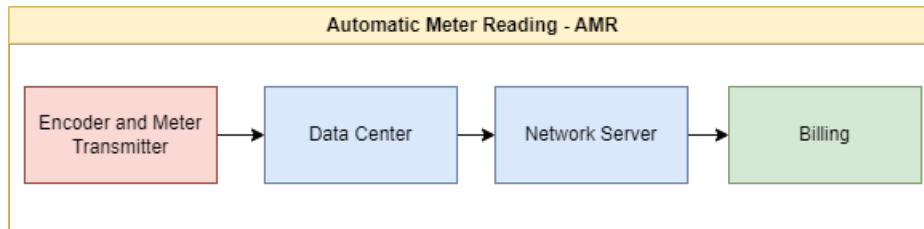


Figure 6.4: AMR data flow chart (Adapted from Nimbargi et al. (2016))

AMI is a more sophisticated version of AMR that comprises a home network system, a communication network connecting meters to a local data concentrator, a database management system, and data integration into current software application platforms. AMI allows two-way communication between the utility provider and the smart meter, allowing for real-time monitoring of energy consumption and more specific information than standard meters. This technology provides customer service benefits such as demand response management, customizable billing cycles, pricing structures, distributed network management, and cost predictions (Nimbargi et al., 2016). Figure 6.5 represents the two-way communication in the AMI data flow.

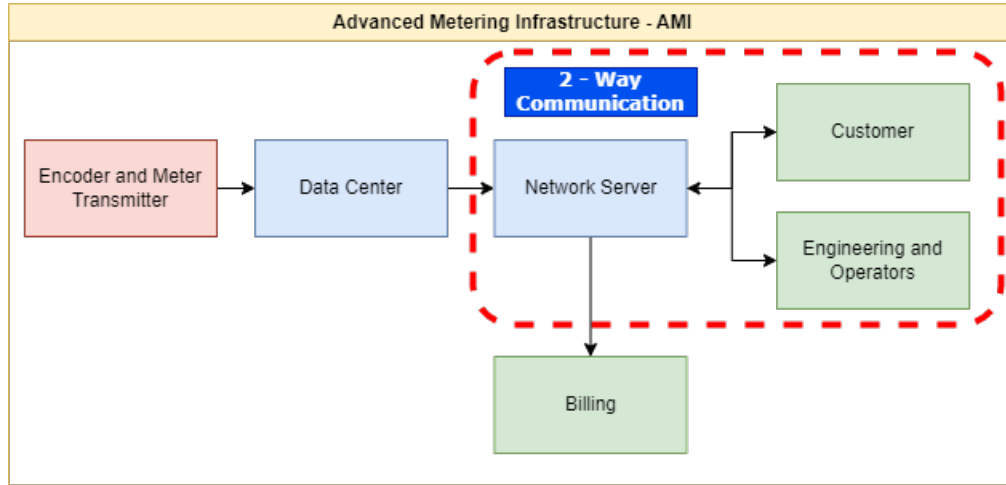


Figure 6.5: AMI data flow chart (Adapted from Nimbargi et al. (2016))

AMI is seen to be superior to AMR since it has more sophisticated features and benefits. AMI, as opposed to AMR, allows for two-way communication between the utility provider and the smart meter, allowing for real-time monitoring of energy use and more specific information than standard meters. This technology provides customer service benefits such as demand response management, customizable billing cycles, pricing structures, distributed network management, and cost predictions. Furthermore, AMI is more reliable and precise than AMR since it transmits data via digital signals rather than decoding mechanical dial motions (Nimbargi et al., 2016).

6.2.2 Low-Power Wide Area Network Related Technologies

The main communication technology associated with AMI is Low-Power Wide Area Network (LPWAN). Because it is a cost-effective and dependable method of connecting smart meters to the utility company's back-end system, LPWAN is utilized in AMI systems. Smart meters are low-power devices that must communicate a small amount of data over vast distances, sometimes in inaccessible areas and can interact with the utility company's network using relatively little power at a cheap cost while yet maintaining reliable connectivity over a large region thanks to LPWAN technology. This makes LPWAN suitable for AMI systems (Chaudhari et al., 2020).

LPWANs can support packet sizes ranging from 10 to 1000 bytes and uplink rates of up to 200 Kbps. The long-distance of an LPWAN varies by technology from 2 km to 1000 km. LPWANs can be subdivided into the following types, shown in Figure 6.6.

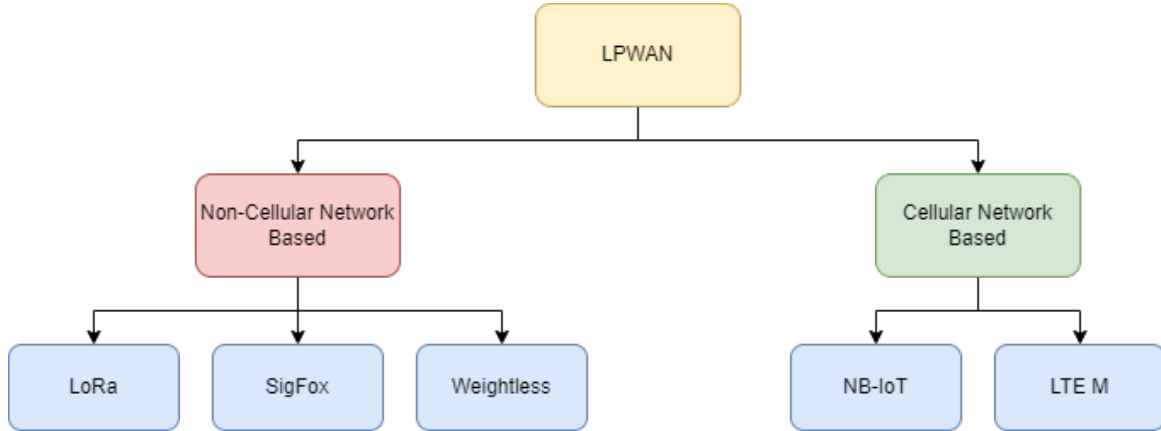


Figure 6.6: Types of LPWANs (Adapted from Shea (2017))

In the research carried out to obtain new smart meters or adapters for digitalizing water consumption, three types of LPWAN stand out: LoRa, SigFox and NB-IoT.

- **LoRa:** Consists of a wireless communication technology that stands for Long Range. It is designed to provide long-range connectivity with low power consumption for IoT devices. LoRa uses spread spectrum modulation to achieve its long-range capability and has advantages in terms of battery lifetime, capacity, and cost (Sinha et al., 2017).
- **NB-IoT:** Narrowband Internet of Things is a wireless communication technology developed for IoT devices. It is part of the 3GPP standard Release 13 and is included in the LTE standard. NB-IoT operates on regulated frequency ranges. Most known for its low power consumption, wide coverage, low cost, high reliability, secure communication and Quality of Service (QoS) support (Sinha et al., 2017).
- **SigFox:** Sigfox is a global supplier of communication services that provides a low-power, wide-area network for the Internet of Things. With a single, standard-based set of APIs, it consistently gathers data from sensors and devices. Sigfox's innovative technology complements standard cellular M2M by delivering low-cost worldwide, universal, ultra-long battery life solutions (SigFox, 2018).

6.2.3 Criteria for LPWANs

A study by Lalle et al. (2019) states that several criteria must be considered when choosing the base technology of a smart water grid system. These are going to be presented, together with a short analysis regarding the three technologies previously mentioned.

- **Energy consumption:** The three technologies allow the use of battery in a prolonged and continuous way, up to a maximum of 10 to 15 years. However, due to its communication

process's peculiarities, NB-IoT tends to consume slightly more than LoRaWAN and SigFox (Lalle et al., 2019).

- **Latency:** NB-IoT promotes the exchange of information with the lowest latency (Lalle et al., 2019).
- **Cost:** NB-IoT has a much higher cost when compared to the others since it operates in licensed bands. LoRaWAN and SigFox use free-licensed bands (Lalle et al., 2019).
- **Scalability:** LoRaWAN and SigFox can support thousands of connections, while NB-IoT only supports up to 50000 (Lalle et al., 2019).
- **Data Range:** Table 6.1 presents the average values for data range for each LPWAN technology. SigFox is the one that covers the most area of communications.

Table 6.1: Data range for each LPWAN technology (Adapted from Lalle et al. (2019))

LPWAN Technology	Urban Range (km)	Rural Range (km)
NB-IoT	1	10
LoRaWAN	5	20
SigFox	10	40

- **Service quality:** Only NB-IoT provides an excellent quality of service, however, at a very high cost. LoRaWAN and SigFox manage to overcome possible interference when issuing information but do not fully guarantee this quality (Lalle et al., 2019).

The spider chart presented in Figure 6.7 resumes all the above parameters for each technology.

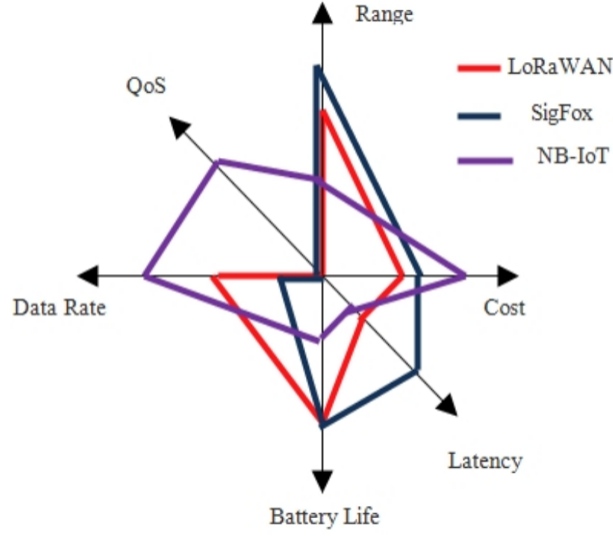


Figure 6.7: Spider chart of the three LPWAN technologies (Lalle et al., 2019)

Another critical aspect of the study is the efficiency of information transmission, which is based on the Packet Error Rate (PER). A packet is a discrete portion of a bigger message. Data is separated into packets when transmitted via computer networks such as the Internet. The receiving computer or device recombines these packets (Lalle et al., 2019). The parameter is calculated as follows:

$$PER = \frac{\sum Packages_{NotReceived}}{Packages_{Total}} \quad (6.2)$$

The authors also concluded that:

- For the same number of devices and gateways in the network, the PER value for NB-IoT is lower than for LoRa and SigFox.
- The error associated with the PER in the three technologies increases with the number of devices on the network.
- The lower the PER linked with each technology, the bigger the number of gateways.

