SANTA CATARINA STATE UNIVERSITY – UDESC
COLLEGE OF TECHNOLOGICAL SCIENCE – CCT
POSTGRADUATE PROGRAM – APPLIED COMPUTING

VINÍCIUS MICHELON GEREMIAS

MACHINE LEARNING FOR PREDICTING ATMOSPHERIC CORROSION RATE IN METALLIC MATERIALS USING REANALYSIS DATA

JOINVILLE
2024

VINÍCIUS MICHELON GEREMIAS

MACHINE LEARNING FOR PREDICTING ATMOSPHERIC CORROSION RATE IN METALLIC MATERIALS USING REANALYSIS DATA

Master thesis submitted to the Computer Science Department at the College of Technological Science of Santa Catarina State University in fulfillment of the partial requirement for the Master's degree in Applied Computing.

Supervisor: Rafael Stubs Parpinelli

JOINVILLE
2024

To generate the catalog card for thesis and dissertations, access the link:
https://www.udesc.br/bu/manuais/ficha

Geremias, Vinícius Michelon
Machine learning for predicting atmospheric corrosion rate in metallic materials using reanalysis data / Vinícius Michelon Geremias. - Joinville, 2024.
72 p. : il. ; 30 cm.

Supervisor: Rafael Stubs Parpinelli.
.
Thesis (Master's Degree) - SANTA CATARINA STATE UNIVERSITY, College of Technological Science, Postgraduate Program in Applied Computing, Joinville, 2024.

1. Atmospheric Corrosion. 2. Machine Learning. 3. Reanalysis Data. 4. Corrosion Map. I. Parpinelli, Rafael Stubs . II. SANTA CATARINA STATE UNIVERSITY, College of Technological Science, Postgraduate Program in Applied Computing. III. Título.

VINÍCIUS MICHELON GEREMIAS

MACHINE LEARNING FOR PREDICTING ATMOSPHERIC CORROSION RATE IN METALLIC MATERIALS USING REANALYSIS DATA

Master thesis submitted to the Computer Science Department at the College of Technological Science of Santa Catarina State University in fulfillment of the partial requirement for the Master's degree in Applied Computing.

Supervisor: Rafael Stubs Parpinelli

EXAMINATION BOARD:

Dr. Rafael Stubs Parpinelli
UDESC

Members:

Dr. Danton Ferreira
UFLA

Dr. Gilmario Barbosa dos Santos
UDESC

Joinville, 11th of December, 2024

I dedicate this work to myself.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my entire family, as without them, I would not be here. To my friends, who have always given me the support to keep going and give my best.

I would also like to thank my supervisor, Prof. Dr. Rafael Stubs Parpinelli, for his guidance and dedication throughout these two years of research, highlighting the care and invaluable lessons that will never be forgotten.

I am also deeply grateful to ArcelorMittal Brasil for making this research possible. A special thanks to Dr. Fabiano Miranda, Dr. José Francisco, and Dr. Guilherme Fischer, from whom I had the opportunity to learn so much throughout this process.

Lastly, a more than extraordinary thank you to my girlfriend. Thank you for being by my side in every moment, both in the good times and the challenging ones. Your support, companionship, and, most importantly, your patience were indispensable in this journey.

ABSTRACT

Atmospheric corrosion is a complex chemical process capable of causing billions of dollars in damage to large metallic structures. To address this, studies focus on understanding the corrosive phenomenon and developing predictive methods to reduce costs in metallurgical projects. This research proposes using machine learning models to predict atmospheric corrosion rates for aluminum, carbon steel, copper, and zinc. The models were trained using corrosion rate data from the ISOCORRAG and MICAT projects, combined with reanalysis data, such as sea salt, sulfur dioxide, temperature, and wind speed, among other environmental factors. The machine learning algorithms Random Forest and Extra Trees were employed and demonstrated significant accuracy, surpassing traditional prediction methods, and serve as valuable tools for better anticipating corrosion impacts. In addition to these models, the development of atmospheric corrosion maps provides a visual representation of corrosivity across different regions of South America, helping specialists to identify high-risk areas and design preventive measures more effectively. The results obtained support more informed decision-making, enabling cost optimization and enhancing the longevity of metallic structures. By offering a more comprehensive understanding of corrosion dynamics, this approach contributes significantly to reducing maintenance costs and improving planning in large-scale metallurgical projects, providing a cost-effective and advanced solution for managing atmospheric corrosion risks.

Keywords: Atmospheric Corrosion, Machine Learning, Reanalysis Data, Corrosion Map

RESUMO

A corrosão atmosférica é um processo químico complexo capaz de causar bilhões de dólares em danos a grandes estruturas metálicas. Para enfrentar esse problema, os estudos focam em compreender o fenômeno corrosivo e desenvolver métodos preditivos para reduzir os custos em projetos metalúrgicos. Esta pesquisa propõe o uso de modelos de aprendizado de máquina para prever as taxas de corrosão atmosférica em materiais como alumínio, aço carbono, cobre e zinco. Os modelos foram treinados com dados de taxas de corrosão provenientes dos projetos ISOCORRAG e MICAT, combinados com dados de reanálise, como sal marinho, dióxido de enxofre, temperatura e velocidade do vento, entre outros fatores ambientais. Os algoritmos de aprendizado de máquina Random Forest e Extra Trees foram utilizados e demonstraram uma precisão significativa, superando métodos tradicionais de predição, servindo como ferramentas valiosas para melhor antecipar os impactos da corrosão. Além desses modelos, o desenvolvimento de mapas de corrosão atmosférica fornece uma representação visual da corrosividade em diferentes regiões da América do Sul, auxiliando especialistas a identificar áreas de alto risco e projetar medidas preventivas de forma mais eficaz. Os resultados obtidos apoiam uma tomada de decisão mais informada, permitindo a otimização de custos e aumentando a longevidade de estruturas metálicas. Ao oferecer uma compreensão mais abrangente das dinâmicas da corrosão, essa abordagem contribui significativamente para a redução de custos de manutenção e para o aprimoramento do planejamento em projetos metalúrgicos de grande escala, fornecendo uma solução avançada e econômica para a gestão de riscos de corrosão atmosférica.

Palavras-chave: Corrosão Atmosférica, Aprendizado de Máquina, Dados de Reanálise, Mapas de Corrosão.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ANN	Artificial Neural Network
r_{corr}	Corrosion per Year
CUP	Count of urbanization points
DEM	Digital Elevation Model
DNC	Distance to Nearest Coastline
DRF	Dose-Response Function
ECMWF	European Centre for Medium-Range Weather Forecasts
ESA	European Space Agency
ET	ExtraTrees
INMET	National Institute of Meteorology
LABICOM	Laboratory of Research in Computational Intelligence
LCCS	Land Cover Classification System
MAPE	Mean Absolute Percentage Error
MERRA-2	The Modern-Era Retrospective Analysis for Research and Applications, Version 2
MICAT	*Mapas de Iberoamérica de Corrosividad Atmosférica*
ML	Machine Learning
MLP	Multi-Layer Perceptron
nc	NetCDF
PacIOOS	Pacific Islands Ocean Observing System
RF	Random Forest
RMSE	Root Mean Squared Error
R^2	Coefficient of Determination
SDG	Sustainable Development Goal
SHAP	SHapley Additive exPlanations
SLR	Systematic Literature Review
SMAPE	Symmetric Mean Absolute Percentage Error
SOM	Self-Organizing Maps
PSV	Percentage of Satisfactory Values
UDESC	Santa Catarina State University

LIST OF SYMBOLS

Al	Aluminium
Cl^-	Chloride Ions
Cu	Copper
Fe	Carbon Steel
km/h	Kilometres per Hour
%	Percentage
°C	Degrees Celsius
SO_2	Sulfur Dioxide
µm	Micrometer
Zn	Zinc

CONTENTS

1 INTRODUCTION

Corrosion results from the interaction between a material and its environment, leading to its deterioration. There are many damages caused by this phenomenon, with the increase in costs being the main concern for companies that use the steel in question or other materials that can also suffer from corrosion. Globally, these numbers reach 4 trillion dollars, which is why industries seek techniques to reduce costs and, primarily, the damages caused by corrosion to their materials (ZHI; YANG; FU, 2020).

In addition to the cost of corrosion, studying and understanding corrosiveness is also costly. Conventional methods for categorizing and assessing corrosive environments can require at least one year of continuous observation. During this period, it is essential to ensure the constant functioning of climatic measurement instruments, as described in ISO 9223 (2012).

Faced with the complexity and resource demands of traditional methods, researchers have sought more accessible methodologies, such as developing formulas and intelligent Models (MIKHAILOV; STREKALOV; PANCHENKO, 2007) (GAVRYUSHINA; PANCHENKO, 2023). This project proposes a more effective use of Machine Learning Models by adopting a different approach to collecting environmental data, avoiding the use of mechanical and physical measurement instruments to predict corrosiveness categories. This new approach reduces the costs of localized case studies and significantly decreases the waiting time to determine the corrosiveness of a given environment.

Several formulas have been developed to predict corrosion, but their accuracy is usually not satisfactory for broader applications in different contexts (SANTANA et al., 2019). An example of this is the ISO9223 Standard, which addresses the prediction of atmospheric corrosion for Carbon Steel, Aluminum, Copper, and Zinc (9223, 2012). In addition to formulas, Machine Learning Models can also be employed to predict atmospheric corrosion.

For Models to perform accurate predictions, it is essential to subject them to training using collected data. In the field of atmospheric corrosion, one of the most significant projects focused on collecting and cataloging environmental corrosiveness was the *Mapas de Iberoamérica de Corrosividad Atmosférica* (MICAT) (MORCILLO et al., 1998). This database covers information on atmospheric corrosion throughout Latin America, Portugal, and Spain, providing environmental data that influence this process.

While the MICAT project provides corrosion data and identifies corrosion agents, the complexity of the corrosion process often leaves crucial information gaps necessary for a comprehensive environmental understanding. One such oversight pertains to wind speed, as windier locations tend to transport higher concentrations of gases and pollutants that catalyze corrosive processes (SCHWEITZER P.E., 2006). Enhancing the analysis by including additional corrosive agents can yield more comprehensive insights for the Machine Learning Models, thereby improving the quality of predicted values.

The data for new corrosive agents can be sourced from Global Reanalysis Datasets. These

datasets synthesize data from diverse sources, including satellite observations, weather balloons, aircraft, and ground-based weather stations (GELARO et al., 2017).

One product that provides these reanalysis data is The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) from NASA, which offers climate and environmental data since 1980 with a focus on the entire globe. Another product used in this work, belonging to the European Space Agency (ESA), is the ERA5. Unlike MERRA-2, ERA5 can provide climate data such as temperature and humidity more accurately. However, it does not include aerosol variables, such as sulfur dioxide concentration, an influential gas in atmospheric corrosion (HERSBACH et al., 2023).

As seen in other research, the utilization of Machine Learning Models often yields superior results compared to traditional formulas in predicting Atmospheric Corrosion (ZHI et al., 2019; GAVRYUSHINA; PANCHENKO, 2023; GEREMIAS et al., 2023). To enhance the predictive capabilities of Machine Learning Models, additional data beyond Atmospheric Corrosion Data can be incorporated. Reanalysis Data, which are climate and meteorological datasets providing a comprehensive and coherent view of atmospheric and oceanic conditions, can significantly contribute to improving the accuracy of environmental corrosiveness predictions.

Another advantage of employing Machine Learning Models and Reanalysis Data is the potential for systematic prediction. Currently, understanding the corrosiveness of extensive areas, such as an entire country, requires the installation of multiple corrosion sites across the studied region. Subsequently, experimental results are collected and extrapolated to create a map, as demonstrated in the methodology outlined in Vera et al. (2012), applied to Chile. However, for large countries with diverse vegetation and climate patterns, such as Brazil, this approach may not be viable.

By utilizing Reanalysis Data, coordinate grids can be developed, enabling the systematic application of Machine Learning Models to each point on the grid. Traditional methods cannot map the corrosivity of a large area due to the difficulty and high cost of corrosion stations and rely on extrapolations from the obtained points to create the maps. Extrapolation methods, such as linear extrapolation, estimate values beyond known data points by extending a linear trend. This involves using a linear equation derived from existing data to predict values at new points. However, this approach assumes the trend continues uniformly, which can lead to inaccuracies over larger distances. The approach used in this work allows for the development of atmospheric corrosion maps without relying on extrapolation methods.

This research utilizes data collected on atmospheric corrosion at ground level, along with additional information obtained through reanalysis data, to develop Machine Learning Algorithms capable of predicting the atmospheric corrosion rate. Using these Models, corrosion maps are created to demonstrate corrosivity at a given resolution.

This project is in partnership with the industrial conglomerate ArcelorMittal Brazil, which contributed data and information on atmospheric corrosion of metallic materials, as well as providing theoretical and technical support to corroborate the advancements of this research.

1.1 MOTIVATION

The cost of installing and monitoring an atmospheric station is expensive, along with the sensitivity factor of each sample, which can be easily damaged, resulting in the loss of all related research. Therefore, finding other methodologies that can provide satisfactory and more accurate results can assist the civil and metallurgical industry in reducing costs and increasing productivity. The use of Machine Learning Models in conjunction with reanalysis data acquisition has proven to be a promising approach to obtaining satisfactory results in assessing material corrosivity.

Aligned with the United Nations Sustainable Development Goal (SDG) number 9, this project focuses on building resilient infrastructure, promoting sustainable industrialization, and fostering innovation (NATIONS; DEVELOPMENT, 2015). By leveraging Machine Learning Models alongside reanalysis data acquisition, it aims to enhance the accuracy of material corrosivity assessments while reducing costs and elevating productivity in civil and metallurgical industries. This initiative supports technological advancement and enhances the industry's capacity to address environmental and economic challenges sustainably.

1.2 OBJECTIVES

The main objective of this work is to develop maps of atmospheric corrosion focusing on South America for the four materials mentioned in ISO9223: carbon steel, aluminum, copper, and zinc. To achieve this goal, Machine Learning Models capable of providing satisfactory and accurate results must be developed. These Models will be applied at different points to form a grid and thus develop maps of atmospheric corrosion.

Thus, the specific objectives of this research are:

- Literature review on Atmospheric Corrosion;
- Literature review on Reanalysis Data;
- Literature review on Machine Learning;
- Develop and follow a Workflow;
- Analyze the databases that will be used in the research;
- Develop Machine Learning Models for predicting Atmospheric Corrosion for the four materials of ISO9223;
- Analyze the results and Develop atmospheric corrosion maps.

1.3 DOCUMENT STRUCTURE

The next chapters of this paper will be organized as follows: In Chapter 2, the essential theoretical foundation to understand this work will be provided, covering the topics of Atmo-

spheric Corrosion, Machine Learning, and Reanalysis Data. Chapter 3 will present a systematic literature review on Machine Learning and Atmospheric Corrosion. In Chapter 4, the proposal for this work will be exposed. Chapter 5 will describe how the experiments were conducted, including an analysis of the results obtained. Finally, in Chapter 6, the conclusions and future works to be carried out will be presented.

2 BACKGROUND

In this chapter, the necessary concepts for understanding this research will be presented, including Atmospheric Corrosion, Reanalysis Data, and Machine Learning.

2.1 ATMOSPHERIC CORROSION

Atmospheric corrosion is a type of deterioration that occurs when metals are exposed to the atmosphere, resulting in interaction with atmospheric and climatic components, leading to material degradation (SCHWEITZER P.E., 2006). Although it can also affect other non-metallic materials, such as plastic, these aspects will not be addressed in this research.

Metals are predominantly found in nature in the form of metal oxides and sulfides since these forms have lower free energy and are more stable in the given environmental conditions. However, after extraction and processing, metals transition to a thermodynamically less stable state, making them more susceptible to reoxidation and other forms of corrosion. This process can be understood as the reverse of steelmaking.(GENTIL, 2011).

For atmospheric corrosion to occur, the presence of an electrolyte, often water, is essential. This can be introduced into the reaction through precipitation, fog, or regions with high humidity. When these electrolytes come into contact with the metal, they trigger a redox reaction in which the metal loses electrons and, consequently, mass (SCHWEITZER, 2006).

In addition to the presence of an electrolyte, such as water, other factors can significantly contribute to atmospheric corrosion. One such factor is the presence of corrosive agents, such as sulfur dioxide (SO_2) and chloride ions (Cl^-).

SO_2 is a gas produced by the combustion of sulfur-containing fossil fuels, such as coal and oil. When combined with atmospheric moisture, it forms sulfuric acid (H_2SO_4), which is highly corrosive to many metals. Contact with the resulting sulfuric acid can significantly accelerate the corrosion process, especially in urban and industrial environments (PANNONI, 2017).