For further clarification, a gateway is a network node that links two networks that employ distinct transmission protocols. Because all data must travel through or connect with the gateway before being routed, gateways serve as a network's entrance and departure point (Lalle et al., 2019).

6.3 Analysis of the potential digitalization of in-use water meters

After the investigation of the technologies related to water digitalization and the Gemba Walks done to review the location and condition of each water meter, the next step was to research the possibility of adapting some of the water meters already in use in the system.

The list of all of the water meters in the factory is presented in the Annex K of this dissertation. The table contains the name of the water meter, its specification, condition and digitalization strategy.

6.3.1 Work partnership with *Go.Charge*

The digitalisation plan will probably have a partnership with the IT company that belongs to Salvador Caetano's group: *Go.Charge*. The company already has a long list of projects applied in TCAP and experience in the digitalization field. Thus, the plan must be carefully detailed with *Go.Charge* in future meetings.

Conclusions And Future Work

This chapter will provide the main conclusions of this master dissertation, as well the proposed future work.

7.1 Conclusions

The work executed on this master's dissertation resulted in the development of a dashboard, a predictive model for the consumption of water, and a strategy for future digitization of the industrial water supply system. The main positive outcomes of each creation are depicted in Figure 7.1.

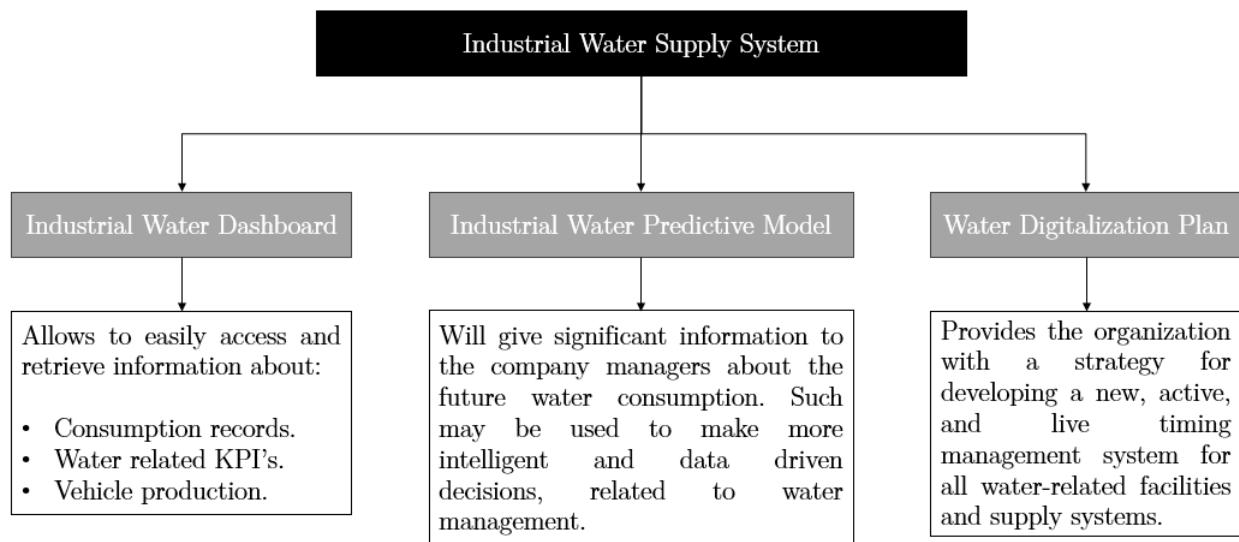


Figure 7.1: Main outcomes from the developed work in the dissertation

7.2 Future Work

Despite the fact that this work will already result in various beneficial improvements, the following future work may give additional benefits in the management and optimization

of the water supply system:

- Integrate an R script in *PowerBI* that will automatically present the forecasts of water consumption using the data that is in the database. The forecast interval is chosen by the user.
- Create a write-back function that enables specific users to input the water meter consumption records in the dashboard manually. Such will be useful while the water meters are not connected to the integrated water metering system.
- Enable the integration of the water meter records that are currently present in the MySQL database in the System Analysis Program Development (SAP) database that the company currently operates.
- Create a dashboard evaluation form in which factory employees are requested to do fundamental activities on the dashboard and provide feedback on the corresponding activity. This form will be utilized as a guideline for future dashboard improvements.
- Refine the industrial water predictive models with the introduction of exogenous variables related to vehicle production. For example, it could be considered both the vehicle production and calendar dummy variables.

Bibliography

- a Rhaffor, K., Azizul, N., Jamian, R., and Azulhisham, J. (2019). Original article the adoption of 5s practice and its impact on safety management performance: A case study in a university environment. 16:9–18.
- ABB Measurement & Analytics (2023). Digitalization in the Water Industry. Retrieved May 28, 2023, from <https://www.azom.com/article.aspx?ArticleID=20323>.
- ACEA (2022). Water used in car production in the EU. Retrieved April 20, 2023, from <https://www.acea.auto/figure/water-used-in-car-production-in-eu/>.
- Adamowski, J., Chan, H., Prasher, S., Ozga-Zielinski, B., and Sliusarieva, A. (2012). Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in montreal, canada. *Water Resources Research*, 48:1528–.
- Alexis, R. D. and Rounds, D. X. (2016). Critical Equipment Identification And Maintenance. Retrieved February 28, 2023, from <https://www.wbdg.org/resources/critical-equipment-identification-and-maintenance>.
- Almetwally, E. and Almongy, H. (2018). Comparison between m-estimation, s-estimation, and mm estimation methods of robust estimation with application and simulation.
- Alwosheel, A., van Cranenburgh, S., and Chorus, C. G. (2018). Is your dataset big enough? sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of choice modelling*, 28:167–182.
- Andreu-Perez, J., Poon, C. C., Merrifield, R. D., Wong, S. T., and Yang, G.-Z. (2015). Big data for health. *IEEE journal of biomedical and health informatics*, 19(4):1193–1208.
- Anele, A. O., Hamam, Y., Abu-Mahfouz, A. M., and Todini, E. (2017). Overview, comparative assessment and recommendations of forecasting models for short-term water demand prediction. *Water*, 9(11):887.
- Arunraj, N. S., Ahrens, D., and Fernandes, M. (2016). Application of sarimax model to forecast daily sales in food retail industry. *International Journal of Operations Research and Information Systems (IJORIS)*, 7(2):1–21.
- Asghari, N. (2012). *Evaluating GQM+ Strategies Framework for Planning Measurement System*. PhD thesis.

- Azuma, R. T. (1997). A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4):355–385.
- Basili, V. R., Caldiera, G., and Rombach, H. D. (1994). The goal question metric approach. *Encyclopedia of software engineering*, pages 528–532.
- Begashaw, G. B. and Yohannes, Y. B. (2020). Review of outlier detection and identifying using robust regression model. *International Journal of Systems Science and Applied Mathematics*, 5(1):4–11.
- Booch, G., Rumbaugh, J., and Jacobson, I. (2005). Unified modeling language user guide, (the 2nd edition).
- Borse, C. (2020). Statistical Methods for Identifying Outliers(Regression analysis Approach) (Part II). Retrieved May 17, 2023, from <https://medium.com/analytics-vidhya/statistical-methods-for-identifying-outliers-regression-analysis-approach-partii-977b035b65a0>.
- Britannica (2023). random walk. Retrieved March 8, 2023, from <https://www.britannica.com/science/random-walk>.
- Bufler, R., Clausnitzer, V., Vestner, R., Werner, U., and Ziemer, C. (2019). Water 4.0—an important element for the germany water industry. germany water partnership. Retrieved March 2, 2023, from https://germanwaterpartnership.de/wp-content/uploads/2019/05/GWP_Brochure_Water_4.0.pdf.
- Carolina Bento (2021). Multilayer Perceptron Explained with a Real-Life Example and Python Code: Sentiment Analysis. Retrieved May 27, 2023, from <https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141>.
- Chang, X., Gao, M., Wang, Y., and Hou, X. (2013). Seasonal autoregressive integrated moving average model for precipitation time series. *Journal of Mathematics and Statistics*, 8:500–505.
- Chaudhari, B. S., Zennaro, M., and Borkar, S. (2020). Lpwan technologies: Emerging application characteristics, requirements, and design considerations. *Future Internet*, 12(3).
- Clary, R. and Wandersee, J. (2010). Fishbone diagrams: Organize reading content with a. *Science Scope*, 33(9):31–37.
- Cleveland, W., Dunn, D., and Terpenning, I. (1978). Seasonal analysis of economic time series: Sabl: A resistant seasonal adjustment procedure with graphical methods for interpretation and diagnosis. *NBER Book Chapters*.
- Daniel H. (2019). Mathematical Representation of a Perceptron Layer (with example in TensorFlow). Retrieved May 27, 2023, from <https://medium.com/>

- @daniel.hellwig.p/mathematical-representation-of-a-perceptron-layer-with-example-in-tensorflow-754a38833b44.
- datacamp (2023). Top Techniques to Handle Missing Values Every Data Scientist Should Know. Retrieved May 17, 2023, from <https://www.datacamp.com/tutorial/techniques-to-handle-missing-data-values>.
- De Pace, F., Manuri, F., and Sanna, A. (2018). Augmented reality in industry 4.0. *American Journal of Computer Science and Information Technology*, 06.
- Deloitte (2014). Industry 4.0: Challenges and solutions for the digital transformation and use of exponential technologies.
- Dickey, D. and Fuller, W. (1979). Distribution of the estimators for autoregressive time series with a unit root. *JASA. Journal of the American Statistical Association*, 74.
- Donkor, E. A., Mazzuchi, T. A., Soyer, R., and Alan Roberson, J. (2014). Urban water demand forecasting: review of methods and models. *Journal of Water Resources Planning and Management*, 140(2):146–159.
- doxee (2023). Digitization in the water industry. Retrieved May 28, 2023, from <https://www.doxee.com/blog/digital-disruption/digitization-in-the-water-industry/>.
- edX (2023). How to Deal with Missing Data . Retrieved May 17, 2023, from <https://www.mastersindatascience.org/learning/how-to-deal-with-missing-data/>.
- Erboz, G. (2017). How To Define Industry 4.0: Main Pillars Of Industry 4.0. pages 761–767.
- Few, S. (2006). *Information dashboard design: The effective visual communication of data*. O'Reilly Media, Inc.
- Gholamy, A., Kreinovich, V., and Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation.
- Guimarães, R. C. and Sarsfield Cabral, J. (2010). *Estatística, 2^a Edição*. Verlag Dashofer.
- Hayes, A. (2022). Heteroscedasticity Definition: Simple Meaning and Types Explained. Retrieved March 10, 2023, from <https://www.investopedia.com/terms/h/heteroskedasticity.asp>.
- Howell, E. (2022a). Partial Autocorrelation for Time Series Analysis. Retrieved May 16, 2023, from <https://towardsdatascience.com/partial-autocorrelation-for-time-series-481a9cfa7526>.
- Howell, E. (2022b). Time Series Decomposition. Retrieved May 15, 2023, from <https://towardsdatascience.com/time-series-decomposition-8f39432f78f9>.
- Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

- Isaiah, D. (2014). Water, water, everywhere in vehicle manufacturing. Retrieved April 20, 2023, from <https://www.automotiveworld.com/articles/water-water-everywhere-vehicle-manufacturing/>.
- Ishwarappa and Anuradha, J. (2015). A brief introduction on big data 5vs characteristics and hadoop technology. *Procedia Computer Science*, 48:319–324. International Conference on Computer, Communication and Convergence (ICCC 2015).
- Isniah, S., Purba, H. H., Debora, F., et al. (2020). Plan do check action (pdca) method: literature review and research issues. *Jurnal Sistem dan Manajemen Industri*, 4(1):72–81.
- Khalid, H., Hashim, S. J., Ahmad, S., Hashim, F., and Chaudary, M. A. (2020). Cybersecurity in industry 4.0 context: Background, issues, and future directions. *The nine pillars of technologies for industry*, 4:263–307.
- Kokoszka, P. and Young, G. (2016). Kpss test for functional time series. *Statistics*, 50(5):957–973.
- Kotu, V. and Deshpande, B. (2018). *Data Science: Concepts And Practice*. Morgan Kaufmann.
- Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, 54(1-3):159–178.
- Lalle, Y., Fourati, L. C., Fourati, M., and Barraca, J. P. (2019). A comparative study of lorawan, sigfox, and nb-iot for smart water grid. In *2019 Global Information Infrastructure and Networking Symposium (GIIS)*, pages 1–6. IEEE.
- Lean Enterprise Institute (2009). Toyota Production System. Retrieved May 23, 2023, from <https://www.lean.org/lexicon-terms/toyota-production-system/>.
- Lean Enterprise Institute (2022). A3 Report. Retrieved May 24, 2023, from <https://www.lean.org/lexicon-terms/a3-report/>.
- Mack, C., Su, Z., and Westreich, D. (2018). Managing missing data in patient registries: addendum to registries for evaluating patient outcomes: a user’s guide.
- Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3):e0194889.
- Malaeb, L. and Ayoub, G. M. (2011). Reverse osmosis technology for water treatment: State of the art review. *Desalination*, 267(1):1–8.
- MasterClass (2022). Stochastic Definition: What Does ‘Stochastic’ Mean? Retrieved March 7, 2023, from <https://www.masterclass.com/articles/stochastic-meaning>.

- MathWorks (2019). hampel - outlier removal using hampel identifier. Retrieved May 28, 2023, from <https://www.mathworks.com/help/signal/ref/hampel.html>.
- McLaughlin, S. (2023). Horizontal and Vertical Integration in Industry 4.0 for Pharmaceutical and Medical Device Manufacturers. Retrieved March 1, 2023, from <https://slcontrols.com/horizontal-and-vertical-integration-in-industry-4-0-for-pharmaceutical-and-medical-device-manufacturers/>.
- Nielsen, H. B. (2005). Non-Stationary Time Series and Unit Root Tests. Retrieved March 9, 2023, from https://web.archive.org/web/20161130133316/http://www.econ.ku.dk/metrics/Econometrics2_05_II/Slides/08_unitroottests_2pp.pdf.
- Nimbargi, S., Mhaisne, S., Nangare, S., and Sinha, M. (2016). Review on ami technology for smart meter. In *2016 IEEE International Conference on Advances in Electronics, Communication and Computer Technology (ICAECCCT)*, pages 21–27.
- Omotayo, A., Telukdarie, A., and Rensburg, N. (2019). Water 4.0: An integrated business model from an industry 4.0 approach. pages 1364–1369.
- Rassoul, G., Al-Alawy, A., and Khudair, W. (2012). Reduction of concentrating poisonous metallic radicals from industrial wastewater by forward and reverse osmosis. *J. Eng.*, 18:784–798.
- Robbins, N. B. (2008). Introduction to cycle plots. *Visual Business Intelligence Newsletter*.
- Roser, C. (2015). Industry 4.0. Retrieved February 28, 2023, from https://commons.wikimedia.org/wiki/File:Industry_4.0.png.
- RTC Lab (2020). Augmented Dickey-Fuller (ADF) Test. Retrieved March 8, 2023, from <https://rtmath.net/assets/docs/finmath/html/93a7b7b9-e3c3-4f19-8a57-49c3938d607d.htm>.
- Rüßmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P., and Harnisch, M. (2015). Industry 4.0: The future of productivity and growth in manufacturing industries. *Boston consulting group*, 9(1):54–89.
- Sachin Date (2019). Understanding Partial Auto-correlation And The PACF. Retrieved May 24, 2023, from <https://timeseriesreasoning.com/contents/partial-auto-correlation/>.
- Salvador Caetano (2021). A nossa história. Retrieved February 28, 2023, from <https://salvadorcaetano.pt/quem-somos/a-nossa-historia/>.
- Sazli, M. (2006). A brief review of feed-forward neural networks. *Communications Faculty Of Science University of Ankara*, 50:11–17.
- Scott Curtis (2022). Understanding Plan-Do-Check-Act (PDCA): Your Complete Guide. Retrieved May 23, 2023, from <https://www.twi-institute.com/plan-do-check-act/>.

- Semmens, J., Bras, B., and Guldberg, T. (2014). Vehicle manufacturing water use and consumption: an analysis based on data in automotive manufacturers' sustainability reports. *The International Journal of Life Cycle Assessment*, 19:246–256.
- Shea, S. (2017). Lpwan (low-power wide area network). Retrieved May 28, 2023, from <https://www.techtarget.com/iotagenda/definition/LPWAN-low-power-wide-area-network>.
- Shokraneh, F., Geoffroy-Gagnon, S., Nezami, M. S., and Liboiron-Ladouceur, O. (2019). A single layer neural network implemented by a 4 times 4 mzi-based optical processor. *IEEE Photonics Journal*, 11(6):1–12.
- SigFox (2018). Sigfox technical overview. Retrieved May 28, 2023, from <https://www.techtarget.com/iotagenda/definition/LPWAN-low-power-wide-area-network>.
- Singh, A. and Ahuja, I. S. (2015). Review of 5s methodology and its contributions towards manufacturing performance. *International journal of process management and benchmarking*, 5(4):408–424.
- Sinha, R. S., Wei, Y., and Hwang, S.-H. (2017). A survey on lpwa technology: Lora and nb-iot. *ICT Express*, 3(1):14–21.
- S.J.M (2021). Advantages of Toyota Production System (TPS) and Lean Manufacturing. Retrieved May 23, 2023, from <https://www.leansixsigmamaster.com/post/advantages-of-toyota-production-system-tps-and-lean-manufacturing>.
- Taboga, M. (2021). Log-likelihood. Retrieved May 28, 2023, from <https://www.statlect.com/glossary/log-likelihood>.
- Tong, W., Li, L., Zhou, X., Hamilton, A., and Zhang, K. (2019). Deep learning pm2.5 concentrations with bidirectional lstm rnn. *Air Quality, Atmosphere Health*, 12:1–13.
- Toyota Europe (2022). Toyota Production System. Retrieved May 23, 2023, from <https://www.toyota-europe.com/about-us/toyota-vision-and-philosophy/toyota-production-system>.
- Toyota Portugal (2008). Fábrica Toyota Ovar. Retrieved February 28, 2023, from <https://www.toyota.pt/mundo-toyota/toyota-em-portugal/fabrica-toyota-ovar>.
- Toyota UK (2013a). Genba – Toyota Production System guide. Retrieved May 23, 2023, from <https://mag.toyota.co.uk/genba-toyota-production-system/>.
- Toyota UK (2013b). Genchi Genbutsu – Toyota Production System guide. Retrieved May 23, 2023, from <https://mag.toyota.co.uk/genchi-genbutsu/>.
- Toyota UK (2013c). Heijunka – Toyota Production System guide. Retrieved May 23, 2023, from <https://mag.toyota.co.uk/heijunka-toyota-production-system/>.