Another common corrosive substance in the atmosphere is Cl^-, often present in the form of sodium chloride (NaCl), better known as sea salt. Chloride is particularly corrosive to metals such as iron and steel, and its presence in coastal environments, where there is a higher concentration of salt aerosols, can drastically increase the rate of corrosion (PANNONI, 2017).

In addition to corrosive substances, other environmental variables can influence the results of corrosion. Wind speed plays a crucial role in atmospheric corrosion. Strong winds can carry abrasive particles, such as sand and dust, which can damage the natural protective layer of metals and accelerate the corrosion process. Additionally, wind speed can influence the evaporation rate of moisture from the metal surface, indirectly affecting corrosion (LEYGRAF et al., 2016).

The influence of ambient temperature on the corrosion process can vary significantly. Elevated temperatures tend to accelerate the rates of various chemical and physical processes,

including chemical reactions, potentially increasing the rate of corrosion. On the other hand, high temperatures can also evaporate the water present on metals, thereby reducing the corrosion process (CAI et al., 2020).

Similarly, precipitation can either accelerate or decelerate the corrosion process. Rainy environments tend to have high levels of atmospheric humidity, which can accelerate corrosion. However, rain can also clean metal surfaces, removing corrosive agents such as Cl^- or SO_2, thereby slowing down the corrosion process. Additionally, rain can bring atmospheric pollutants, a phenomenon known as wet deposition, which can introduce additional corrosive substances to metal surfaces, influencing the corrosion process in a complex manner (COLE et al., 2004; TOWNSEND, 2002).

Therefore, in addition to the presence of electrolytes such as water, atmospheric corrosion can be significantly influenced by the presence of corrosive substances like SO_2 and Cl^-, as well as by temperature, precipitation, relative humidity, and wind speed. All these factors play important roles in the corrosion process of metals exposed to the external environment.

2.2 ISO 9223, 9224, 9225, 9226

The ISOs 9223 (2012), 9224 (2012), 9225 (2012), 9226 (2012) constitute a set of international guidelines governing procedures related to atmospheric corrosion. These standards establish a unified classification for atmospheric corrosivity in different regions of the world, providing guidelines for assessing the corrosive potential of the atmospheric environment concerning metallic materials.

Specifically, ISO 9226 (2012) defines the procedures for determining the corrosivity of the environment, requiring the exposure of carbon steel, aluminum, copper, and zinc samples for a minimum of one year. These samples must be placed on a rack with an angle of 45° relative to the surface. After one year of exposure, the mass difference is checked, and the mass loss is calculated in μm/year based on the original weight. The equation used for this calculation is provided in equation 1 (9225, 2012).

$$r_{\text{corr}} = \frac{\Delta m}{A \cdot \rho \cdot t} \tag{1}$$

where:

- r_{corr} is the corrosion rate, expressed in μm/year;
- ρ is the density of the metal (Fe: 7.86 g/cm^3; Zn: 7.14 g/cm^3; Cu: 8.96 g/cm^3; Al: 2.70 g/cm^3);
- Δm is the mass loss, expressed in grams (g);
- A is the surface area, expressed in square meters (m^2);

- t is the exposure time, expressed in years (a).

In Figure 1, a rack with different samples following the ISO methodologies can be observed.

Figure 1 – Example of a rack that follows the methodology of international atmospheric corrosion standards.

Source: (ALMEIDA; PANOSSIAN, 1999)

In addition to the exposure of metals, ISO 9225 (2012) requires the installation of instruments to measure environmental variables such as temperature (°C), relative humidity (%), chloride ion deposition (mg/(m²·day)), and sulfur dioxide deposition (mg/(m²·day)). Temperature and humidity measurements can be performed using electronic devices, while chloride ion and sulfur dioxide deposition measurements utilize candle methods described in ISO 9225.

The candle method for collecting chloride ions involves exposing a moistened textile surface protected from rain for a specific period. The textile, soaked in a glycerol and water solution, captures the chlorides present in the air. After exposure, the amount of deposited chloride is determined by chemical analysis and used to calculate the chloride ion deposition rate. For sulfur dioxide deposition, the candle method uses a lead dioxide surface that reacts with sulfur dioxide in the air, forming lead sulfate, which is quantified to calculate the sulfur dioxide deposition rate (9225, 2012).

In Figure 2, a corrosion station can be seen where, on the left side of the sample racks, the wet candle for chloride ions is represented by the blue box, and on the right, represented by the yellow box, is the wet candle for sulfur dioxide.

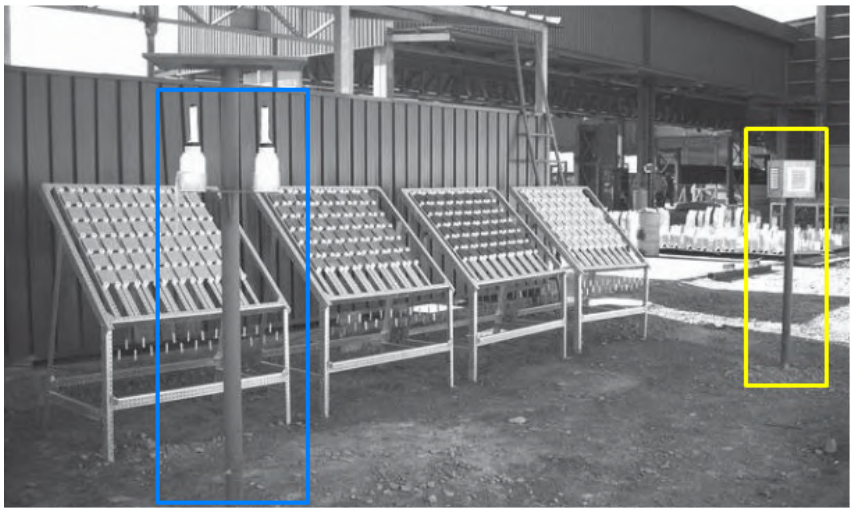

Figure 2 – Atmospheric corrosion station. The sample racks are in the center, with the candle for chloride ions highlighted by the blue box on the left and the candle for sulfur dioxide highlighted by the yellow box on the right.

Source: Source: (VERA et al., 2012)

To estimate the corrosion (r_{corr}) of an environment, ISO 9223 (2012) proposes Dose-Response Functions (DRFs), which can be observed in Equations 2, 3, 4 and 5.

Carbon Steel

$$r_{\text{corr}} = 1.77 \cdot P_d^{0.52} \cdot \exp(0.020 \cdot RH + f_{\text{Fe}}) + 0.102 \cdot S_d^{0.62} \cdot \exp(0.033 \cdot RH + 0.04 \cdot T) \quad (2)$$

$$f_{\text{Fe}} = \begin{cases} 0.150 \cdot (T-10), & \text{if } T \leq 10^{\circ}C \\ -0.054 \cdot (T-10), & \text{otherwise} \end{cases}$$

Aluminium

$$r_{\text{corr}} = 0.0042 \cdot P_d^{0.73} \cdot \exp(0.025 \cdot RH + f_{\text{Al}}) + 0.0018 \cdot S_d^{0.60} \cdot \exp(0.020 \cdot RH + 0.094 \cdot T) \quad (3)$$

$$f_{\text{Al}} = \begin{cases} 0.009 \cdot (T-10), & \text{if } T \leq 10^{\circ}C \\ -0.043 \cdot (T-10), & \text{otherwise} \end{cases}$$

Copper

$$r_{\text{corr}} = 0.0053 \cdot P_d^{0.26} \cdot \exp(0.059 \cdot RH + f_{\text{Cu}}) + 0.01025 \cdot S_d^{0.27} \cdot \exp(0.036 \cdot RH + 0.049 \cdot T) \quad (4)$$

$$f_{\text{Cu}} = \begin{cases} 0.126 \cdot (T-10), & \text{if } T \leq 10^{\circ}C \\ -0.080 \cdot (T-10), & \text{otherwise} \end{cases}$$

Zinc

$$r_{\text{corr}} = 0.0129 \cdot P_d^{0.44} \cdot \exp(0.046 \cdot RH + f_{\text{Zn}}) + 0.00175 \cdot S_d^{0.57} \cdot \exp(0.008 \cdot RH + 0.085 \cdot T) \quad (5)$$

$$f_{\text{Zn}} = \begin{cases} 0.038 \cdot (T - 10), & \text{if } T \leq 10°C \\ -0.071 \cdot (T - 10), & \text{otherwise} \end{cases}$$

As can be observed, the DRFs estimate the corrosion rate based on parameters such as Relative Humidity (RH), Temperature (T), Sulfur Dioxide Deposition (Pd), and Chloride Ion Deposition (Sd). However, some significant variables, such as distance from the sea, wind speed, and precipitation, are not considered in these formulas, leading to potential loss of information. Additionally, the equations were developed based on data collected in the ISOCORRAG project, whose collection stations are concentrated in Europe. Due to these limitations, other authors, such as Rios-Rojas et al. (2017), Tidblad et al. (2001), Mikhailov, Strekalov e Panchenko (2007), and Santana et al. (2019), have developed their own prediction formulas, suggesting that the DRFs may not be suitable for all cases.

2.3 ATMOSPHERIC CORROSION PROJECTS

Determining the corrosivity of an environment requires at least one year of research, as corrosion is evaluated in terms of annual mass loss, usually measured in micrometers per year (μm/year). The one-year exposure period is necessary to account for the full range of seasonal cycles, including variations in temperature and relative humidity, which significantly affect corrosion rates. Due to this complexity, establishing and maintaining a study on atmospheric corrosion is expensive and demands considerable effort. Any incident that damages the samples, such as sensor failure or metal plates falling to the ground, necessitates the disposal of the affected samples. This difficulty results in fragmented and localized studies in different regions, making it challenging to collect large-scale corrosion data. Despite these adversities, two of the largest atmospheric corrosion projects, ISOCORRAG and MICAT, were developed in the 1980s.

The ISOCORRAG Project was conducted between 1986 and 1998, involving the participation of 13 countries and the establishment of 53 corrosion study sites in various parts of the world. This project was fundamental for the creation of the first version of ISO 9223 and also defined guidelines for conducting case studies. In addition to the samples, sensors were installed to collect atmospheric data, including temperature, relative humidity, sulfur dioxide concentration, and chloride ion deposition. At each monitoring station, samples of four different materials were provided: Carbon Steel, Aluminum, Copper, and Zinc. This diversity is crucial because materials react differently to environmental conditions (DEAN DAGMAR KNOTKOVA, 2010).

In contrast to ISOCORRAG, the MICAT project was conducted between 1988 and 1994, focusing on the Americas, Portugal, and Spain, with the participation of 14 countries. Following the procedures adopted by ISOCORRAG, MICAT also collected samples of the four previously

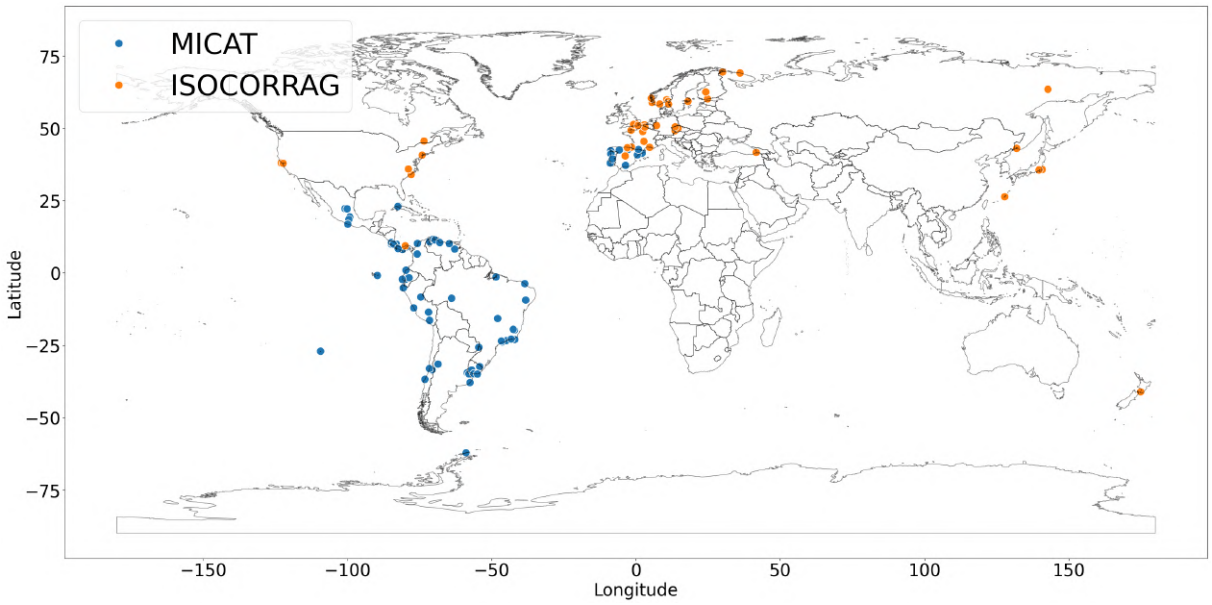

Figure 3 – Location of corrosion sites for the MICAT and ISOCORRAG projects.

Source: Developed by the author

mentioned materials and monitored the same four environmental variables (MORCILLO et al., 1998).

Figure 3 presents the countries participating in the mentioned projects. In some cases, countries participated in both projects. As can be observed, the projects explore different regions of the globe, each possessing distinct properties.

Another point to highlight is that both projects collect only four environmental variables as corrosive factors. As clarified in this Chapter, atmospheric corrosion is a complex chemical process in which various factors influence the mass loss of materials. Therefore, the omission of data collection for variables such as wind speed and precipitation may compromise the comprehensiveness and integrity of the analyzed samples.

climate and meteorological datasets providing a comprehensive and coherent view of atmospheric and oceanic conditions

2.4 REANALYSIS DATA AND COMPLEMENTARY DATASETS

Reanalysis data consists of extensive climate and meteorological datasets that integrate historical observations with advanced numerical models to generate a consistent, high-resolution representation of atmospheric and oceanic conditions over time. These datasets provide detailed information on variables such as temperature, humidity, wind speed, and atmospheric pressure, offering a reliable foundation for climate studies, weather forecasting, and environmental modeling (HERSBACH et al., 2023). These data are generated through a complex process of assimilating observations from satellites, meteorological stations, and other measuring instruments combined with substance transport models. By integrating these diverse sources of information

over a specific period, usually decades, reanalysis data enable the retrospective reconstruction of the climate and environmental system's state on a regular spatial and temporal grid (GELARO et al., 2017). Various companies, institutes, and organizations develop their reanalysis datasets for different uses, each with its distinct methodology. Examples of reanalysis datasets include ERA5 and MERRA-2. Here in this dissertation, it was used other datasets that are not considered Reanalysis datasets but are also important: Distance to Nearest Coastline (DNC), Global 1-km Digital Elevation Model (DEM), and LandCover.

Developed by the European Centre for Medium-Range Weather Forecasts (ECMWF), ERA5 provides detailed information on a variety of climate variables, such as temperature, relative humidity, wind, and precipitation, with a spatial resolution of 100 km². These data are widely used in various applications, including scientific research, weather forecasting, climate studies, environmental monitoring, and modeling climate impacts across different sectors such as agriculture, energy, and public health (HERSBACH et al., 2023).

The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2), is an atmospheric and oceanic reanalysis dataset developed by NASA, offering a detailed and accurate reconstruction of the climate system's state from 1980 to the present. MERRA-2 utilizes a variety of observational sources, including satellite observations, atmospheric sondes, ocean buoys, and meteorological station data, to assimilate information into advanced numerical substance transport models. This combination of observations and modeling allows for a more precise representation of atmospheric and oceanic conditions across a variety of variables, such as temperature, humidity, wind, sulfur dioxide concentration, and chloride ion concentration. The dataset features a distribution of points with 50 km latitude and 70 km longitude, resulting in a resolution of 3,500 km² (GELARO et al., 2017).

The Pacific Islands Ocean Observing System (PacIOOS), located in Hawaii, United States, provides two datasets relevant to this research, as follows:

The first is the Distance to Nearest Coastline (DNC), which indicates the distance in kilometers to the nearest coastline for any geographic coordinate. This dataset has a resolution of 1 km². The second is the Global 1-km Digital Elevation Model (DEM), which also has a resolution of 1 km² and provides the elevation in meters for any point on the planet (SANDWELL; SMITH; BECKER, 2014).