- Toyota UK (2013d). Kanban – Toyota Production System guide. Retrieved May 23, 2023, from <https://mag.toyota.co.uk/kanban-toyota-production-system/>.
- Toyota UK (2013e). Muda, Muri, Mura – Toyota Production System guide. Retrieved May 23, 2023, from <https://mag.toyota.co.uk/muda-muri-mura-toyota-production-system/>.
- Toyota UK (2013f). Nemawashi – Toyota Production System guide. Retrieved May 23, 2023, from <https://mag.toyota.co.uk/nemawashi-toyota-production-system/>.
- Toyota UK (2013g). What is kaizen and how does Toyota use it? Retrieved May 23, 2023, from <https://mag.toyota.co.uk/kaizen-toyota-production-system/>.
- Toyota UK (2016a). Andon – Toyota Production System guide. Retrieved May 23, 2023, from <https://mag.toyota.co.uk/andon-toyota-production-system/>.
- Toyota UK (2016b). Hansei – Toyota Production System guide. Retrieved May 23, 2023, from <https://mag.toyota.co.uk/hansei-toyota-production-system/>.
- Toyota UK (2016c). Konnyaku Stone – Toyota Production System. Retrieved May 23, 2023, from <https://mag.toyota.co.uk/konnyaku-stone/>.
- Toyota UK (2016d). Poka-yoke – Toyota Production System guide. Retrieved May 23, 2023, from <https://mag.toyota.co.uk/poka-yoke/>.
- Uhlemann, T. H.-J., Lehmann, C., and Steinhilper, R. (2017). The digital twin: Realizing the cyber-physical production system for industry 4.0. *Procedia Cirp*, 61:335–340.
- United Nations Educational, Scientific and Cultural Organization (2023). The united nations world water development report 2023 - partnerships and cooperation for water.
- Vaidya, S., Ambad, P., and Bhosle, S. (2018). Industry 4.0 – a glimpse. *Procedia Manufacturing*, 20:233–238. 2nd International Conference on Materials, Manufacturing and Design Engineering (iCMMD2017), 11-12 December 2017, MIT Aurangabad, Maharashtra, INDIA.
- Visual Paradigm (2022). "What is Unified Modeling Language (UML)?". Retrieved April 20, 2023, from <https://www.visual-paradigm.com/guide/uml-unified-modeling-language/what-is-uml/>.
- Vogel-Heuser, B. and Hess, D. (2016). Guest editorial industry 4.0–prerequisites and visions. *IEEE Transactions on Automation Science and Engineering*, 13(2):411–413.
- Wang, J., Lu, S., Wang, S.-H., and Zhang, Y.-D. (2022). A review on extreme learning machine. *Multimedia Tools and Applications*, 81(29):41611–41660.
- West, M. (1997). Time series decomposition. *Biometrika*, 84(2):489–494.
- Witkowski, K. (2017). Internet of things, big data, industry 4.0 – innovative solutions in

logistics and supply chains management. *Procedia Engineering*, 182:763–769. 7th International Conference on Engineering, Project, and Production Management.

Zakaria, A., Howard, N. K., and Nkansah, B. K. (2014). On the detection of influential outliers in linear regression analysis.

Zhang, S., Zhang, C., and Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, 17(5-6):375–381.

Gantt Diagram

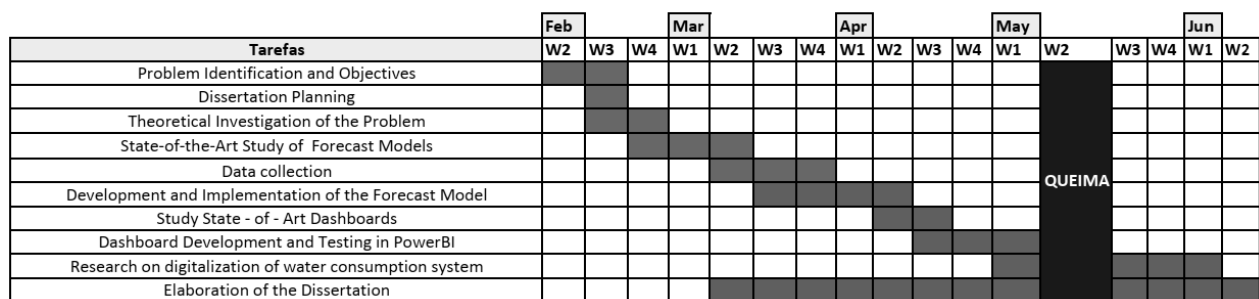


Figure A.1: Gantt diagram

A3 Problem Solving Template

[illegible]

Figure B.1: A3 problem solving company template

Industrial Water And RIA Tank Supply System Technical Drawing

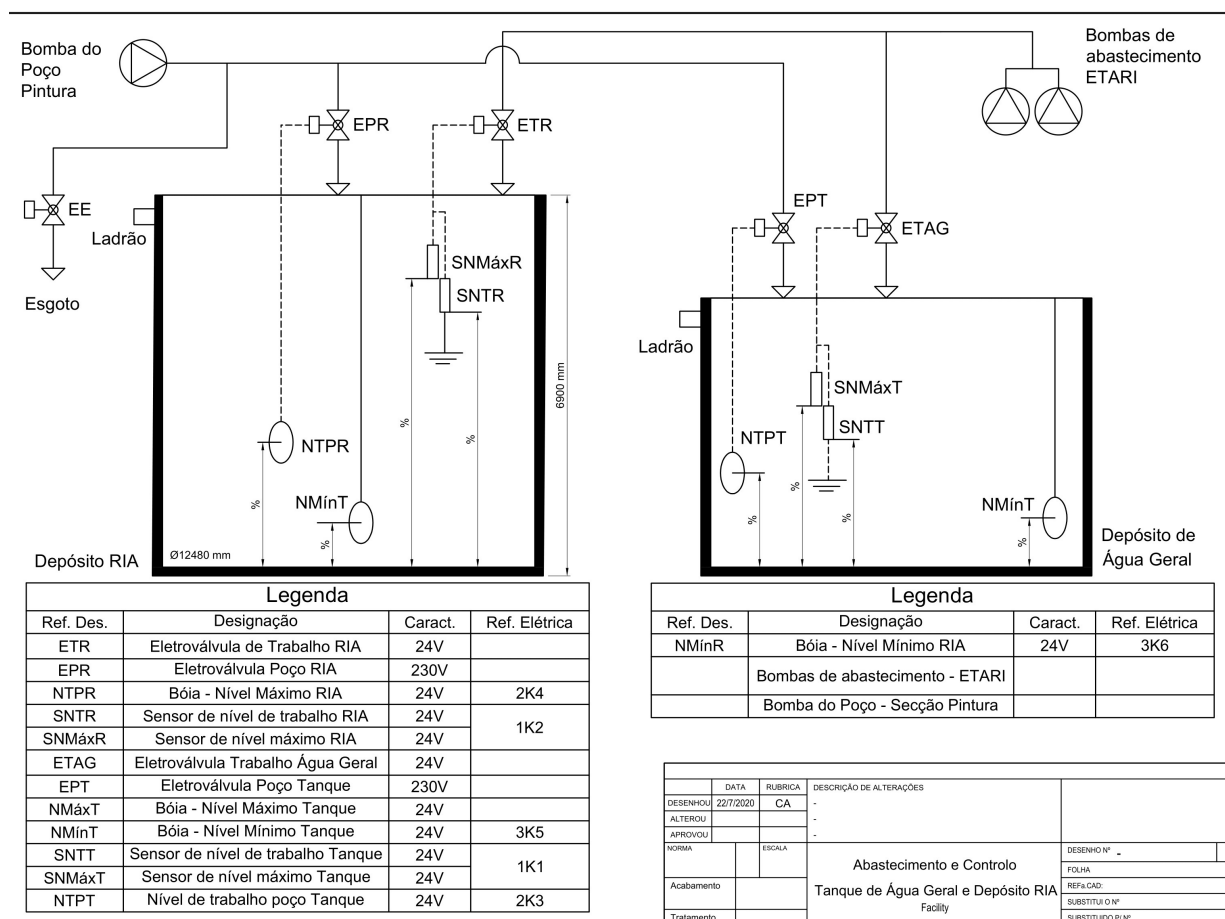


Figure C.1: Industrial Water And RIA Tank Supply System Technical Drawing

First UML iteration

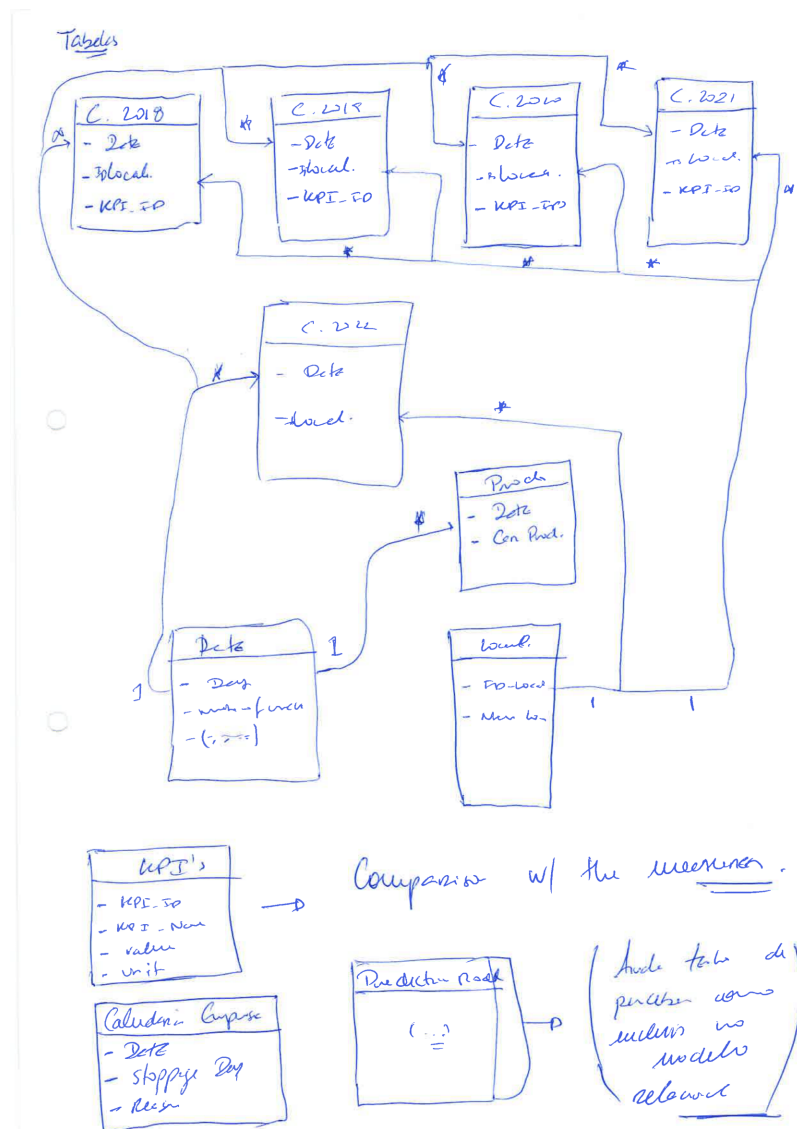


Figure D.1: First handwriting UML iteration for the industrial water dashboard

Water Consumption Analysis - Robust Regression Code

```
library(olsrr)
library(robustbase)
library(sjPlot)

IndustrialWater = Dados_agua_pintura$Agua_Industrial_F1
DemineralizedWater = Dados_agua_pintura$Agua_Desmineralizada
ProdCarros = Dados_agua_pintura$Prod_Carros

m1 <- lm(IndustrialWater ~ ProdCarros)
m2 <- lm(DemineralizedWater ~ ProdCarros)

# Residuals diagnostics

outliers_m1 <- ols_plot_resid_lev(m1)
outliers_m2 <- ols_plot_resid_lev(m2)

# Conducting the robust regression

lm1 <- lm(IndustrialWater ~ ProdCarros)
summary(lm1)
rm1 <- lmrob(IndustrialWater ~ ProdCarros)
summary(rm1)

lm2 <- lm(DemineralizedWater ~ ProdCarros)
summary(lm2)
rm2 <- lmrob(DemineralizedWater ~ ProdCarros)
summary(rm2)

#plotting residuals
```

```
plot_residuals(m1)
plot_residuals(m2)

# plotting both models

plot_model(m1, type = "pred", show.data = T)
plot_model(rm1, type = "pred", show.data = T)

plot_model(m2, type = "pred", show.data = T)
plot_model(rm2, type = "pred", show.data = T)

#compare R2

tab_model(m1)
tab_model(rm1)
tab_model(m2)
tab_model(rm2)
```


Water Consumption In The Paintshop - Robust Regression Code

```
library(olsrr)
library(robustbase)
library(sjPlot)

lmrob.control(setting = "KS2014", max.it = 500)

IndustrialWater_PT = consumos_pintura$Consumo_PT_Agua_Geral
DemineralizedWater_PT = consumos_pintura$Consumo_PT_Agua_Desm
DemineralizedWater_ED = consumos_pintura$`Consumo_ED_Agua Desm`
ProdCarros = consumos_pintura$Prod_Carros

m1_PT <- lm(IndustrialWater_PT ~ ProdCarros)
m2_PT <- lm(DemineralizedWater_PT ~ ProdCarros)
m2_ED <- lm(DemineralizedWater_ED ~ ProdCarros)

# Residuals diagnostics

outliers_m1_PT <- ols_plot_resid_lev(m1_PT)
outliers_m2_PT <- ols_plot_resid_lev(m2_PT)
outliers_m2_ED <- ols_plot_resid_lev(m2_ED)
# Conducting the robust regression

lm1_PT <- lm(IndustrialWater_PT ~ ProdCarros)
summary(lm1_PT)
rm1_PT <- lmrob(IndustrialWater_PT ~ ProdCarros)
summary(rm1_PT)

lm2_PT <- lm(DemineralizedWater_PT ~ ProdCarros)
summary(lm2_PT)
```

```
rm2_PT <- lmrob(DemineralizedWater_PT ~ ProdCarros)
summary(rm2_PT)

lm2_ED <- lm(DemineralizedWater_ED ~ ProdCarros)
summary(lm2_ED)
rm2_ED <- lmrob(DemineralizedWater_ED ~ ProdCarros)
summary(rm2_ED)

#plotting residuals

plot_residuals(m1_PT)
plot_residuals(m2_PT)
plot_residuals(m2_ED)

# plotting the models

plot_model(m1_PT, type = "pred", show.data = T)
plot_model(rm1_PT, type = "pred", show.data = T)
plot_model(m2_PT, type = "pred", show.data = T)
plot_model(rm2_PT, type = "pred", show.data = T)
plot_model(m2_ED, type = "pred", show.data = T)
plot_model(rm2_ED, type = "pred", show.data = T)

#compare R2

tab_model(m1_PT)
tab_model(rm1_PT)
tab_model(m2_PT)
tab_model(rm2_PT)
tab_model(m2_ED)
tab_model(rm2_ED)
```

Time Series Decomposition, Statistical Tests And SARIMAX Forecastings Code

```
library(pracma)
library(forecast)
library(seastests)

#time series creation
temperature <- ts(temperaturas_2019_2021_2022$tempmax, frequency = 7)
humidity <- ts(temperaturas_2019_2021_2022$humidity, frequency = 7)
precipitation <- ts(temperaturas_2019_2021_2022$precip, frequency = 7)

#hampel filtering for outlier detection
water_hampel <- hampel(consumos_industrial_2019_2021_2022$Total,k=1.4826)
water <- ts(water_hampel$y, frequency = 7)

#data plotting
plot(temperature)
plot(humidity)
plot(precipitation)
plot(water)

# Get the total number of observations in the time series
total_obs_water <- length(water)

# Calculate the number of observations for the training and testing sets
train_obs_water <- round(0.7 * total_obs_water)
test_obs_water <- total_obs_water - train_obs_water

# Split the time series into training and testing sets
train_water <- water[1:train_obs_water]
test_water <- water[(train_obs_water + 1):total_obs_water]
```

```
ts_train_water <- ts(train_water, frequency = 7)
ts_test_water <- ts(test_water, frequency = 7)

plot(ts_train_water)
plot(ts_test_water)

#Splitting the temperature dataset

total_obs_temperature <- length(temperature)

train_obs_temperature <- round(0.7 * total_obs_water)
test_obs_temperature <- total_obs_temperature - train_obs_water

train_temperature <- temperature[1:train_obs_temperature]
test_temperature <- temperature[(train_obs_temperature + 1):total_obs_temperature]

ts_train_temperature <- ts(train_temperature, frequency = 7)
ts_test_temperature <- ts(test_temperature, frequency = 7)

plot(ts_train_temperature)
plot(ts_test_temperature)

#Splitting the humidity dataset

total_obs_humidity <- length(humidity)

train_obs_humidity <- round(0.7 * total_obs_humidity)
test_obs_humidity <- total_obs_humidity - train_obs_humidity

train_humidity <- humidity[1:train_obs_humidity]
test_humidity <- humidity[(train_obs_humidity + 1):total_obs_humidity]

ts_train_humidity <- ts(train_humidity, frequency = 7)
ts_test_humidity <- ts(test_humidity, frequency = 7)

plot(ts_train_humidity)
plot(ts_test_humidity)

#Splitting the precipitation dataset
```

```

total_obs_precipitation <- length(precipitation)

train_obs_precipitation <- round(0.7 * total_obs_precipitation)
test_obs_precipitation <- total_obs_precipitation - train_obs_precipitation

train_precipitation <- precipitation[1:train_obs_precipitation]
test_precipitation <- precipitation[(train_obs_precipitation + 1):total_obs_precipitation]

ts_train_precipitation <- ts(train_precipitation, frequency = 7)
ts_test_precipitation <- ts(test_precipitation, frequency = 7)

plot(ts_train_precipitation)
plot(ts_test_precipitation)

#test for time series seasonality

stl_decomposition <- stl(water, s.window = 7, robust = TRUE)
stl_decomposition
plot(stl_decomposition)

seasonal <- isSeasonal(water, freq = 7)
seasonal

#test for time series differencing
diff_level <- ndiffs(water)
diff_level

water_diff <- diff(water)
plot(water_diff, main="Industrial Water - Differenced", xlab="Week")

stl_decomposition_diff <- stl(water_diff, s.window = 7, robust = TRUE)
stl_decomposition_diff
plot(stl_decomposition_diff)

#acf and pacf

acf_water <- acf(water)
plot(acf_water, main = "ACF Test - Industrial Water", col = "red")