Figure 4 presents the DNC and DEM data in map format. For the DEM, the greener the region, the higher its elevation. In contrast, for the DNC, the greener the area, the greater the distance to the nearest coastline.

The Copernicus Land Cover dataset provides global land cover data derived from the Sentinel-3 satellite at a 300-meter spatial resolution. It focuses on 22 land cover classifications, which include forests, croplands, urban areas, water bodies, and wetlands, among others. The classification system is built using the FAO's Land Cover Classification System (LCCS), ensuring standardization across different geographic areas. The dataset is updated annually and has data from 1993 until 2024 (Copernicus Climate Change Service, 2019).

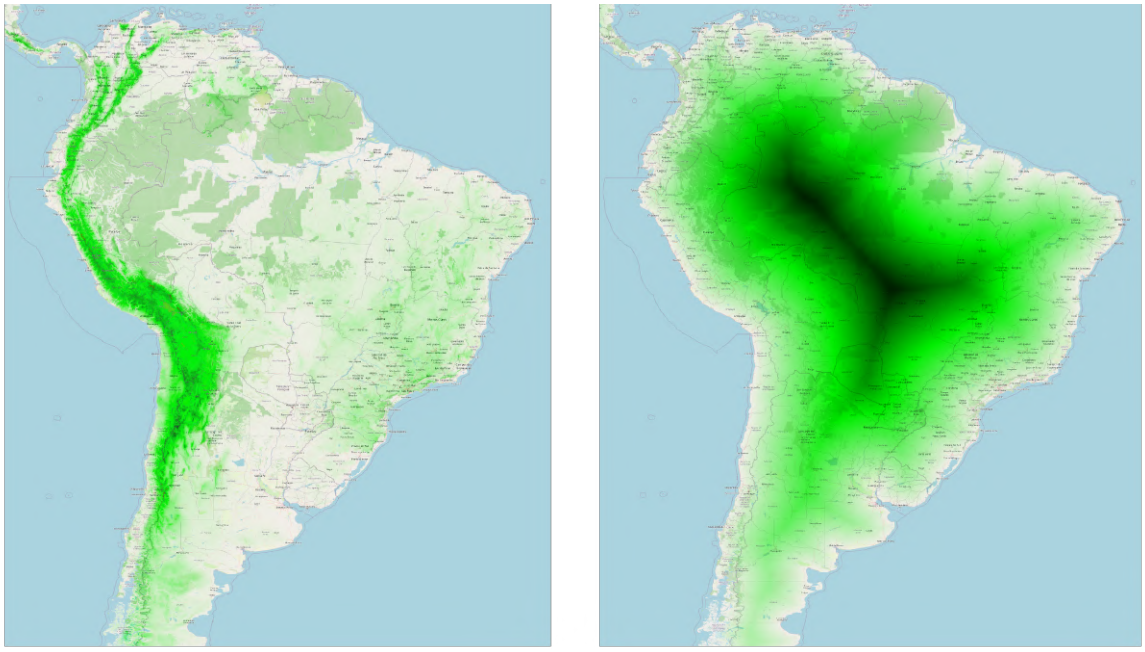

Figure 4 – Maps of South America with a resolution of 1 km², showing the Digital Elevation Model (DEM) on the left and the Distance to Nearest Coastline (DNC) on the right. Darker green areas represent higher values.

Source: Developed by the author

This data is widely applied across various fields, such as environmental monitoring, agriculture, forestry, and urban planning. It provides valuable insights into changes in land cover over time, enabling decision-makers and researchers to evaluate the impact of human activities and natural processes on ecosystems and biodiversity. Beyond these applications, the Land Cover dataset also facilitates the analysis of terrain roughness using the WINDPRO roughness classification, as demonstrated in Table 1. The roughness parameter, z0, quantifies the effect of the terrain's surface irregularities on the flow of wind near the ground. Higher z0 values indicate greater roughness, which is typical for environments such as urban areas and dense forests (EMD, 2005).

All the datasets mentioned are publicly accessible, making them available to any institution or individual interested, thus facilitating the widespread and inclusive use of this information. Additionally, a significant aspect of these projects is that both provide their data in the NetCDF (.nc) file format, which is commonly used in climate and meteorological analyses. These files can be loaded and manipulated by modern libraries, such as xarray (HOYER; HAMMAN, 2017), and high-level programming languages like Python (ROSSUM; DRAKE, 2011). This accessibility and compatibility with modern analysis tools contribute to the effective dissemination and use of these datasets, allowing researchers, scientists, and analysts to extract detailed information about the climate and its trends.

Value	Land Cover Type	Roughness Class
11	Post-flooding or irrigated croplands (or aquatic)	z0=0.100
14	Rainfed croplands	z0=0.100
20	Mosaic cropland (50-70%) / vegetation (20-50%)	z0=0.07
30	Mosaic vegetation (50-70%) / cropland (20-50%)	z0=0.07
40	Closed to open broadleaved evergreen or semi-deciduous forest	z0=0.5
50	Closed broadleaved deciduous forest (>5m)	z0=0.4
60	Open broadleaved deciduous forest/woodland (>5m)	z0=0.4
70	Closed needleleaved evergreen forest (>5m)	z0=0.5
90	Open needleleaved deciduous or evergreen forest (>5m)	z0=0.4
100	Closed to open mixed broadleaved and needleleaved forest (>5m)	z0=0.4
110	Mosaic forest or shrubland (50-70%) / grassland (20-50%)	z0=0.07
120	Mosaic grassland (50-70%) / forest or shrubland (20-50%)	z0=0.07
130	Closed to open shrubland (<5m)	z0=0.07
140	Closed to open herbaceous vegetation	z0=0.05
150	Sparse (<15%) vegetation	z0=0.07
160	Closed to open broadleaved forest regularly flooded	z0=0.1
170	Closed broadleaved forest or shrubland permanently flooded	z0=0.1
180	Closed to open grassland or woody vegetation on waterlogged soil	z0=0.03
190	Artificial surfaces and associated areas (Urban areas >50%)	z0=0.4
200	Bare areas	z0=0.02
210	Water bodies	z0=0.0002
220	Permanent snow and ice	z0=0.001
230	No data (burnt areas, clouds, . . .)	z0=N/A

Table 1 – Land Cover Types and their corresponding Roughness Classes.

2.5 ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Artificial intelligence (AI) is a field of computer science dedicated to the development of systems and algorithms capable of performing a variety of tasks that would normally require human intervention. These tasks range from pattern recognition to natural language processing and computer vision, among others (RUSSELL; NORVIG, 2010).

One of the most prominent subfields of AI is Machine Learning (ML), which focuses on creating algorithms and techniques that enable computers to learn from data and past experiences without the need for explicit programming. Within ML, a significant approach is supervised learning, where the model is trained with a labeled dataset. Each training data point consists of an input and the corresponding desired output, allowing the algorithm to identify patterns in the data and produce accurate predictions (NASTESKI, 2017).

Figure 5 illustrates the workflow of supervised learning models. First, the data is collected and then divided into training and testing datasets. The training data is used to train the model,

which is subsequently evaluated with the testing data. If the Model performs well, the process is concluded. If not, the cycle is restarted, refining the data and the Model until the desired performance is achieved.

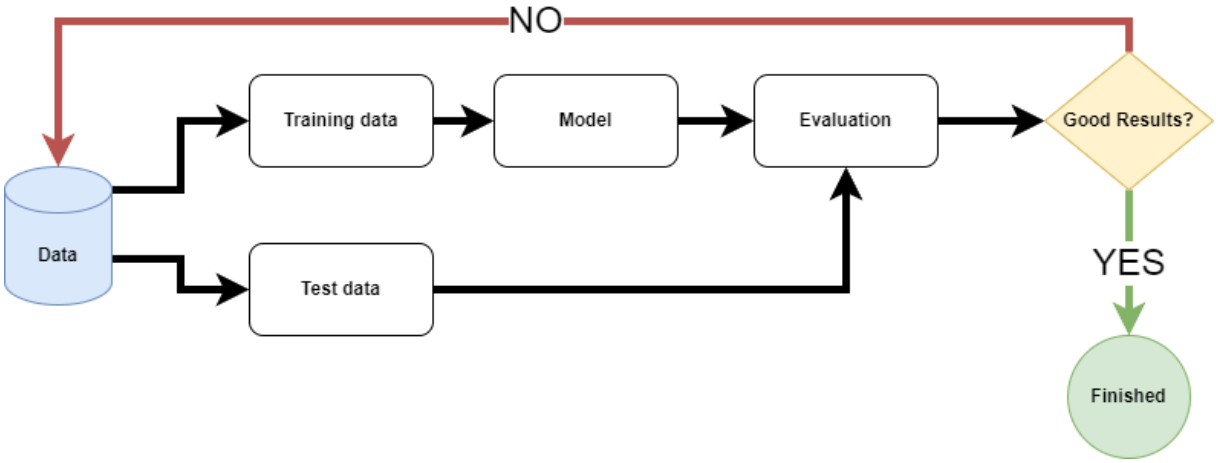

Figure 5 – Workflow of supervised learning Models.

Source: Developed by the author

In supervised learning, there are two types of problems: classification and regression. In classification, the objective is to assign a category or class to a dataset based on its features. For example, determining whether an email is spam or not spam, or identifying whether an image contains a cat or a dog. Classification algorithms aim to learn patterns in the training data that allow for correct assignments for new, unseen examples. On the other hand, in regression, the focus is on predicting a numerical value based on a set of independent variables. For instance, predicting the price of a house based on characteristics such as size, number of rooms, and location. In this case, regression algorithms seek to find a pattern in the training data that allows for good predictions of numerical values for new examples (JAMES et al., 2023).

During model training, it is necessary to identify and avoid two behaviors that can occur: overfitting and underfitting. Overfitting happens when the model fits the training data too well, capturing even the noise in the data but failing to generalize to new data, resulting in poor performance. Underfitting, on the other hand, occurs when the model fails to identify a trend in the data, leading to unsatisfactory performance on both the training data and new data (GERON, 2017).

There are various Machine Learning algorithms applicable to regression and classification problems. These algorithms were developed without a specific focus on a defined problem, meaning they can be applied in various contexts with flexibility and adaptability. This versatility allows them to be used in a wide range of domains, respecting the requirements and particularities of each problem. Two Machine Learning Algorithms were used in this work, ExtraTrees (regression method applied in atmospheric corrosion rate for Aluminum, Carbon Steel, Copper, and Zinc) and Random Forest (to enhance the accuracy of SO_2 deposition).

2.5.1 ExtraTrees

The ExtraTrees (ET) Algorithm is a technique in the field of supervised learning that is applicable to both classification and regression problems. This method distinguishes itself from traditional decision tree-based algorithms by its high level of randomization in choosing cut points and attributes during node splitting. ExtraTrees construct an ensemble of decision or regression trees using the entire training sample instead of bootstrap samples.

At each node, K attributes are randomly selected from the total set of features. For each of these attributes, a cut point is randomly determined rather than chosen based on an optimization criterion. A cut point is a threshold value that splits the data into two subsets. Among the K generated cut points, the one that optimizes a predefined evaluation metric, such as variance reduction for regression, is selected. The process is repeated recursively until a stopping criterion, such as a minimum number of samples per leaf or a maximum tree depth, is met (GEURTS; ERNST; WEHENKEL, 2006).

By combining high randomization with the ensemble process, ExtraTrees increases the diversity of individual trees and reduces the variance of the final model without significantly increasing bias. This technique proves particularly effective in problems where individual decision trees exhibit high variability (BUI; NGUYEN; SOUKHANOUVONG, 2022).

Key hyperparameters for ExtraTrees include the number of trees and the maximum depth of each tree. Generally, increasing the number of trees enhances model stability and reduces variance by averaging results across more trees. Controlling the maximum depth determines how deep each tree can grow. Deeper trees can capture more complex patterns but may also lead to overfitting. Proper tuning of these hyperparameters is essential for the model to make better predictions (ASIF et al., 2023).

At the end of the training, regression predictions are determined by combining the predictions of all the decision trees in the ensemble, where the final prediction is obtained by averaging the individual predictions of each tree. This approach results in a robust regression model capable of handling various types of data, beneficial when aiming to avoid overfitting and achieve accurate predictions in complex datasets with many features (GEURTS; ERNST; WEHENKEL, 2006).

Figure 6 represents the functioning of the ExtraTrees Algorithm. The diagram illustrates how the dataset is used to build multiple decision trees, each with a random selection of attributes. Each of these trees generates an individual prediction. The individual predictions from all the trees are then aggregated by averaging the predictions to produce the model's final prediction.

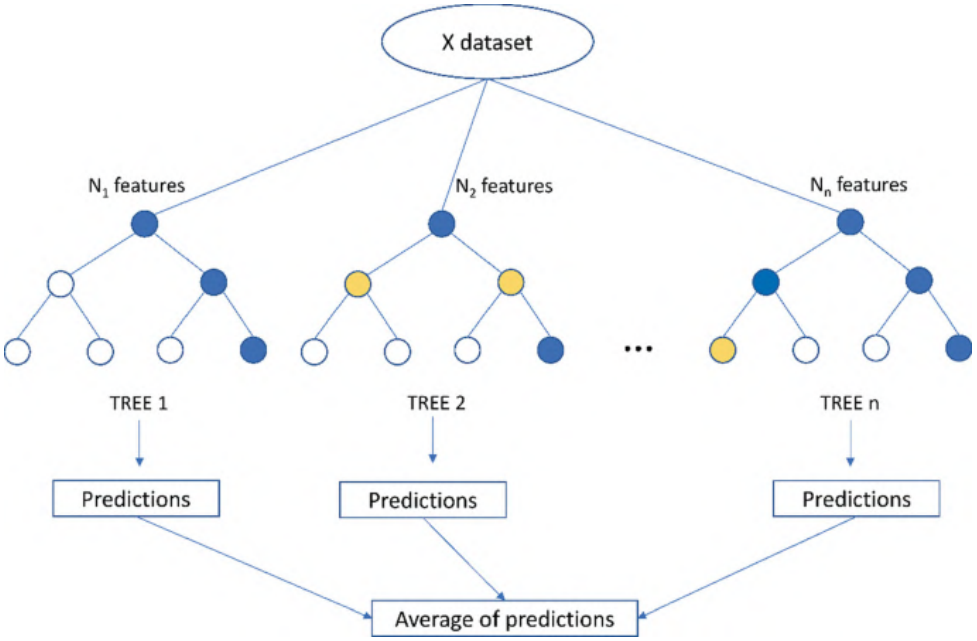

Figure 6 – Functioning of the Extra-Trees Algorithm, where multiple decision trees are constructed with random selections of features, and their predictions are aggregated to produce the final prediction.

Source: (BUI; NGUYEN; SOUKHANOUVONG, 2022)

2.5.2 Random Forest

Random Forest is a supervised learning algorithm, similar to ExtraTrees, that creates an ensemble of decision trees, where each tree is generated from a random sample of the training dataset. Additionally, during the construction of each tree, a random subset of variables is selected for each node split, which reduces the correlation between trees and improves the model's generalization. The process of node splitting is fundamental to the model's performance. At each split, the algorithm evaluates multiple candidate features and selects the one that results in the best partition of the data based on a chosen criterion. For regression tasks, the usual criterion is the Mean Squared Error (MSE), which evaluates the variance reduction after the split. The tree construction continues recursively until a predefined stopping condition is met, such as a maximum depth, a minimum number of samples per leaf node, or when further splitting does not significantly improve the prediction. This technique combines the predictions of all trees to arrive at a single final prediction (BREIMAN, 2001).

The main difference between Random Forest and ExtraTrees lies in the level of randomness introduced during tree construction. In Random Forest, the trees are built from bootstrap samples, and the selection of variables for node splitting is done randomly, but the split point is still determined in a more optimized manner. In ExtraTrees, in addition to the random selection of variables, the split point is also chosen randomly, without necessarily seeking the optimal split, further increasing the model's randomness (ALJAMAAN; ALAZBA, 2020).

2.5.3 K-fold Cross-validation

K-fold cross-validation is a technique used to estimate the performance of a machine learning model by dividing the dataset into K parts. In each iteration, K - 1 parts are used to train the model, while the remaining part is used to test it. This process is repeated K times, ensuring that all parts are used for both training and testing. The evaluation metrics from each iteration are stored, and at the end, the average of these metrics is calculated (BERRAR, 2019). In Figure 7, an example of 5-fold Cross-validation is shown, where the data is split into 5 folds, and each fold is used once as a test set and four times as a training set. At the end of the five iterations, the average performance of the model is calculated based on the metrics obtained in each fold.

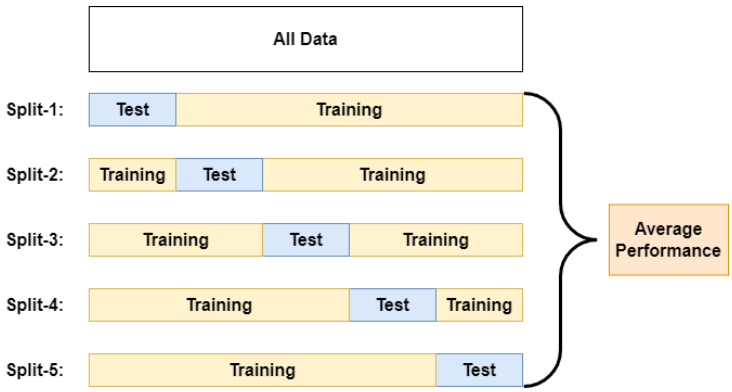

Figure 7 – Example of a 5-fold Cross-validation, where the yellow boxes represent the training data and the blue boxes represent the test data

Source: Developed by the author

2.5.4 Model Evaluation

Various metrics are employed to assess the performance of predictions relative to the actual values in regression model evaluation. Some of the most common metrics include Root Mean Squared Error (RMSE), the coefficient of determination (R^2), and Spearman's correlation.

a) RMSE: is a measure of the difference between the values predicted by the model and the actual values. It calculates the square root of the mean of the squared errors between the predictions ($\hat{y}_i$) and the true values (y_i). RMSE provides a measure of the dispersion of residuals and is useful for determining the average magnitude of the model's errors (JAMES et al., 2023). RMSE can be visualized in equation 6:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{6}$$

The value of $\hat{y}_i$ represents the predicted value, y_i represents the actual value, and n represents the number of samples.

b) R^2: It is a metric that quantifies the proportion of data variance explained by the constructed model. Thus, it indicates how well the model fits the real application

data. The value of R^2 ranges from 0 to 1, where 0 indicates a completely poorly fitted model and 1 indicates a perfectly fitted model (GERON, 2017). The equation for R^2 is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{7}$$

c) Spearman's Correlation (ρ): it is a measure of association between variables. Spearman's Correlation assesses the relationship between two variables in a way that does not depend on linearity, as it is based on the ranks of the data points rather than their raw values. Each variable is ranked, and the correlation is calculated using these ranks. The ranking is determined by ordering the values of each variable in ascending order and assigning ranks accordingly, with the smallest value receiving rank 1, the second smallest rank 2, and so on. This correlation ranges from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. The equation for Spearman's Correlation is given by:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{8}$$

The value of d_i is the difference between the ranks of each pair of samples.

2.5.5 SHapley Additive exPlanations

Shapley Additive exPlanations (SHAP) is a methodology grounded in the Shapley value theory from cooperative game theory, applied to the field of Machine Learning Model interpretability. The primary objective of SHAP is to assign the contribution of each feature to a Model's prediction, thereby enabling a better understanding of how variables influence the predicted outcome (MARCíLIO; ELER, 2020).

The application of Shapley values in Machine Learning involves constructing an explanatory Model that approximates the original predictive Model's function, using only the features of interest. SHAP combines the predictions of this explanatory Model to calculate the importance of each feature, resulting in values that can be directly interpreted. These Shapley values are visualized in various ways, such as bar charts or dependency plots, allowing analysts to understand which features are most influential and how they interact with each other (LUNDBERG; LEE, 2017).

Figure 8 shows an example of a SHAP plot for the California housing prices dataset [1]. Each line represents a variable, ordered by its influence on the Model, from the most to the least

[1] Californa housing prices: <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html>

influential. The color of the points indicates the feature value, with blue representing low values and red representing high values. Each point on the plot represents an instance from the dataset, indicating the impact of each feature on the Model's output.

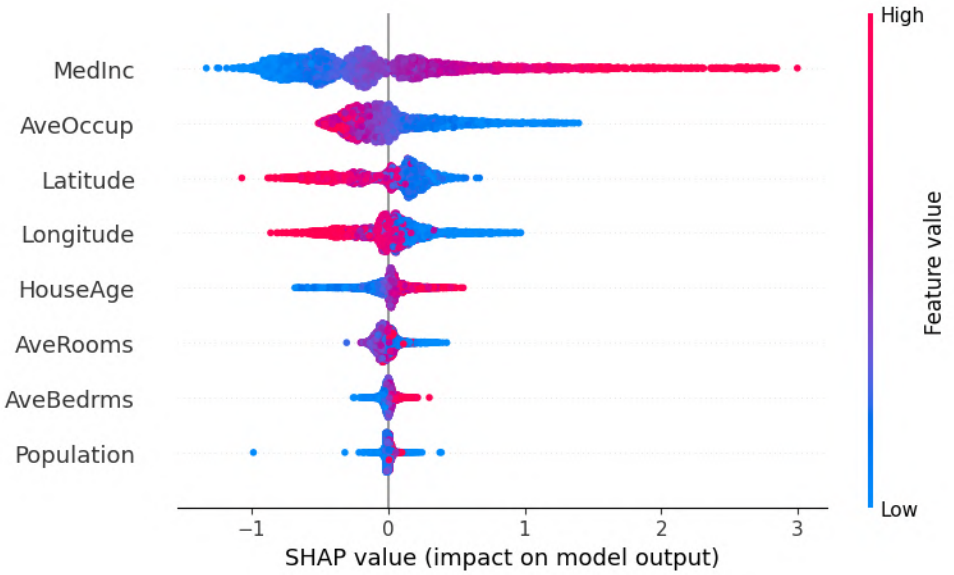

Figure 8 – SHAP plot for the California housing prices dataset, showing the influence of variables on the Model.

Source: Developed by the author

3 SYSTEMATIC LITERATURE REVIEW

This work proposes conducting a Systematic Literature Review to identify and analyze related studies that have already been conducted, highlighting their methodologies, results, and contributions to the field of Machine Learning applied to Atmospheric Corrosion.

The motivation for conducting this Systematic Literature Review (SLR) arose from the absence of a specific SLR on the topic of "Machine Learning applied to Atmospheric Corrosion." Although a similar review was found in Coelho et al. (2022), that publication focuses on the application of Machine Learning to Corrosion in general, without specifically concentrating on Atmospheric Corrosion.

In the review conducted by Coelho et al. (2022), various articles on corrosion that utilize Machine Learning techniques for predictions were cataloged. Their research was conducted exclusively in the Web of Science database, using the search terms: "machine learning" AND "corrosion". Although the identified articles cover various types of corrosion, 9 specific articles on atmospheric corrosion were found. These articles are relevant but were not included in the scope of the present work.

3.1 METHODOLOGY

The Systematic Literature Review was conducted in 2023 using three different academic search engines: Springer Link, Scopus, and IEEE Xplore. Various search terms were tested, and the ones that returned the most interesting and relevant articles for this research topic was: "machine learning" AND "atmospheric corrosion". Although other keywords were tested, the scarcity of specific articles indicated that expanding the search terms did not yield significant results. The selected search term was applied to the titles, abstracts, and bodies of the articles.

To select the articles, the following exclusion criteria were adopted:

- Not regression-based
- Does not follow the ISO9223 collection standards
- Not focused on atmospheric corrosion prediction

In addition to these criteria, only articles in English and Portuguese were included in the final results due to the author's linguistic limitations.

During the research, it was observed that some authors published several similar articles, merely adding new features to their prediction Models. To avoid redundancy, only one of these similar articles was arbitrarily selected.

3.2 RESULTS

The first search engine used was IEEE Xplore; however, no results were found.

Through the Scopus search engine, 32 articles covering the period from 2018 to 2023 were initially identified. After applying the exclusion criteria, 7 relevant articles remained. It is important to note that some identified articles could not be accessed and thus were not included.

In the final academic search engine used, Springer Link, 17 articles were found. After applying the exclusion criteria and removing duplicates identified through other search engines, 3 articles remained. The majority of the articles found cited other studies that utilized Machine Learning methods to predict atmospheric corrosion. However, many of these articles were not directly related to the specific topic of this study, leading to their exclusion.

3.3 ARTICLES ANALYSIS

In the article by Tran et al. (2021), a Multi-Layer Perceptron (MLP) Artificial Neural Network (ANN) was developed to predict the Corrosion Rate of Carbon Steel. The author utilized a private dataset containing corrosion samples from Vietnam. The variables considered by the Model included Temperature, Rainfall, Hours of Sunlight Exposure, Relative Humidity, and other pollutants. According to the author, the most influential variables for the Model were Rainfall and Hours of Sunlight Exposure, notable for their unconventionality in the literature. The evaluation metrics used included RMSE, MAPE, and R^2, revealing that the Model achieved a coefficient of determination of 0.99 for the test data.

Terrados-Cristos et al. (2021) employed Self-Organizing Maps (SOM) to develop a Model for predicting the Corrosion Rate of Zinc. The ISOCORRAG dataset used by the author includes samples from various regions around the world, with an emphasis on European countries. Unlike other methods, the author used environmental categories, such as marine, industrial, and urban, as input variables, thereby broadening the range of characteristics considered in the Model. The only evaluation metric presented was R^2, with the Model achieving a coefficient of determination of 0.77 for the test data.

Unlike the previously mentioned articles, Zhi et al. (2021) uses the Random Forest (RF) Algorithm to identify the most important parameters for atmospheric corrosion of Carbon Steel. The author employs a dataset from corrosion stations located in China, which have a different climate compared to Brazil. To identify the variables, the correlation coefficient metric was used, and sampling was conducted for different corrosion periods.

A similar approach to the one presented by Zhi et al. (2021) is adopted by Yan, Diao e Gao (2020). However, the author employs a private dataset containing information from Japanese stations. Additionally, the correlation coefficient is applied to evaluate the performance of the Random Forest Algorithm in predicting the corrosion rate of Carbon Steel. A distinction between this work and others is the use of the Explainable AI tool called SHAP. This tool identifies the main variables influencing the Model, helping to understand the reasons behind certain results.

Other authors also use Random Forests for the problem of atmospheric corrosion of Carbon Steel, such as Pei et al. (2020). In addition to using this Algorithm, they found that, with

Author	Year	Algorithm	Metrics	Material	Region	Database
Ben Seghier et al.	2021	ANN	RMSE, R^2	Carbon Steel	Global	Karanci and Betti
Chae et al.	2022	RF	RMSE, MAE	Nickel	China	Authorship
Gavryushina et al.	2023a	RF	MAPE, SMAPE, PSV, R^2	Carbon Steel	Global	ISOCORRAG and MICAT and Mikhailov
Gavryushina et al.	2023b	RF	MAPE, SMAPE, PSV, R^2	Aluminum	Global	ISOCORRAG and MICAT and Mikhailov
Terrados-Cristos et al.	2021	SOM	R^2	Zinc	Global	ISOCORRAG
Tran et al.	2021	ANN	RMSE, MAPE, R^2	Carbon Steel	Vietnam	Authorship
Pei et al.	2020	RF	RMSE, R^2	Carbon Steel	China	Authorship
Yan et al.	2020	RF	R^2, MAE	Carbon Steel	Japan	National Institute of Materials Science (NIMS)
Yang et al.	2022	RF	MSE	Carbon Steel	China	National Material Environmental Corrosion Platform of China
Zhi et al.	2021	RF	RE	Carbon Steel	China	Authorship
This Work	2024	ET	RMSE, R^2	Al, Fe, Cu and Zn	Global	ISOCORRAG and MICAT

Table 2 – Summary identifying the main attributes found in each article.

their dataset, this Model performed better than other Models using different Algorithms, such as Artificial Neural Networks and Support Vector Machines. In their article, the instantaneous corrosion rate was predicted, which differs from others that forecast corrosion for one or more years. RMSE and R^2 were used as evaluation metrics, with the best case showing an R^2 value of 0.52.

Using a database with stations in various regions, Gavryushina e Panchenko (2023) developed Models for predicting atmospheric corrosion of Carbon Steel. Their unique approach involved using data from major corrosion studies such as MICAT, ISOCORRAG, ICP/UNESP, and a Russian database. Like previous authors, they used Random Forest as the Algorithm for the Models and achieved good results. The evaluation metrics included Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), Percentage of Satisfactory Values (PSV), and R^2.

In Gavryushina, Marshakov e Panchenko (2023), the author applied the same methodology used in Gavryushina e Panchenko (2023), but for aluminum. In this article, other Dose-Response Functions (DRFs) not included in ISO9223 were also developed. Despite using these new functions, the Machine Learning Models indicated better results.

Yang et al. (2022) produced a somewhat different work compared to the others. The study utilized electronic sensors to measure the atmospheric corrosion of Carbon Steel at Chinese stations. Using this data, along with a Random Forest Algorithm, the author developed a regression Model for atmospheric corrosion.

Chae et al. (2022) utilized a different Chinese dataset containing atmospheric corrosion data for Nickel. In this study, the author employed a Random Forest Algorithm to identify the variables that most influence the corrosion of this material. A distinguishing feature of this work is that all samples were exposed to extreme conditions, with an average temperature of 650°C. For this reason, this study will not be related to the present work, which focuses solely on samples at ambient temperature.

In the last selected work, Seghier et al. (2021) developed Models to predict the annual corrosion rate for suspension bridge cables. The author used 303 samples located in different parts of the world to train their Artificial Neural Network Models. These Models utilize some variables already mentioned in other articles, such as Temperature and Chloride Ions levels. The evaluation metrics used were RMSE and R^2, with the best Models achieving an R^2 of 0.95.

Table 2 presents a summary of the identified articles, each categorized by year, Algorithm used, evaluation metrics, material analyzed, and associated region. All the retrieved articles were published within the last five years, suggesting the contemporary relevance of the topic. However, it was initially expected to find a higher proportion of older articles, given that the major atmospheric corrosion projects date back to the 1980s.

Upon analyzing the table, it is evident that the Random Forest (RF) Algorithm is the most commonly used and demonstrates superior results. Regarding the evaluation metrics, there is no standardization among the authors, although all the employed metrics are commonly used in the Machine Learning literature.

It is relevant to note that most studies focus on predictions related to Carbon Steel. This alignment is beneficial for this work, which aims to make predictions for specific materials such as Carbon Steel, Copper, Zinc, and Aluminum, as these are the materials established by ISO9223.

3.4 SLR CONSIDERATIONS

The systematic review highlights the growing interest in applying Machine Learning techniques to the analysis of atmospheric corrosion. The reviewed studies demonstrate that the results obtained through these techniques have the potential to reduce costs related to atmospheric corrosion, enabling the identification of risk factors and the implementation of more effective preventive measures. The analysis of the selected articles reveals the relevance and diversity of the approaches adopted, underscoring the capability of Machine Learning Algorithms to recognize the corrosive trend of a specific location and identify patterns that result in more accurate predictions compared to traditional mathematical methods.

However, it is important to acknowledge that the application of Machine Learning in the analysis of atmospheric corrosion still faces challenges and limitations. The need for high-quality and representative data, along with the interpretation of the obtained results, are aspects that require careful attention. Additionally, the selection of relevant environmental variables and the validation of the developed Models are factors to consider to ensure the reliability of the predictions.

Several relevant articles were identified during the research. However, while conducting manual searches of previous related works, many equally interesting articles were found. One challenge that impacts the efficiency of the systematic search is the limitation imposed by publication dates. Given that corrosion was a widely discussed topic in the 1990s, academic search engines struggle to automatically locate these works, making it necessary to refer to specific sources.

Despite these challenges, the systematic review reinforces that Machine Learning yields significant results in predicting atmospheric corrosion rates. The diversity of approaches and the ongoing evolution of Machine Learning techniques indicate a promising field for future research

and practical applications in the context of atmospheric corrosion, which will be explored in this master's thesis.

This study stands out by proposing the development of four distinct atmospheric corrosion models specific to the metals Al, Fe, Cu, and Zn. An innovative feature of this work is the creation of atmospheric corrosion maps based on the developed models, providing a visual tool that can assist specialists in decision-making.

4 PROPOSED APPROACH

In this chapter, the Machine Learning Models proposed for atmospheric corrosion prediction will be presented, along with the methodology for developing corrosion maps. To facilitate understanding, a flowchart covering all stages of the proposal has been prepared, which can be seen in Figure 9. Furthermore, throughout this chapter, the flow will be explained in detail.