```

```
pacf_water <- pacf(water)
plot(pacf_water, main = "PACF Test - Industrial Water", col="red")

#ARIMA with no exogenous variable

fit_water_arima <- auto.arima(ts_train_water, seasonal = TRUE)
summary(fit_water_arima)
checkresiduals(fit_water_arima)

autoplot(ts_train_water, series="Industrial Water Time Series") +
  autolayer(fitted(fit_water_arima),
    series="Training ARIMA ")

forecast_water_arima <- forecast(fit_water_arima, h=test_obs_water)
ts_test_water_arima <- ts(forecast_water_arima$mean, frequency=7)

autoplot(water, series="Industrial Water Time Series") +
  autolayer(forecast_water_arima,
    series="Forecast ARIMA")

accuracy(ts_test_water_arima, ts_test_water)

#ARIMA with max temperature as exogenous variable

fit_temperature_arima <- auto.arima(ts_train_water, xreg = ts_train_temperature,
seasonal = TRUE)
summary(fit_temperature_arima)
checkresiduals(fit_temperature_arima)

autoplot(ts_train_water, series="Training Industrial Water Time Series") +
  autolayer(fitted(fit_temperature_arima),
    series="Training ARIMAX - Temperature")

forecast_temperature_arima <- forecast(fit_temperature_arima, xreg = ts_test_temperature,
h=test_obs_water)
ts_test_temperature_arima <- ts(forecast_temperature_arima$mean, frequency=7)

autoplot(water, series="Industrial Water Time Series") +
  autolayer(forecast_water_arima,
    series="Forecast ARIMAX - Temperature")
```

```

accuracy(ts_test_temperature_arima, ts_test_water)

#ARIMA with humidity as exogenous variable

fit_humidity_arima <- auto.arima(ts_train_water, xreg = ts_train_humidity,
seasonal = TRUE)
summary(fit_humidity_arima)
checkresiduals(fit_humidity_arima)

autoplot(ts_train_water, series="Test Industrial Water Time Series") +
  autolayer(fitted(forecast_humidity_arima),
            series="Forecast ARIMAX - Humidity")

forecast_humidity_arima <- forecast(fit_humidity_arima, xreg = ts_test_humidity,
h=test_obs_water)
ts_test_humidity_arima <- ts(forecast_humidity_arima$mean, frequency=7)

autoplot(water, series="Industrial Water Time Series") +
  autolayer(forecast_humidity_arima,
            series="Forecast ARIMAX - Humidity")

accuracy(ts_test_humidity_arima, ts_test_water)

#ARIMA with precipitation as exogenous variable

fit_precipitation_arima <- auto.arima(ts_train_water, xreg = ts_train_precipitation,
seasonal = TRUE)
summary(fit_precipitation_arima)
checkresiduals(fit_precipitation_arima)

autoplot(ts_train_water, series="Training Industrial Water Time Series") +
  autolayer(fitted(fit_precipitation_arima),
            series="Training ARIMAX - Precipitation")

forecast_precipitation_arima <- forecast(fit_precipitation_arima,
xreg = ts_test_precipitation, h=test_obs_water)
ts_test_precipitation_arima <- ts(forecast_precipitation_arima$mean, frequency=7)

```

```

autoplot(water, series="Industrial Water Time Series") +
  autolayer(forecast_precipitation_arima,
    series="Forecast ARIMAX - Precipitation")

accuracy(ts_test_precipitation_arima, ts_test_water)

#ARIMA with 3 exogenous variables

exogenous_data_train <- cbind("Temperature" = ts_train_temperature,
  "Humidity" = ts_train_humidity, "Precipitation" = ts_train_precipitation)
exogenous_data_test <- cbind("Temperature" = ts_test_temperature,
  "Humidity" = ts_test_humidity, "Precipitation" = ts_test_precipitation)

fit_exogenous_arima <- auto.arima(ts_train_water, xreg = exogenous_data_train,
  seasonal = TRUE)
summary(fit_exogenous_arima)
checkresiduals(fit_exogenous_arima)

autoplot(ts_train_water, series="Training Industrial Water Time Series") +
  autolayer(fitted(fit_exogenous_arima),
    series="Training ARIMAX - Precipitation")

forecast_exogenous_arima <- forecast(fit_exogenous_arima, xreg = exogenous_data_test,
  h=test_obs_water)
ts_test_exogenous_arima <- ts(forecast_exogenous_arima$mean, frequency=7)

autoplot(water, series="Industrial Water Time Series") +
  autolayer(forecast_exogenous_arima,
    series="Forecast ARIMAX - 3 variables")

accuracy(ts_test_exogenous_arima, ts_test_water)

#ARIMA with temperature and humidity as exogenous variables

exogenous_data_train_2 <- cbind("Temperature" = ts_train_temperature,
  "Humidity" = ts_train_humidity)
exogenous_data_test_2 <- cbind("Temperature" = ts_test_temperature,
  "Humidity" = ts_test_humidity)

fit_exogenous_arima_2 <- auto.arima(ts_train_water, xreg = exogenous_data_train_2,

```

```
seasonal = TRUE)
summary(fit_exogenous_arima_2)
checkresiduals(fit_exogenous_arima_2)

autoplot(ts_train_water, series="Training Industrial Water Time Series") +
  autolayer(fitted(fit_exogenous_arima_2),
    series="Training SARIMAX - Exogenous 2 ")

forecast_exogenous_arima_2 <- forecast(fit_exogenous_arima_2,
xreg = exogenous_data_test_2, h=test_obs_water)
ts_test_exogenous_arima_2 <- ts(forecast_exogenous_arima_2$mean, frequency=7)

autoplot(water, series="Industrial Water Time Series") +
  autolayer(forecast_exogenous_arima_2,
    series="Forecast SARIMAX - Exogenous 2")

accuracy(ts_test_exogenous_arima_2, ts_test_water)
```

Exponential Smoothing Forecastings Code

```
library(pracma)
library(forecast)
library(seastests)
library(stats)

#hampel filtering for outlier detection
water_hampel <- hampel(consumos_industrial_2019_2021_2022$Total,k=1.4826)
water <- ts(water_hampel$y, frequency = 7)

#data plotting
plot(water)

# Get the total number of observations in the time series
total_obs_water <- length(water)

# Calculate the number of observations for the training and testing sets
train_obs_water <- round(0.7 * total_obs_water)
test_obs_water <- total_obs_water - train_obs_water

# Split the time series into training and testing sets
train_water <- water[1:train_obs_water]
test_water <- water[(train_obs_water + 1):total_obs_water]

ts_train_water <- ts(train_water, frequency = 7)
ts_test_water <- ts(test_water, frequency = 7)

plot(ts_train_water)
plot(ts_test_water)

#Holt-Winters Training
fit_water_hw <- HoltWinters(ts_train_water)
```

```
summary(fit_water_hw)
checkresiduals(fit_water_hw)

autoplot(ts_train_water, series="Training Industrial Water Time Series") +
  autolayer(fit_water_hw$fitted,
            series="Holt-Winters Training")

#Holt Winters Testing

forecast_water_hw <- forecast(fit_water_hw, h=test_obs_water)
summary(forecast_water_hw)
checkresiduals(forecast_water_hw)

test_forecast_water_hw <- forecast_water_hw$mean
ts_test_forecast_water_hw <- ts(test_forecast_water_hw, frequency=7)

autoplot(water, series="Industrial Water Time Series") +
  autolayer(forecast_water_hw,
            series="Holt-Winters Forecast")

accuracy(ts_test_forecast_water_hw, ts_test_water)
accuracy(forecast_water_hw)

#ETS training

fit_water_ets <- ets(ts_train_water)
summary(fit_water_ets)
checkresiduals(fit_water_ets)

autoplot(ts_train_water, series="Training Industrial Water Time Series") +
  autolayer(fitted(fit_water_ets),
            series="ETS Training")

#ETS testing

forecast_water_ets <- forecast(fit_water_ets, h=test_obs_water)

test_forecast_water_ets <- forecast_water_ets$mean
ts_test_forecast_water_ets <- ts(test_forecast_water_ets, frequency=7)

autoplot(water, series="Test Industrial Water Time Series") +
```

```
    autolayer(forecast_water_ets,  
              series="ETS Test")  
  
accuracy(ts_test_forecast_water_ets,ts_test_water)
```

NNETAR (Neural Networks) Forecasting Code

```
library(pracma)
library(forecast)
library(seastests)

#time series creation
temperature <- ts(temperaturas_2019_2021_2022$tempmax, frequency = 7)
humidity <- ts(temperaturas_2019_2021_2022$humidity, frequency = 7)
precipitation <- ts(temperaturas_2019_2021_2022$precip, frequency = 7)

#hampel filtering for outlier detection
water_hampel <- hampel(consumos_industrial_2019_2021_2022$Total,k=1.4826)
water <- ts(water_hampel$y, frequency = 7)

#data plotting
plot(temperature)
plot(humidity)
plot(precipitation)
plot(water)

# Get the total number of observations in the time series
total_obs_water <- length(water)

# Calculate the number of observations for the training and testing sets
train_obs_water <- round(0.7 * total_obs_water)
test_obs_water <- total_obs_water - train_obs_water

# Split the time series into training and testing sets
train_water <- water[1:train_obs_water]
test_water <- water[(train_obs_water + 1):total_obs_water]

ts_train_water <- ts(train_water, frequency = 7)
```

```
ts_test_water <- ts(test_water, frequency = 7)

plot(ts_train_water)
plot(ts_test_water)

#Splitting the temperature dataset

total_obs_temperature <- length(temperature)

train_obs_temperature <- round(0.7 * total_obs_water)
test_obs_temperature <- total_obs_temperature - train_obs_water

train_temperature <- temperature[1:train_obs_temperature]
test_temperature <- temperature[(train_obs_temperature + 1):total_obs_temperature]

ts_train_temperature <- ts(train_temperature, frequency = 7)
ts_test_temperature <- ts(test_temperature, frequency = 7)

plot(ts_train_temperature)
plot(ts_test_temperature)

#Splitting the humidity dataset

total_obs_humidity <- length(humidity)

train_obs_humidity <- round(0.7 * total_obs_humidity)
test_obs_humidity <- total_obs_humidity - train_obs_humidity

train_humidity <- humidity[1:train_obs_humidity]
test_humidity <- humidity[(train_obs_humidity + 1):total_obs_humidity]

ts_train_humidity <- ts(train_humidity, frequency = 7)
ts_test_humidity <- ts(test_humidity, frequency = 7)

plot(ts_train_humidity)
plot(ts_test_humidity)

#Splitting the precipitation dataset

total_obs_precipitation <- length(precipitation)
```

```

train_obs_precipitation <- round(0.7 * total_obs_precipitation)
test_obs_precipitation <- total_obs_precipitation - train_obs_precipitation

train_precipitation <- precipitation[1:train_obs_precipitation]
test_precipitation <- precipitation[(train_obs_precipitation + 1):total_obs_precipitation]

ts_train_precipitation <- ts(train_precipitation, frequency = 7)
ts_test_precipitation <- ts(test_precipitation, frequency = 7)

plot(ts_train_precipitation)
plot(ts_test_precipitation)

#nnetar with no exogenous variable

fit_water_nnetar <- nnetar(ts_train_water)
summary(fit_water_nnetar)
checkresiduals(fit_water_nnetar)
accuracy(fit_water_nnetar)

autoplot(ts_train_water, series="Industrial Water Time Series") +
  autolayer(fitted(fit_water_nnetar),
    series="Training NNETAR ")

forecast_water_nnetar <- forecast(fit_water_nnetar, h=test_obs_water)
ts_test_water_nnetar <- ts(forecast_water_nnetar$mean, frequency=7)

autoplot(water, series="Industrial Water Time Series") +
  autolayer(forecast_water_nnetar$mean,
    series="Forecast NNETAR")

accuracy(ts_test_water_nnetar, ts_test_water)

#nnetar with max temperature as exogenous variable

fit_temperature_nnetar <- nnetar(ts_train_water, xreg=ts_train_temperature)
summary(fit_temperature_nnetar)
checkresiduals(fit_temperature_nnetar)
accuracy(fit_temperature_nnetar)