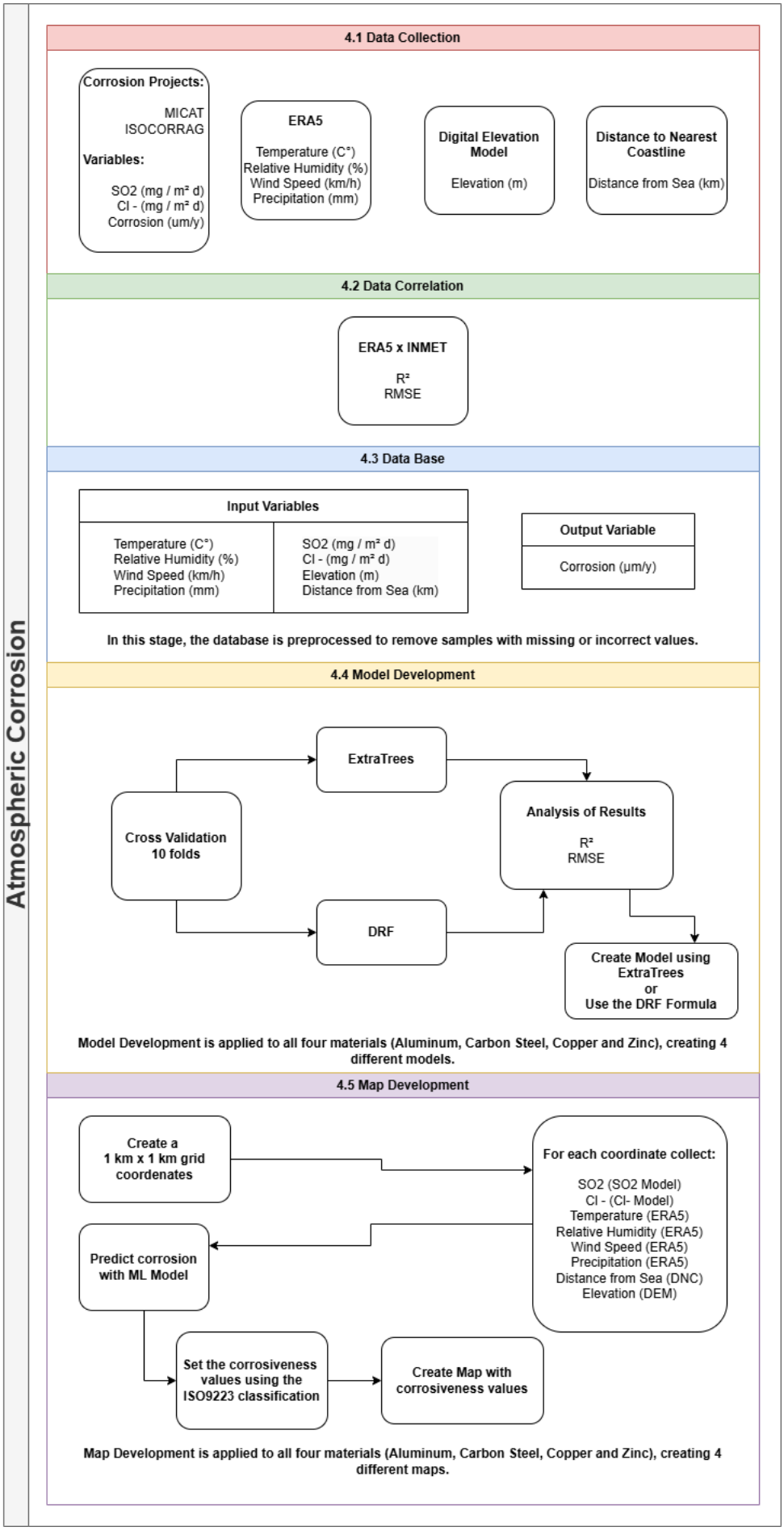

Figure 9 – Research Workflow.

Source: Developed by the author

4.1 DATA COLLECTION

The first stage of development, represented by the red box in Figure 9, consisted of selecting the atmospheric corrosion information and reanalysis data that will compose the atmospheric corrosion Models. Both MICAT and ISOCORRAG are well-established references in the corrosion literature, and their widespread use by other researchers (GAVRYUSHINA; PANCHENKO, 2023; MIKHAILOV; STREKALOV; PANCHENKO, 2007; PINTOS et al., 2000; ALMEIDA; ROSALES, 2000) is an indicator of the reliability and relevance of these databases. By opting for these projects, access to a significant amount of data is ensured, enriching corrosion research. The information for each sample in these projects includes Sulfur Dioxide, Chloride Ions, Carbon Steel Mass Loss, Aluminum Mass Loss, Copper Mass Loss, and Zinc Mass Loss.

In addition to the data collected from the projects, environmental and geographical variables from reanalysis data will be introduced, which, as mentioned in Chapter 2, can influence corrosion values. The ERA5 reanalysis dataset was selected to add information on Temperature, Relative Humidity, Wind Speed, and Precipitation to the samples, which have a resolution of 100km² per point. For the geographic values of Elevation and Distance from the Sea, the Digital Elevation Model and Distance to the Nearest Coastline datasets will be used, both provided by PacIOOS, each with a resolution of 1 km² per point. This entire process can be visualized in Figure 10.

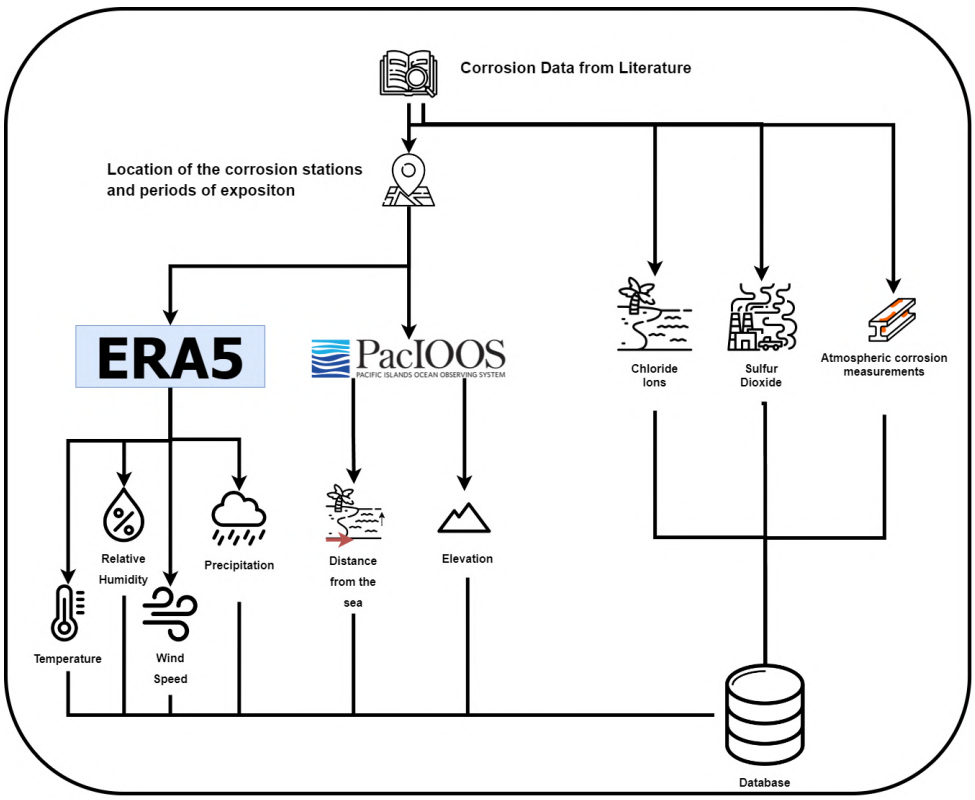

Figure 10 – Data collection process for Atmospheric Corrosion analysis, integrating environmental and corrosion data from various sources into a database.

Source: Developed by the author

4.2 DATA CORRELATION

Unlike other Machine Learning studies that rely on homogeneous datasets, which use data from a single source, our approach aims to enhance results by combining corrosion project data with reanalysis data. Correlation in this context refers to assessing the relationship between data from on-site samples and reanalysis data. This step, represented by the green box in Figure 9, involves verifying if the reanalysis data provides reliable values when compared to samples collected on-site.

A dataset from the National Institute of Meteorology (INMET) of Brazil was used to correlate and verify the quality of temperature, relative humidity, wind speed, and precipitation data. This entity conducts studies at various meteorological stations distributed throughout Brazil, collecting climatic data from different points. All these values can be compared with those obtained through ERA5 to verify their precision using evaluation metrics such as R^2 and RMSE.

4.3 DATA BASE

In this stage, represented by the blue box in Figure 9, preprocessing of the data collected in the Data Collection Section and then correlated in the Data Correlation Section will be conducted. This process is essential for removing samples with missing or incorrect values, such as entries that should be numerical but are not or other anomalies that may compromise data integrity.

In addition to data with incorrect values, a search is conducted for samples that exhibit disproportionate values, commonly classified as outliers. Given the limited number of samples in the database and the fact that each includes its geographical location, it was possible, in collaboration with ArcelorMittal Brazil experts, to identify and eliminate these samples.

After preprocessing, the dataset is divided into four groups, each represented by a distinct metal: Carbon Steel, Aluminum, Copper, and Zinc. This separation is necessary because four Machine Learning Models will be developed to generate four atmospheric corrosion maps.

The final database contains 8 numeric input variables and 1 regression output variable, categorizing the problem as a univariate regression task. These input variables include Temperature, Relative Humidity, Wind Speed, SO_2 deposition, Cl^- Deposition, Elevation, and Distance from the Sea, all of which are known to influence the corrosion process significantly.

Chapter 5 will provide all details of the samples, including the initial data quantity and after preprocessing, as well as the proportionality for each type of metal. The preprocessing stage ensures that the data quality is suitable for training accurate machine learning models, ultimately leading to more reliable atmospheric corrosion predictions.

4.4 MODELS DEVELOPMENT

After structuring the database to be used, the process of training the Machine Learning Models is initiated, as indicated in Figure 9 by the yellow box.

First, the ExtraTrees Algorithm and DRFs are applied to the dataset. A 10-fold cross-validation process is performed simultaneously on both the ExtraTrees and DRFs to evaluate their predictions.

Following cross-validation, the results are analyzed using metrics such as R^2 and RMSE to determine which Model provides better predictions. This analysis helps in comparing the performance of the ExtraTrees and DRFs.

Based on the analysis, the best-performing methodology, either ExtraTrees or DRF, is selected. If the ExtraTrees Model is found to be statistically superior, it is then used to train the final Models on the entire dataset. If not, the DRF formulas are used instead.

The Model development process is repeated for each of the four materials: Aluminum, Carbon Steel, Copper, and Zinc, resulting in the creation of four distinct Atmospheric Corrosion Models.

4.5 MAP DEVELOPMENT

In the stage represented by the purple box in Figure 9, atmospheric corrosion maps are generated for the four materials: Carbon Steel, Aluminum, Copper, and Zinc. These maps have a resolution of 1 km by 1 km or 0.01 degrees of latitude by 0.01 degrees of longitude and are focused on the South American region. An example of a 1 km^2 resolution map can be seen in Figure 11, which shows the elevation at each point in meters in South America.

Due to the low resolution of 1 km^2 and the extensive area of South America, predicting atmospheric corrosion involves processing a large number of points. To manage this, parallelism techniques are utilized. This approach allows the simultaneous process of corrosion at multiple points, with each processor core handling the prediction for individual points.

Before the beginning of map development, a SO_2 deposition model was created. This model is essential as it provides more accurate pollution values than MERRA-2. This improvement is mainly due to MERRA-2's high resolution, which fails to capture the microclimate of the region.

The map development process begins by selecting the boundary coordinates of the maps. Afterward, points within the grid are identified, and input variables from reanalysis data and machine learning models are collected for each point on the map. These variables, which include environmental and atmospheric factors, are input into individual corrosion Models for each material: Carbon Steel, Aluminum, Copper, and Zinc. The output of these Models is the annual rate of Atmospheric Corrosion, and the values are stored. To generate the maps, these stored corrosion rates are converted into corrosiveness categories based on ISO9223 standards, enhancing clarity in visualization. This individualized approach per material is essential due to

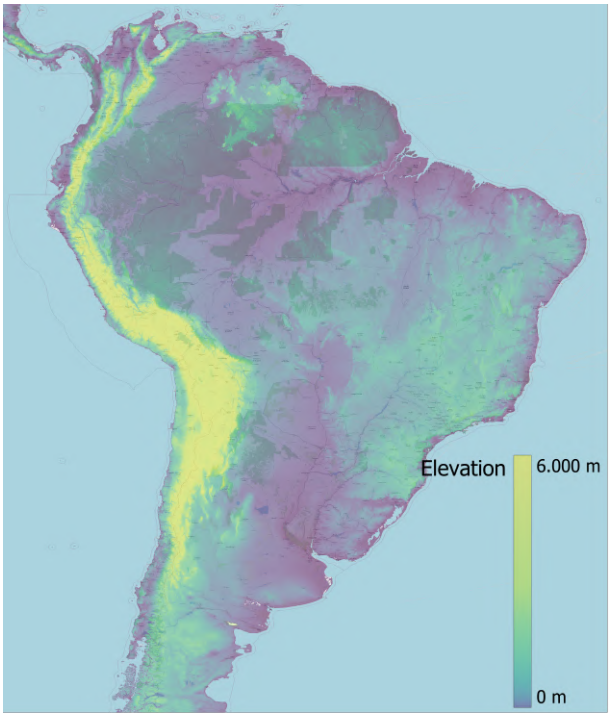

Figure 11 – Elevation map in meters of the South America region with 1 km² resolution

Source: Developed by the author

their unique responses to environmental variables, each necessitating distinct Machine Learning Models. All this process can be visualized in Figure 12, and the machine setup used in the experiments can be found in the next section.

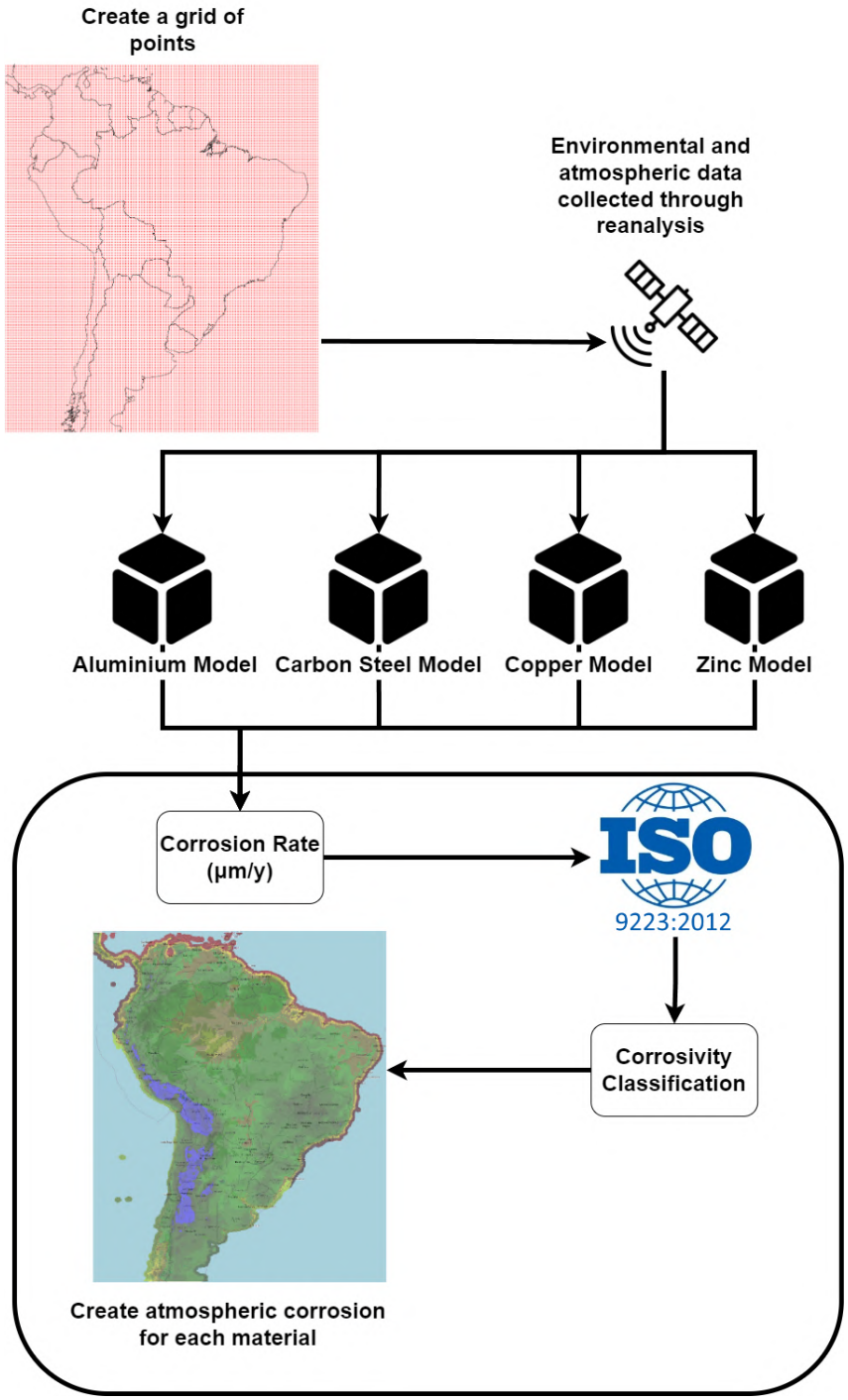

Figure 12 – The workflow for creating atmospheric corrosion maps involves creating a grid of points across South America, collecting environmental and atmospheric data, and feeding the data into material Models (Aluminium, Carbon Steel, Copper, Zinc) to calculate corrosion rates (µm/y), classifying the corrosivity, and generating atmospheric corrosion maps for each material.

Source: Developed by the author

4.6 TOOLS AND LIBRARIES

Python was used for the construction of all algorithms due to its simplicity in reading CSV and .nc files and its various libraries that are useful for building Machine Learning Models. Python can be defined as an interpreted and object-oriented programming language. It features various technologies such as exceptions, native data structures, and access to several native libraries of great utility (ROSSUM; DRAKE, 2011).

In the development of the Models, the Python library scikit-learn [1] was used, which features ExtraTree and Random Forest Algorithm utilized in this research. The Pandas [2] library was used for data organization and preprocessing. This library can read data from an xlsx file and transform it into usable data in the program. The library also includes other auxiliary functions that will help in the development of the work. The Xarray [3] library was also used, which is responsible for structuring and processing reanalysis data, more specifically in the .nc format.

A server from the Laboratory of Research in Computational Intelligence (LABICOM) at UDESC was used to execute the entire development workflow. The machine has the following specifications:

- Processor: Intel(R) Xeon(R) Gold 6330N CPU @ 2.20GHz
- Number of Cores: 80
- Operational System: Ubuntu 22.04.3 LTS
- Principal Memory: 1.5 Tb

[1] scikit-learn: Machine Learning in Python - (PEDREGOSA et al., 2011)
[2] Pandas: Python Data Analysis Library - (PANDAS, 2020)
[3] Xarray: N-D labeled arrays and datasets in Python - (HOYER; HAMMAN, 2017)

5 RESULTS AND ANALYSIS

This chapter will present the development of atmospheric corrosion Models and the results that were obtained.

5.1 DATA COLLECT

To develop atmospheric corrosion Models, training data must be collected. These data can be classified into two categories: Atmospheric Corrosion Projects and Reanalysis and Complementary Data.

5.1.1 Atmospheric Corrosion Projects

Two major atmospheric corrosion projects, MICAT and ISOCORRAG, were developed between 1980 and 1995. These studies collected environmental and corrosivity information from different regions of the globe, which was essential for developing a unified database.

At each corrosion station of the two projects, samples of different materials were exposed to the environment, representing different exposure series.

Figure 13 illustrates the division of each series. The MICAT project was divided into three series of one year of exposure and other series for more than one year. The ISOCORRAG project adopted the same approach, with six series of one year of exposure and other series lasting up to eight years.

The models being developed aim to predict atmospheric corrosion for one year of exposure. Therefore, only the data from the one-year series were used, totaling 512 samples. For each sample, atmospheric corrosion values were collected for Aluminum, Carbon Steel, Copper, and Zinc, in addition to values for the corrosive agents Sulfur Dioxide (SO_2) and Chloride Ions (Cl^-).

All documentation and results from ISOCORRAG can be found in the digital catalog Dean Dagmar Knotkova (2010), in PDF format. For data collection, Excel was used, which allows tables to be transformed into CSV documents (Microsoft Corporation, 2018).

The MICAT dataset (MORCILLO et al., 1998) is available digitally but in an incomplete form. A full printed version can be found in the library of the Polytechnic School at the University of São Paulo. Due to the large number of pages, the book was divided into 2 parts. Part 2 contains the tables of atmospheric corrosion experiments, which were manually extracted. To ensure the accuracy and consistency of the data, two independent collections were conducted.

5.1.2 Reanalysis and Complementary Data

Corrosion results from interaction with the environment; thus, data that describe it are necessary. These data can be obtained from three datasets: ERA5, DEM and DNC.

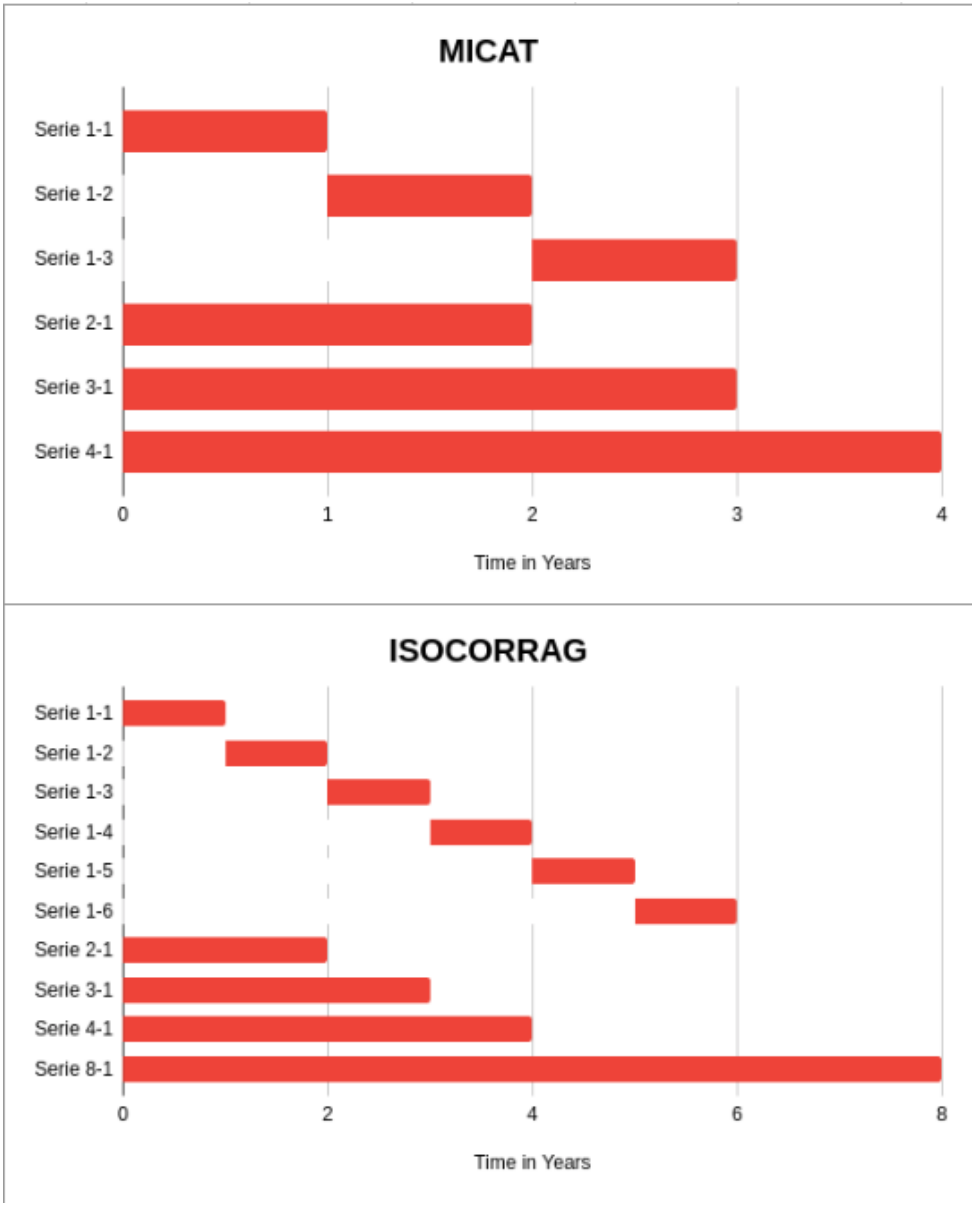

Figure 13 – Comparison of exposure series durations for the MICAT and ISOCORRAG projects. The MICAT series ranges from 1 to 4 years, while the ISOCORRAG series ranges from 1 to 8 years.

Source: Developed by the author

ERA5 provides values for Temperature, Relative Humidity, Wind Speed, and Precipitation with a resolution of 100 km². Since the corrosion stations provide their exact locations, these environmental values can be collected and associated with each station. This combination of information is fundamental for understanding the environment's corrosive behavior.

Additional information on Distance from the Sea and Elevation was collected through the DNC and DEM datasets, as these factors are directly related to the concentration of Cl^-, which is a corrosiveness agent.

Each data point in the final database contains eight pieces of information describing the environment, in addition to the values of mass loss of the materials exposed in these environments.

To ensure the quality of externally provided information, such as reanalysis data, a correlation was performed between reanalysis collections and on-site collections, which can be seen in the next section, Data Correlation.

5.2 DATA CORRELATION

As mentioned in section 5.1 (Data Collect), the corrosion datasets do not contain sufficient information about corrosive agents for a comprehensive understanding of the environment. Therefore, reanalysis data were used to enrich the content of the existing data in corrosion datasets. To ensure the quality of these data, they were compared with field-collected values.

For environmental corrosive agents, such as Temperature, Relative Humidity, Precipitation, and Wind Speed provided by ERA5, the INMET database was used to perform correlations and verify the accuracy of the reanalysis data. INMET annually provides data on these variables from 567 stations spread across the Brazilian territory. Each station records the values of each variable at hourly intervals, resulting in up to 8,760 values per station over a year.

However, some stations may not provide all 8760 annual values due to problems with the measure, and in some cases, entire months may lack recorded data. Since environmental data vary according to the seasons, stations that provided less than 80% of the total data, fewer than 7,008 values, were excluded from the analysis.

All wind speed data collected by INMET were measured at 2 meters above ground, whereas the ERA5 data were recorded at 10 meters, as noted by Siefert et al. (2021). To ensure a consistent comparison, this difference in measurement height needs to be adjusted. The transformation can be done using the formula provided by Allan, Pereira e Smith (1998):

$$u_2 = u_z \frac{4.87}{\ln(67.8z - 5.42)} \tag{9}$$

where:

- u_2: wind speed at 2 meters above ground [m/s],
- u_z: wind speed measured at z meters above ground [m/s],
- z: height of the wind speed measurement above ground [m].

Thus, to convert wind speed data from 10 meters to 2 meters, the values can be multiplied by 0.74.

After data collection, the annual average value of each variable per station was used to make comparisons with the reanalysis data. The RMSE and R^2 metrics were employed for this comparison, as R^2 provides the overall trend of the values, while RMSE indicates the average error, particularly considering large errors. The results can be visualized in Figure 14, where the red line in each of the four charts represents the line of perfect agreement between the observed

INMET data and the ERA5 reanalysis data. This line serves as a theoretical reference, illustrating the ideal scenario in which the ERA5 data would perfectly replicate the observed measurements. The closer the data points are to this red line, the stronger the concordance between the two datasets, indicating a high level of accuracy in the reanalysis data.

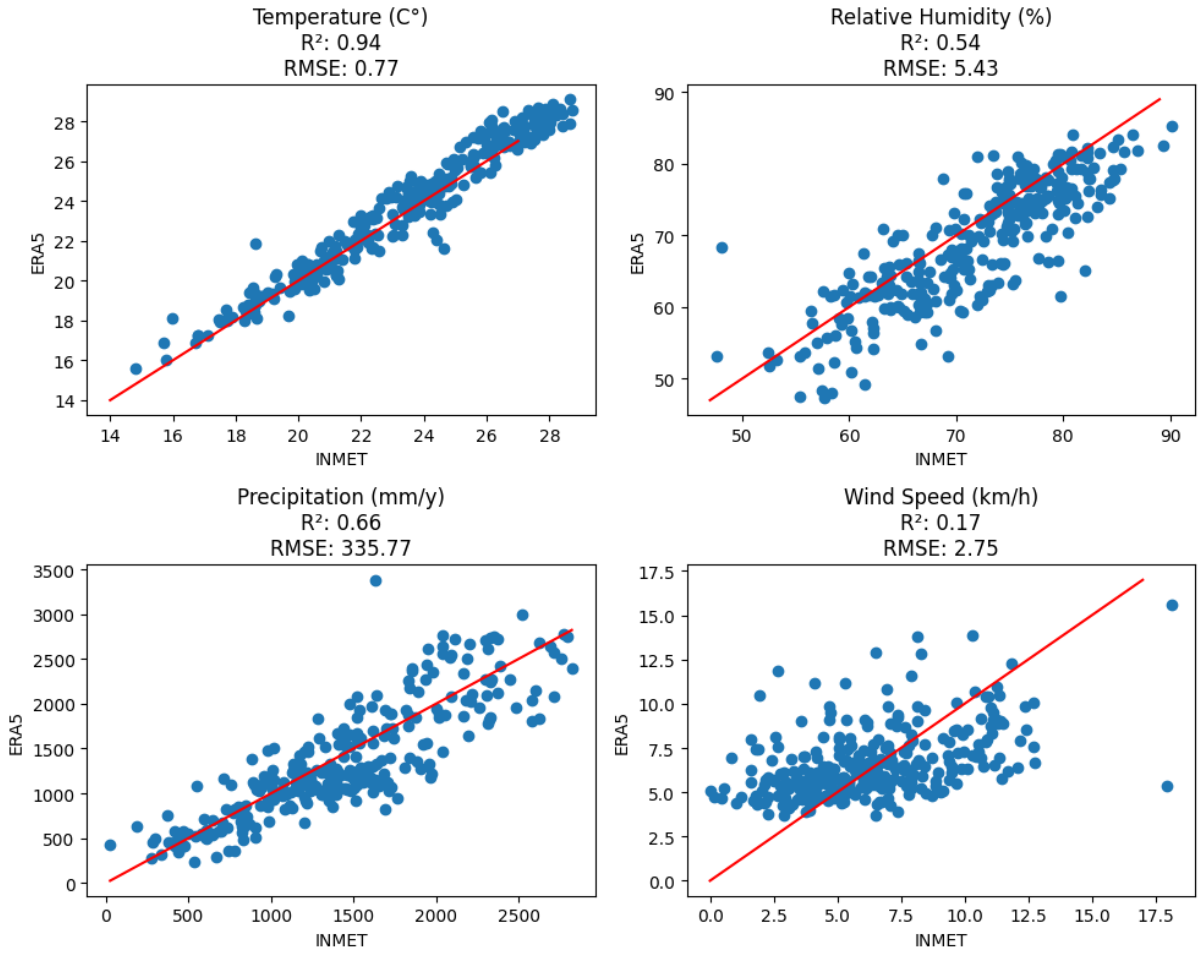

Figure 14 – Graphs comparing INMET values with ERA5 reanalysis data of Temperature, Relative Humidity, Precipitation, and Wind Speed

Source: Developed by the author

As can be observed, the Temperature values from ERA5, when compared with INMET collections, are quite similar, presenting an average error of 0.77 °C. This error is small for the application in question. Relative Humidity and Precipitation show larger errors than Temperature, but these errors are acceptable for future corrosion models, given that precipitation can vary from 0 up to 5000, and Relative Humidity can range from 0 to 100. Wind Speed, on the other hand, exhibited a different pattern. It is noted that ERA5 does not record low Wind Speed values, resulting in a higher error. Another contributing factor to this error is mentioned in the ERA5 documentation itself, which states that Wind Speed is influenced by terrain and time of day. Since the annual average value was used and ERA5 has a resolution of 100 km², the comparisons may be discrepant. Since the largest discrepancy is in values below 5 km/h, and these values represent low wind speeds in any case, ERA5 data can still be used for corrosion models.

5.3 DATA BASE

With the corrosion database collected and the reanalysis data verified through correlations, it was necessary to unify them. For each corrosion data point, corresponding environmental information was associated with each station. Some data series from certain stations did not present Cl^- or SO_2 values and were therefore discarded. All values were reviewed by a group of specialists from ArcelorMittal. In some cases, samples exhibited values that were discrepant with their geographical reality and were also discarded. In total, 1,084 samples were selected to compose the consolidated database that will be used for Model training.

In Table 3, the environmental and atmospheric variables that will be used for training are shown in blue, while the atmospheric corrosion variable, which will be the output of the corrosion Models, is highlighted in red.