```

```

autoplot(ts_train_water, series="Training Industrial Water Time Series") +
  autolayer(fitted(fit_temperature_nnetar),
    series="Training NNETAR - Temperature ")

forecast_temperature_nnetar <- forecast(fit_temperature_nnetar, h=test_obs_water,
xreg=ts_test_temperature, PI =TRUE)
ts_test_temperature_nnetar <- ts(forecast_temperature_nnetar$mean, frequency=7)

autoplot(water, series="Industrial Water Time Series") +
  autolayer(forecast_temperature_nnetar,
    series="Forecast NNETAR - Temperature")

accuracy(ts_test_temperature_nnetar, ts_test_water)

#nnetar with humidity as exogenous variable

fit_humidity_nnetar <- nnetar(ts_train_water, xreg=ts_train_humidity)
summary(fit_humidity_nnetar)
checkresiduals(fit_humidity_nnetar)
accuracy(fit_humidity_nnetar)

autoplot(ts_train_water, series="Training Industrial Water Time Series") +
  autolayer(fitted(fit_humidity_nnetar),
    series="Training NNETAR - Humidity ")

forecast_humidity_nnetar <- forecast(fit_humidity_nnetar, h=test_obs_water,
xreg=ts_test_humidity, PI = TRUE)
ts_test_humidity_nnetar <- ts(forecast_humidity_nnetar$mean, frequency=7)

autoplot(water, series="Industrial Water Time Series") +
  autolayer(forecast_humidity_nnetar, series= "Forecast NNETAR")

accuracy(ts_test_humidity_nnetar, ts_test_water)

#nnetar with precipitation as exogenous variable

fit_precipitation_nnetar <- nnetar(ts_train_water, xreg=ts_train_precipitation)
summary(fit_precipitation_nnetar)

```

```

checkresiduals(fit_precipitation_nnetar)
accuracy(fit_precipitation_nnetar)

autoplot(ts_train_water, series="Training Industrial Water Time Series") +
  autolayer(fitted(fit_precipitation_nnetar),
    series="Training NNETAR - Precipitation")

forecast_precipitation_nnetar <- forecast(fit_precipitation_nnetar,
h=test_obs_water, xreg=ts_test_precipitation, PI = TRUE)
ts_test_precipitation_nnetar <- ts(forecast_precipitation_nnetar$mean, frequency=7)

autoplot(water, series="Industrial Water Time Series") +
  autolayer(forecast_precipitation_nnetar, series="Forecast NNETAR")

accuracy(ts_test_precipitation_nnetar, ts_test_water)

#nnetar with the 3 exogenous factors

exogenous_data_train <- data.frame(temperature = ts_train_temperature,
  humidity = ts_train_humidity,
  precipitation = ts_train_precipitation)

exogenous_data_test <- data.frame(temperature = ts_test_temperature,
  humidity = ts_test_humidity,
  precipitation = ts_test_precipitation)

fit_exogenous_nnetar <- nnetar(ts_train_water, xreg=exogenous_data_train)
summary(fit_exogenous_nnetar)
checkresiduals(fit_exogenous_nnetar)
accuracy(fit_exogenous_nnetar)

autoplot(ts_train_water, series="Training Industrial Water Time Series") +
  autolayer(fitted(fit_exogenous_nnetar),
    series="Training NNETAR - Exogenous")

forecast_exogenous_nnetar <- forecast(fit_exogenous_nnetar, h=test_obs_water,
xreg=exogenous_data_test, PI = TRUE)
ts_test_exogenous_nnetar <- ts(forecast_exogenous_nnetar$mean, frequency=7)

```

```
autoplot(water, series="Industrial Water Time Series") +
  autolayer(forecast_exogenous_nnetar, series="Forecast NNETAR - Exogenous")

accuracy(ts_test_exogenous_nnetar, ts_test_water)

#nnetar with the 2 exogenous factors - temperature and humidity

exogenous_data_train_2 <- data.frame(temperature = ts_train_temperature,
                                     humidity = ts_train_humidity
                                   )

exogenous_data_test_2 <- data.frame(temperature = ts_test_temperature,
                                   humidity = ts_test_humidity
                                   )

fit_exogenous_2_nnetar <- nnetar(ts_train_water, xreg=exogenous_data_train_2)
summary(fit_exogenous_2_nnetar)
checkresiduals(fit_exogenous_2_nnetar)
accuracy(fit_exogenous_2_nnetar)

autoplot(ts_train_water, series="Training Industrial Water Time Series") +
  autolayer(fitted(fit_exogenous_2_nnetar),
            series="Training NNETAR - Exogenous 2")

forecast_exogenous_2_nnetar <- forecast(fit_exogenous_2_nnetar, h=test_obs_water,
xreg=exogenous_data_test_2, PI = TRUE)
ts_test_exogenous_2_nnetar <- ts(forecast_exogenous_2_nnetar$mean, frequency=7)

autoplot(water, series="Industrial Water Time Series") +
  autolayer(forecast_exogenous_2_nnetar, series="Forecast NNETAR - Exogenous 2")

accuracy(ts_test_exogenous_2_nnetar, ts_test_water)
```

A3 Solving of the water digitalization process

Toyota Caetano Portugal, S.A.

DIGITALIZAÇÃO DA REDE DE ÁGUA GERAL

Ambito		Tipo de Redução			Dept:	No Projecto:	Data:	Ateração:
Mão-de-Obra	Gastos	Eficiência	Potencial	Impacto Económico	Manutenção	PRCJ	06-06-2023	00
X	X	X						PT

1. Background / Caracterização:

Current Situation / Situação Atual:

O sistema de abastecimento de água industrial, bem como todo o seu circuito de tubagem, é um dos pontos mais importantes na TCAPI. No entanto, existem algumas situações cuja gestão é anómala:

- Existem erros de contagem de consumos de água, aquando da sua introdução na base de dados em Excel;
- Método de introdução dos registos é arcaico - método manual;
- Não existe uma análise cuidada e crítica aos consumos registados;
- Inexistência de métodos de análise de consumo preditivos;
- Falta de manutenção das tubagens do sistema de abastecimento de água;
- Inexistência de sistema PokeYoke para aviso de possíveis fugas no sistema de abastecimento de água;
- Inexistência de dashboards para visualização de consumos a tempo real.

Para uma empresa do nível da TCAPI espera-se uma gestão dos consumos de água bastante rigorosa, tendo em conta as métricas previstas para o Toyota Environmental Challenge 2050 e as auditorias externas. Dá-se justificar um novo processo de gestão e otimização de consumo de água, baseado nos princípios da digitalização e Indústria 4.0.

2. Action / Acções:

2.1 Project Detail / Detalhes Projeto

A proposta de digitalização tem como base os princípios fundamentais da digitalização de processos, baseada em alguns pilares que sustentam a Indústria 4.0: IoT, cloud, cibersegurança, Big Data & Analytics e integração com sistemas. Os principais passos a seguir no projeto são os seguintes:

- Avaliação geral do processo de abastecimento de água industrial - estado dos equipamentos e avaliação dos KPIs atuais;
- Levantamento dos contadores de água já existentes para adaptação no processo de digitalização;
- Estudo de layout do sistema de abastecimento: layout dos locais mais críticos de instalação; avaliação de localização dos contadores e eletroválvulas existentes, com base em indicações da ERSAR; avaliação das tecnologias de contadores existentes, no caso de necessidade de compra de novos contadores;
- Estudo das tecnologias passíveis e serem utilizadas no sistema de comunicação entre contadores e base de dados (LoRa, Sigfox, NB-IoT etc.);
- Avaliação de performance da base de dados da água geral atual e apresentação de novas estratégias para a nova base de dados;
- Criação de sistema PokeYoke para aviso de possíveis anomalias na rede de água geral: fugas, consumos excessivos num determinado setor, etc;
- Criação de dashboards de visualização de processo e de KPIs com redação de resultados em tempo real;
- Criação do plano de manutenção dos equipamentos da rede.

Todo este processo terá colaboração com a Go Charge, que conta com a experiência necessária para avaliar as tecnologias e metodologias que poderão ser utilizadas neste projeto.

2.2 Custos / Custos

Neste momento, ainda não é possível obter a redução de custos estimada com a implementação deste projeto, devido a ainda não terem sido pedidos cotizações. No entanto:

1. Considerando 2 pessoas a fazer registo de consumos, num total de 2h por dia e a receber 7,56h, durante 46 semanas de trabalho: $46 \times 2 \times 7,56 / 11 = 788,27$ horas passadas de redução;
2. Considerando uma taxa de recomendação numa tubagem passado 24h do seu começo, em que o caudal da tubagem é 3m3/h com um custo associado de 1,14€/m3,04 de captação/distribuição + 0,19€ de tratamento: $788,27 \times 720h \text{ de água desperdiçada} = \text{Considerando KPI de 5m3 desperdiçado, são 14,4 m3 de água que poderiam ter sido produzidos com essa água}$ ou seja, um dia de produção em geral. Em termos de custo ambiental é **insustentável pelo standard do Desafio Toyota R2050**.