Variable	Unit	Source	Resolution
Annual Mean Temperature	°C	ERA5	1 km²
Annual Mean Relative Humidity	%	ERA5	1 km²
Annual Mean Wind Speed	km/h	ERA5	1 km²
Annual Mean Precipitation	mm	ERA5	1 km²
Distance from the Sea	km	DNC	1 km²
Elevation	m	DEM	1 km²
Annual Mean Chloride Ions Deposition	mg/(m²·day)	Corrosion Database	—
Annual Mean Sulfur Dioxide Deposition	mg/(m²·day)	Corrosion Database	—
Atmospheric Corrosion	µm/year	Corrosion Database	—

Table 3 – Environmental and atmospheric variables used for model training, with the atmospheric corrosion variable highlighted in red.

The variables Distance from the Sea and Elevation were transformed to a logarithmic scale, as performed by Geremias et al. (2023), which resulted in improvements in the Models. This transformation proved effective due to the behavior of Cl^-: near the sea and in low elevation areas, the values tend to be higher, while a small increase in Distance from the Sea and Elevation can result in an exponential decrease in these values.

After consolidating the database, the data were separated into four groups according to the material, as four distinct Machine Learning Models will be developed. After separation, each data group comprised the following number of samples:

- Aluminium: 252
- Carbon Steel: 276
- Copper: 272

- Zinc: 284

After preprocessing and grouping the data, it was necessary to verify the correlations between the variables. According to the literature on atmospheric corrosion, corrosive agents tend to corrode the material when they have high values. Therefore, an analysis using Spearman's correlation was performed, and the results can be visualized in Table 4.

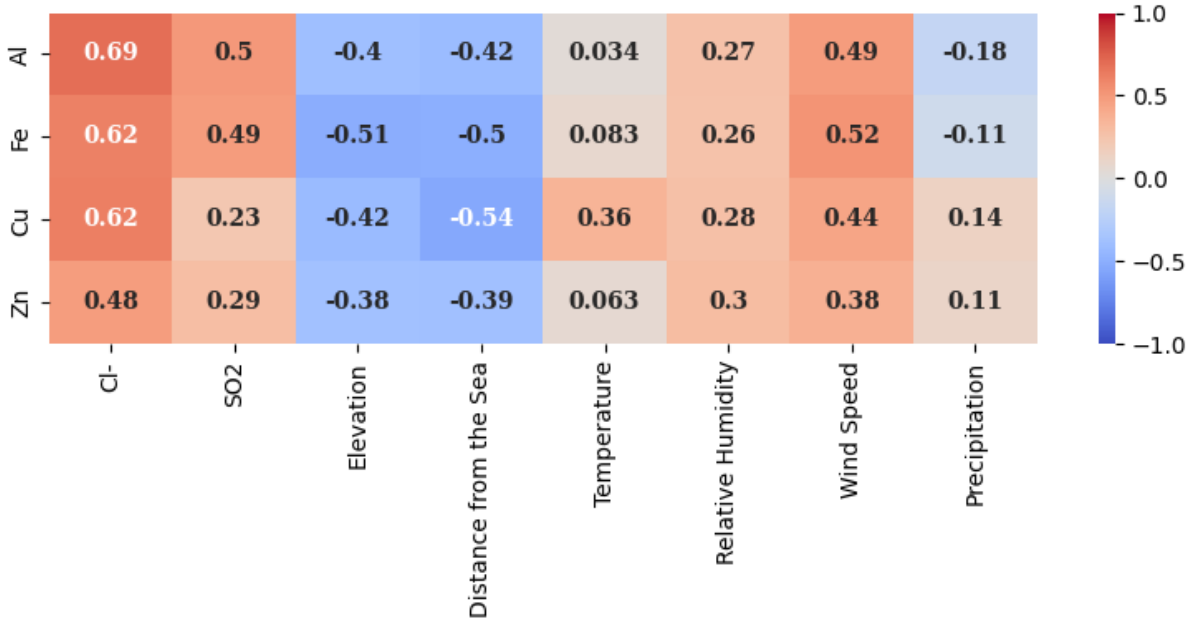

Table 4 – Spearman correlations between atmospheric corrosion values and corrosive agents for the materials: Aluminum (Al), Carbon Steel (Fe), Copper (Cu), and Zinc (Zn)

Source: Developed by the author

As described in Chapter 2, the main corrosive agents are Cl^- and SO_2. Figure 4 illustrates that the increase in these values is directly related to the increase in atmospheric corrosion of the four materials studied: Aluminum, Carbon Steel, Copper, and Zinc. It can be observed that the Spearman correlation between Cl^- and the materials ranges from 0.48 to 0.69, while for SO_2, the correlations range from 0.23 to 0.5, reinforcing how critical these agents are for corrosion.

The variables Distance from the Sea and Elevation have negative correlations with corrosion, indicating that higher values correspond to lower corrosion rates. This aligns with the theory, as regions closer to the coast and with low elevation exhibit higher corrosivity due to the greater concentration of Cl^- in the atmosphere. The correlations for Distance from the Sea range from -0.38 to -0.51, while for Elevation, range from -0.39 to -0.54.

Temperature, Relative Humidity, and Precipitation present less pronounced correlations. The correlations for Temperature range from 0.034 to 0.36, Relative Humidity from 0.26 to 0.52, and Precipitation from -0.18 to 0.14. Although these variables influence the corrosion process, the presence of an acid, such as Cl^- or SO_2, is fundamental for corrosion to occur. Thus, even in conditions of high temperature, humidity, or precipitation, the absence of acids results in less aggressive corrosion.

Lastly, Wind Speed showed a relatively high correlation with the materials, ranging from 0.38 to 0.49. This can be explained by the wind's ability to transport corrosive agents such as Cl^- and SO_2 to the surface of the materials, thereby intensifying the corrosion process.

The analysis of Spearman's Correlation demonstrates how different environmental factors interact and influence the atmospheric corrosion of various materials, highlighting the complexity of the process and the importance of considering multiple variables in the assessment of corrosivity.

5.4 MODELS DEVELOPMENT

With the data consolidated and separated into groups by material, the training of the Models begins. ISO9223 presents the Dose Response Functions (DRFs), which are mathematical formulas for predicting the atmospheric corrosion rate for Aluminum, Carbon Steel, Copper, and Zinc. During this project, the article Geremias et al. (2023) was published, demonstrating how Machine Learning Models outperform the DRFs. In that article, the variables Wind Speed, Precipitation, Elevation, and Distance from the Sea were not used. By incorporating these new variables, new Models (ML) were developed.

For the development of the new Models, the ExtraTrees Algorithm was used, with the following hyperparameter configurations:

- Maximum Depth: 15
- Number of Trees: 50

These configurations were determined empirically based on comprehensive testing and analysis.

A 10-fold Cross-Validation method was employed for training. This approach was applied during both the model development and the evaluation of the DRFs, with the results illustrated in Figure 15. The Cross-Validation was performed 50 times, and the mean and standard deviation of the results were recorded. The models were evaluated using the RMSE and R^2 metrics.

The results showed that the Machine Learning Models outperformed the DRFs in all analyzed cases. This is due to the ability of Machine Learning Models to handle a larger number of input variables and capture more complex relationships between them. The DRFs are limited because they use only four input variables: Temperature, Relative Humidity, SO_2, and Cl^-. As a result, they cannot capture the influence of other important climatic and geographical variables, such as Wind Speed, Precipitation, Elevation, and Distance from the Sea. This limitation prevents the DRFs from providing predictions as accurate as the Machine Learning Models, which incorporate a more comprehensive analysis of the corrosive environment.

SHAP plots were used to better understand the importance of each variable in the Model. These plots helped identify the variables that most influence the predictions, providing interpretability of the individual contributions of each variable in the Machine Learning Models. The plots can be visualized in Figure 16.

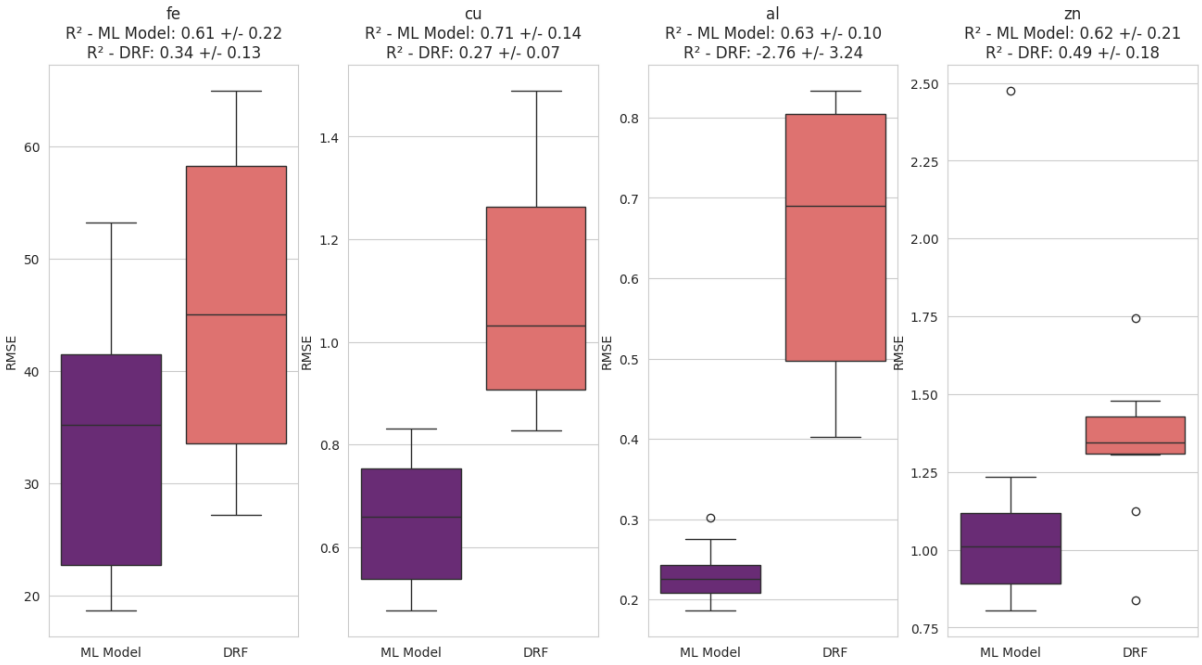

Figure 15 – Cross-Validation results for 10 folds comparing the Machine Learning Models and the DRFs. Results are presented in terms of mean/standard deviation and boxplots showing distributions of data values

Source: Developed by the author

For all materials, the main corrosive agents, SO_2 and Cl^-, consistently appear at the top, indicating that they have a significant impact on corrosion predictions. This confirms the well-documented importance of these agents in atmospheric corrosion. The presence of high SHAP values for SO_2 and Cl^- indicates that these variables are crucial for determining the corrosion rate, reinforcing the need to monitor and control these agents in environments susceptible to corrosion.

Another important observation is the influence of Temperature, which has a significant impact on the corrosion of Copper but is less relevant for other materials. This suggests that higher temperatures may accelerate the corrosion of Copper. In contrast, temperature variations appear to have a more limited effect on materials such as aluminum, iron, and zinc, indicating that other environmental factors play a more dominant role in their corrosion processes. This underscores how different features interact uniquely with each material, reinforcing the importance of material-specific corrosion models.

Additionally, Distance from the Sea also appears as an important variable in all the plots, especially for Copper and Aluminum. This can be attributed to the influence of marine air salinity, which increases the presence of Cl^-, a known factor that significantly contributes to corrosion.

Wind Speed also plays a relevant role, particularly in materials such as Aluminum and Zinc. Higher wind speeds can enhance the deposition of airborne pollutants, including Cl^- and SO_2, accelerating corrosion in exposed environments.

On the other hand, variables such as Precipitation and Relative Humidity are consistently

positioned lower in the SHAP plots for all materials. This suggests that, while these variables still impact corrosion, they are less influential compared to the main corrosive agents and Distance from the Sea. The lesser influence of Precipitation and Relative Humidity may indicate that their effects are more indirect or dependent on interactions with other variables.

The development of new Machine Learning Models incorporating additional variables resulted in more precise predictions of atmospheric corrosion rates. The analyses demonstrated the superiority of these Models over traditional DRFs, highlighting the importance of considering a broader range of environmental variables to achieve a more comprehensive and detailed understanding of the factors influencing corrosion.

The Cross-Validation method, although useful for evaluating the performance of the Models, does not generate final usable Models, as its approach is focused on verifying how the Models perform with different datasets. Through the experiments, it was possible to observe that the Models produced good results with this dataset. Subsequently, the Models were retrained for each material using the entire available database. This approach was necessary due to the limited number of samples in the database, making it unfeasible to discard any data. This is why Cross-Validation was utilized.

The new Models, trained using all available data, are now ready for use. The next Section, Map Development, will demonstrate how these Models will be applied and utilized in practice.

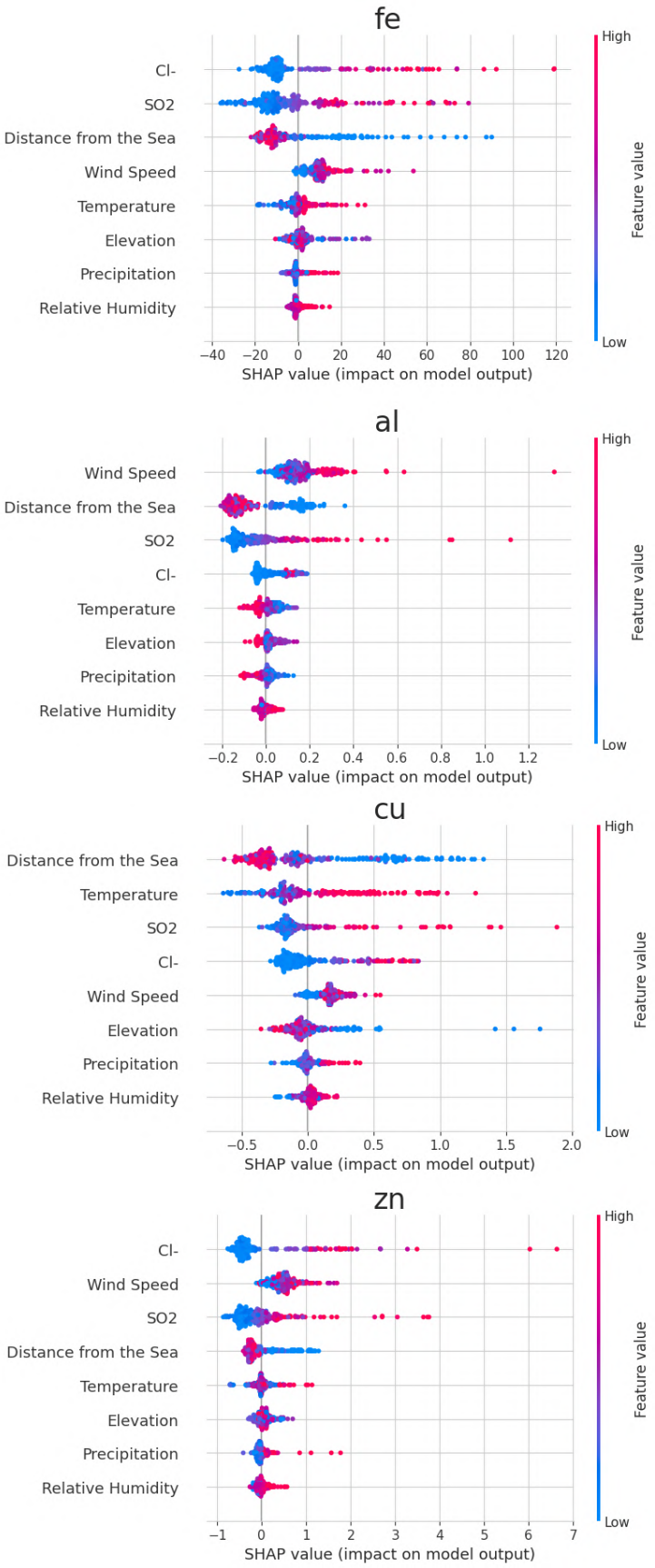

Figure 16 – SHAP plots showing the importance of variables for corrosion predictions in the materials Carbon Steel (Fe), Aluminum (Al), Copper (Cu), and Zinc (Zn).

Source: Developed by the author

5.5 MAP DEVELOPMENT AND AUXILIARY MODELS

Four atmospheric corrosion Models have been developed and are ready for use in the project. One of the main objectives of this project is to create Atmospheric Corrosion Maps for South America, using these Models to represent the corrosion of materials in different regions.