2.3 Identificação e avaliação de riscos / Risk Identification and assessment

Possíveis riscos / Possible risks	Definição de estratégia para eliminação ou redução dos riscos identificados / Definition of strategy for elimination or reduction of identified risks
Segurança	Risco associado à cibersegurança.
	Definição de plano estratégico de cibersegurança com a Go Charge.
Ambiente	Risco associado à contaminação pluvial.
	Definição de plano contra risco de contaminação pluvial com o departamento de ambiente.
Processo	Nada a apontar.
Produto	Nada a apontar.
Outros	Nada a apontar.

2.4 Vantagens & Desvantagens/ Advantage & Disadvantage

VANTAGENS / ADVANTAGES:	DESADVANTAGES / DISADVANTAGES:
- Monitorização do processo em tempo real.	- Resistência à mudança.
- Fácil obtenção dos dados de todo o processo.	- Necessidade de sub-contratação de equipas para intervenção técnica.
- Criação de relatórios de análise de forma automática, com indicação de KPIs.	- Aumento de custo relacionado com investimento inicial.
- Poupança relacionada com os pilares financeiros e ambiente.	- Necessidade de planos de formação de equipas e de melhoria de contínuo.
- Implementação de manutenção preditiva, preventiva e curativa.	

3. Evaluation / Avaliação

3.1 Schedule / Plano

Actividade / Activity	% Concl.												Resp.	Status	Remarks		
	Q1 2023	Q2 2023	Q3 2023	Q4 2023	Q1 2024	Q2 2024	Q3 2024	Q4 2024	Q1 2025	Q2 2025	Q3 2025	Q4 2025					
Avaliação geral da rede de água geral																	
Levantamento e avaliação do estado dos contadores																	On Going
Estudo da tecnologia de comunicação entre contadores																	On Going
Estudo da tecnologia existente dos novos contadores																	On Going
Estudo de layout da rede de água geral																	TBD
Início da implementação das medidas após estudos																	TBD
Criação do plano de manutenção dos equipamentos																	TBD
Testes e avaliação ao novo sistema																	TBD
Fim da implementação do novo sistema																	TBD

Planned / Planado Realizado / Realized

TBD - To Be Done

3.2 Follow up / Seguimento

Reunions every week to evaluate knowledge gap and project implementation status.

Create forms for each collaborator to report on their problems, doubts and proposals for improvement.

Analysis of Cost - Benefit after introduction.

4. General comment

Nada a apontar.

5. Communicate

Made: André Domingues Approved: Date: 06/06/2023 Cost Control

Impresso aprovado DOT - entra em vigor em 06-02-2019

Figure J.1: A3 Solving of the water digitalization process

Water meter specifications and digitalization strategy

Contadores de Água – Estratégia de Digitalização

Identificação	Especificações	Estratégia de Digitalização
Industrial F1 	SENSUS MeiStream Plus 65 $Q3 = 40\text{m}^3/\text{h}$; $Q_{\text{max}} = 60\text{m}^3/\text{h}$	Adaptação com sensores de impulso da SENSUS: • HRI-Lei • OD 01 • OD 03
Poço 	SENSUS 620 Volumetric Meter Dry Dial – DN40 $Q3 = 16\text{m}^3/\text{h}$; $Q_{\text{max}} = 20\text{m}^3/\text{h}$	Adaptação com sensor de impulso da SENSUS HRI-A + HRI -B (Data Unit). Ou Adaptação com radio module (telemetria) CompactRF ou PulseRF-A3.

Figure K.1: Water meter specifications and digitalization strategy - page 1

Tanque RIA 	Sem informação.	Sem adaptação possível. Solução passa pela compra de um novo contador.
Osmose Para Produção 	JANZ JV400 - DN 40 $Q_3 = 16\text{m}^3/\text{h}$; $Q_{\text{max}} = 20\text{m}^3/\text{h}$	Adaptação com sensor de impulsos JANZ JI (Sensor Impulsos) ou similar. <u>Ou</u> Adaptação com sistema de telemetria LPWA MYWATER ou similar.
Osmose Para ED e Restante 	JANZ MST 15050 – DN 50 $Q_3 = 15\text{m}^3/\text{h}$; $Q_{\text{max}} = 30\text{m}^3/\text{h}$	Não foi encontrada qualquer informação sobre o contador permitir a colocação de um dispositivo de medição (de notar que a tampa do contador parecer indicar que de facto poderá haver um dispositivo). Contactar fornecedor e, em última instância, proceder à compra de um novo contador.

Figure K.2: Water meter specifications and digitalization strategy - page 2

Osmose Pré-Tratamento 	AquaMetro ???? DN 40 ou 50 (por confirmar) Q3 = 40m3/h	Contador em mau estado de conservação, sem possibilidade de adaptação para digitalização de consumo. Solução passa pela compra de um novo contador.
Industrial Para Osmose 	G2 AFM-R100H – DN 15 ou 20 (por confirmar) Q3 = 3,5 m3/h	Já se encontra adaptado com um sensor de impulsos.
Torres de Refrigeração 	Sem informação.	Contador em mau estado de conservação, sem possibilidade de adaptação para digitalização de consumo. Solução passa pela compra de um novo contador.

Figure K.3: Water meter specifications and digitalization strategy - page 3

Cantina F1 Lavatórios/WC Norte Balneários H 	(3x) JANZ MSV 2525 – DN 20 Q3 = 2,5m³/h; Qmax = 5m³/h	Não foi encontrada qualquer informação sobre o contador permitir a colocação de um dispositivo de medição. Contactar fornecedor e, em última instância, proceder à compra de um novo contador.
Potável F2/Parque 	(2x) JANZ MST 15050 – DN 50 Q3 = 25m³/h; Qmax = 31,5m³/h	Adaptação com sensor de impulsos JANZ JI (Sensor Impulsos) ou similar. <u>Ou</u> Adaptação com sistema de telemetria LPWA MYWATER ou similar.
Jardim 	JANZ MST 15050 – DN 50 Q3 = 25m³/h; Qmax = 31,5m³/h	Adaptação com sensor de impulsos JANZ JI (Sensor Impulsos) ou similar. <u>Ou</u> Adaptação com sistema de telemetria LPWA MYWATER ou similar.

Figure K.4: Water meter specifications and digitalization strategy - page 4

Montagem Final (Exterior)/ Prova de Água 	JANZ MST 15050 – DN 50 (2013) $Q_3 = 16\text{m}^3/\text{h}$; $Q_{\text{max}} = 32\text{m}^3/\text{h}$	Não foi encontrada qualquer informação sobre o contador permitir a colocação de um dispositivo de medição. Contactar fornecedor e, em última instância, proceder à compra de um novo contador.
Maquinas de Café Máquina de Lavar 	JANZ JV400 - DN 15 ou 20 (por confirmar) $Q_3 = 2,5\text{m}^3/\text{h}$; $Q_{\text{max}} = 3,125\text{m}^3/\text{h}$	Adaptação com sensor de impulsos JANZ JI (Sensor Impulsos) ou similar. <u>Ou</u> Adaptação com sistema de telemetria LPWA MYWATER ou similar.

Figure K.5: Water meter specifications and digitalization strategy - page 5

Fossa Do Esmalte 	<i>SENSUS WP-DYNAMIC DN 65</i> $Q_3 = 25\text{m}^3/\text{h}$; $Q_{\text{max}} = 50\text{m}^3/\text{h}$	Adaptação com sensores de impulso da SENSUS: • RD 01 • OD 01 • OD 03
Cuba Tinta ED (Água Desmin.) 	<i>(Marca???) WP50 – DN 50 ou 60 (por confirmar)</i> $Q_3 = 35\text{m}^3/\text{h}$; $Q_{\text{max}}=?$	Contador em mau estado de conservação, sem possibilidade de adaptação para digitalização de consumo. Solução passa pela compra de um novo contador.

Figure K.6: Water meter specifications and digitalization strategy - page 6

		
Lavagem de Carro antes da Cuba ED – Posto 1 ED 	JANZ JV400 - DN 15 ou 20 (por confirmar) Q3 = 2,5m³/h; Qmax = 3,125m³/h	Adaptação com sensor de impulsos JANZ JI (Sensor Impulsos) ou similar. <u>Ou</u> Adaptação com sistema de telemetria LPWA MYWATER ou similar.
Tanque 3 e 4 do PT 	(2x) JANZ MST6030 - DN 30 Q3 = 6m³/h; Qmax = 12m³/h	Não foi encontrada qualquer informação sobre o contador permitir a colocação de um dispositivo de medição. Contactar fornecedor e, em última instância, proceder à compra de um novo contador.

Figure K.7: Water meter specifications and digitalization strategy - page 7

Tanque 2 (Água Desmin.) PT 	<i>SENSUS MeiStream Plus 65 – DN50 ou 60 (por confirmar)</i> $Q_3 = 40\text{m}^3/\text{h}$; $Q_{\text{max}} = 60\text{m}^3/\text{h}$	Adaptação com sensores de impulso da SENSUS: • HRI-Lei • OD 01 • OD 03
Tanque 2 PT 	<i>JANZ MSV 2525 – DN 20</i> $Q_3 = 2,5\text{m}^3/\text{h}$; $Q_{\text{max}} = 5\text{m}^3/\text{h}$	Não foi encontrada qualquer informação sobre o contador permitir a colocação de um dispositivo de medição. Contactar fornecedor e, em última instância, proceder à compra de um novo contador.
Tanque 1 Desengorduramento PT 	<i>JANZ MST6030 - DN 30</i> $Q_3 = 6\text{m}^3/\text{h}$; $Q_{\text{max}} = 12\text{m}^3/\text{h}$	Não foi encontrada qualquer informação sobre o contador permitir a colocação de um dispositivo de medição. Contactar fornecedor e, em última instância, proceder à compra de um novo contador.

Figure K.8: Water meter specifications and digitalization strategy - page 8