Initially, it was necessary to define the projection quadrant for the maps, which was delimited as follows:

- Latitude: -48° to 13°
- Longitude: -85° to -33°

With the grid of points defined, it was necessary to establish the resolution of the maps. The chosen resolution was 1 km², meaning each point on the map represents an area of 1 km by 1 km. With this definition, each corrosion map will have a total of 31,833,100 points.

Due to the quadrant's location, many points are situated in the ocean and do not need to be considered for corrosion calculations. Therefore, only points located on land and up to 50 km offshore will be included in the calculations. The choice to include points up to 50 km offshore is based on the adopted resolution. Without this approach, points very close to the seashore or located on small islands would not be calculated.

It was necessary to define the reference year for the Corrosion Maps, considering that some corrosive agents vary from year to year. It was decided that the reference year will be 2023, and the collection of all corrosive agents to be used as input values for the prediction Models will follow this year.

With the points defined, the process of capturing corrosive agents for each point began. For environmental data such as Temperature, Wind Speed, Relative Humidity, and Precipitation, the ERA5 reanalysis dataset was used, which has a resolution of 100 km². For Distance from the Sea, the DNC dataset was used, and for Elevation, the DEM dataset was utilized. Both datasets have a resolution of 1 km² and are timeless, as these two variables undergo few changes over the years.

For the atmospheric variables Cl^- and SO_2, two distinct models were used. It's possible to use data from the MERRA-2 dataset, but this dataset has a high resolution, and the values can have a significant error.

For Cl^- values, the model from Brandenburg (2024) was applied. The model, based on a Random Forest algorithm, incorporates inputs such as wind speed, distance from the sea, roughness, elevation, Cl^- concentration from MERRA-2, temperature, and relative humidity. These variables can be obtained using datasets like ERA5, MERRA-2, DEM, DNC, and LandCover. The model's output is Cl^- deposition. This approach offers better results compared to other Cl^- deposition estimation methods and is well-suited for use in grid mapping applications.

For SO_2 values, a new model was developed for this proposed, which utilizes values of SO_2 concentration and levels of urbanization.

5.5.1 SO_2 Model

MERRA-2 provides SO_2 concentration values that can be converted into deposition estimates. However, the high resolution of MERRA-2, at 50 km x 70 km, makes these values less accurate for understanding the microclimate of a specific region, particularly in industrial areas.

By employing a Random Forest model, along with MERRA-2 SO_2 concentration data and the LandCover dataset, it is possible to enhance the accuracy of SO_2 deposition estimates.

5.5.1.1 Data Base

The output for this model is SO_2 deposition. Therefore, the same dataset used for corrosion models was applied here, but with the addition of satellite data corresponding to the coordinates and SO_2 deposition values as the model output.

For the input variables, MERRA-2 SO_2 concentration was utilized. Despite the differing units, it is possible to convert concentrations into deposition values, as described in ISO 9223. The standard indicates that a constant deposition rate for SO_2 can be used, and it provides the following conversion formula:

$$P_d = 0.8 \cdot P_c$$

P_d is the Deposition value of SO_2 in mg/(m²·day) and P_c is the Concentration value of SO_2 in µg/m³.

In addition to the converted concentration data, LandCover data, particularly urbanization data, was incorporated. Given that LandCover has a resolution of 0.3 km x 0.3 km, it is possible to assess how much of a coordinate falls within an urbanized area by counting the number of urban area points surrounding that coordinate. This is a crucial factor, as higher SO_2 concentrations tend to occur in more urbanized areas.

Four different urbanization radiuses were selected as input variables, with the following values:

- Count of urbanization points (CUP) within a 1.2 km radius.
- Count of urbanization points (CUP) within a 12 km radius.
- Count of urbanization points (CUP) within a 30 km radius.
- Count of urbanization points (CUP) within a 90 km radius.

These radiuses were chosen to capture varying degrees of urbanization, allowing the model to account for different environments such as small towns, rural areas, large metropolitan regions, or industrial sites located in more remote areas.

Table 5 presents the variables used to train the model. The input variables are highlighted in blue, while the SO_2 deposition variable, which serves as the model's output, is highlighted in red.

Variable	Unit	Source	Resolution
Annual Mean Sulfur Dioxide Deposition	mg/(m²·day)	MERRA-2	1 km²
CUP within a 1.2 km radius	—	LandCover	0.09 km²
CUP within a 12 km radius	—	LandCover	0.09 km²
CUP within a 30 km radius	—	LandCover	0.09 km²
CUP within a 90 km radius	—	LandCover	0.09 km²
Annual Mean Sulfur Dioxide Deposition	mg/(m²·day)	Corrosion Database	—

Table 5 – Count of urbanization points (CUP) from LandCover and MERRA-2 variables used for model training, with the SO_2 depostion variable highlighted in red.

5.5.1.2 Model Development

For the development of the SO_2 Model, the RandomForest Algorithm was used, with the following hyperparameter configurations:

- Maximum Depth: 15
- Number of Trees: 50

These configurations were determined empirically based on comprehensive testing and analysis.

A 10-fold Cross-Validation method was employed for training, and the results are illustrated in Figure 17. The Cross-Validation was performed 50 times, and in each iteration, the mean and standard deviation of the results were computed across the 10 folds. The models were evaluated using the RMSE and R^2 metrics.

As shown, the model provides good performance, with an R^2 score of 0.85 and errors around 7.43, which is an acceptable margin for SO_2 values. With this model, it is possible to generate pollution estimates on a grid map or use it as input for atmospheric corrosion models.

Figure 18 illustrates the comparison between the deposition values from MERRA-2 with a resolution of 3,500 km², which were converted from concentration, and the predicted values using the SO_2 model with a resolution of 100 km². The model-generated map reveals varying levels of pollution across the region, capturing more spatial detail compared to the MERRA-2 map. In the MERRA-2 representation, only the values for major cities such as São Paulo, Rio de Janeiro, and Buenos Aires are distinguishable, as their high population density leads to increased emissions.

The SO_2 model provides more realistic values. According to ISO 9223, values between 4 and 24 mg/(m²·day) are classified as urban areas, while values exceeding 24 mg/(m²·day) are classified as industrial areas. If only MERRA-2 values were used, all of South America would be classified as rural, rendering these values unsuitable for the atmospheric corrosion model. This outcome is in clear contrast to the results obtained from the SO_2 model, which can differentiate between various areas and deliver more accurate results.

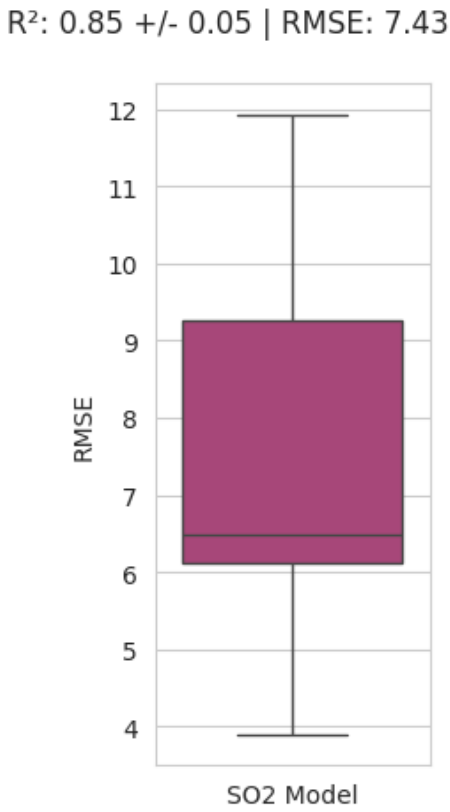

Figure 17 – Cross-Validation results for 10 folds for the SO_2 Model

Source: Developed by the author

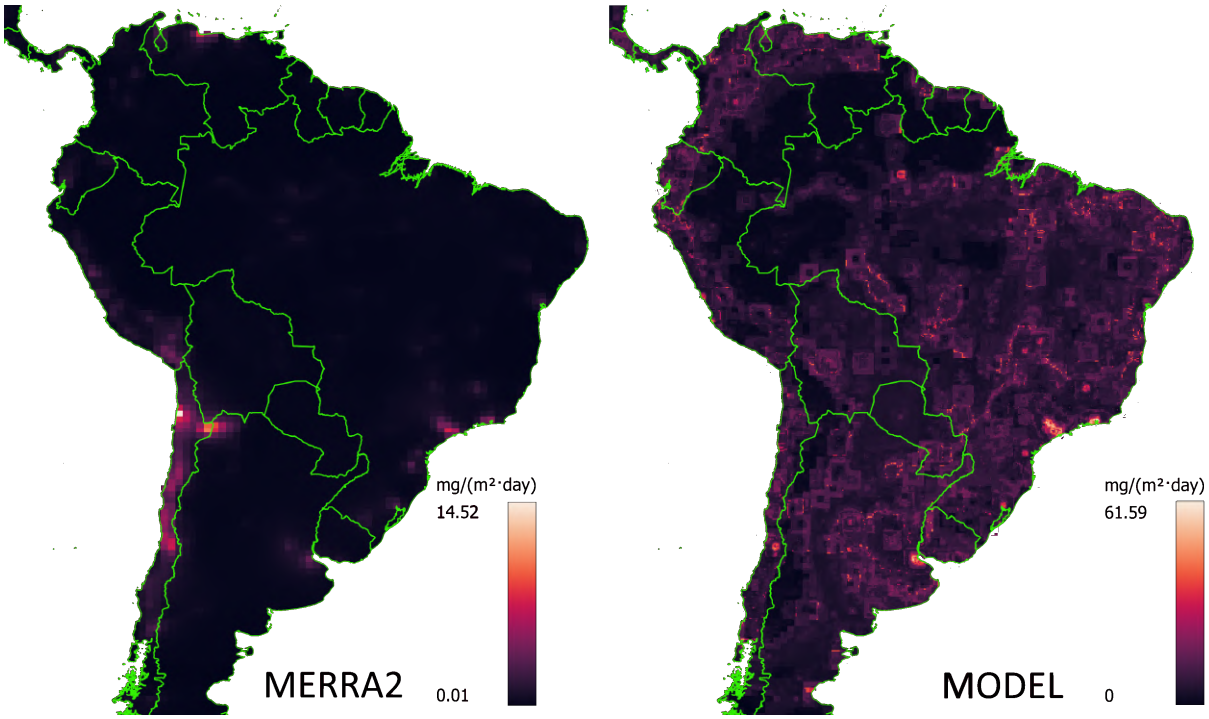

Figure 18 – SO_2 deposition maps. The left map, with a 3,500 km² of resolution, was created using MERRA-2 data, and the right map, with a 100 km² of resolution, was created using the SO_2 Model

Source: Developed by the author

5.5.2 Map Development

The results showed that the Machine Learning Models outperformed the DRFs in all analyzed cases. This is due to the ability of Machine Learning Models to handle a larger number of input variables and capture more complex relationships between them. The DRFs are limited because they use only four input variables: Temperature, Relative Humidity, SO_2, and Cl^-. As a result, they cannot capture the influence of other important climatic and geographical variables, such as Wind Speed, Precipitation, Elevation, and Distance from the Sea. This limitation prevents the DRFs from providing predictions as accurate as the Machine Learning Models, which incorporate a more comprehensive analysis of the corrosive environment.

With the corrosive agents defined, the process of predicting atmospheric corrosion for each point began. Utilizing the available 80 core processor architecture, each point was calculated in parallel to reduce processing time. In addition to parallel processing, the atmospheric corrosion value for the four different materials is calculated for each point, as the only difference in prediction is the Model being used. This approach saves time, as it eliminates the need to iterate each point separately for each material. The prediction flow for the points can be visualized in Figure 19.

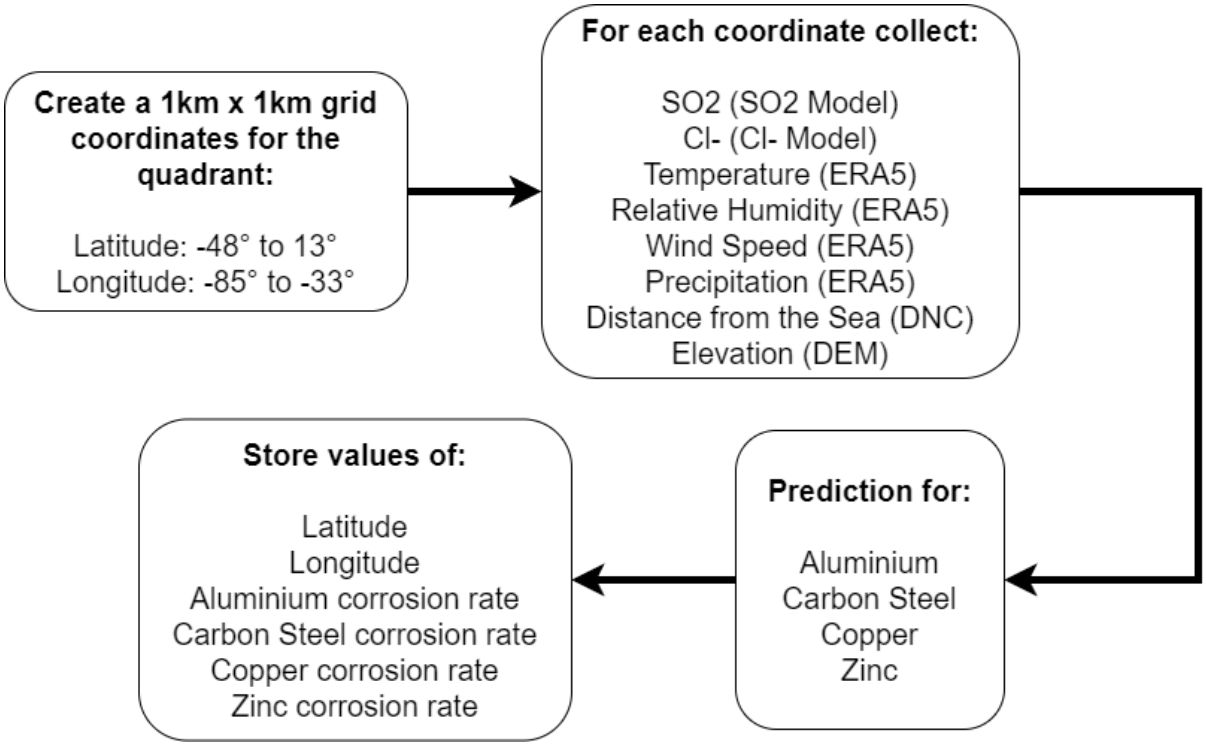

Figure 19 – Prediction flow of the points to generate atmospheric corrosion maps.

Source: Developed by the author

Predicting the points took 7 hours and consumed 250 MB of storage per map, totaling 1 GB. 79 cores were used for parallel processing, with one core reserved for the operating system to avoid server crashes. To optimize performance, all datasets of corrosive agents were loaded into memory, with peak principal memory usage reaching 128 GB.

After processing and storing each point, it is necessary to visualize them in a geographic map format. For better visualization, the corrosion values were classified into corrosivity categories using the conversion table from ISO 9223. The Table 20 defines six corrosion categories, ranging from C1 to CX. Category C1 indicates low corrosivity, while CX indicates extreme corrosivity.

Corrosivity category	Corrosion rates of metals r_{corr}				
	Unit	Carbon steel	Zinc	Copper	Aluminium
C1	g/(m²·a)	$r_{corr} \leq 10$	$r_{corr} \leq 0,7$	$r_{corr} \leq 0,9$	negligible
	µm/a	$r_{corr} \leq 1,3$	$r_{corr} \leq 0,1$	$r_{corr} \leq 0,1$	—
C2	g/(m²·a)	$10 < r_{corr} \leq 200$	$0,7 < r_{corr} \leq 5$	$0,9 < r_{corr} \leq 5$	$r_{corr} \leq 0,6$
	µm/a	$1,3 < r_{corr} \leq 25$	$0,1 < r_{corr} \leq 0,7$	$0,1 < r_{corr} \leq 0,6$	—
C3	g/(m²·a)	$200 < r_{corr} \leq 400$	$5 < r_{corr} \leq 15$	$5 < r_{corr} \leq 12$	$0,6 < r_{corr} \leq 2$
	µm/a	$25 < r_{corr} \leq 50$	$0,7 < r_{corr} \leq 2,1$	$0,6 < r_{corr} \leq 1,3$	—
C4	g(m²·a)	$400 < r_{corr} \leq 650$	$15 < r_{corr} \leq 30$	$12 < r_{corr} \leq 25$	$2 < r_{corr} \leq 5$
	µm/a	$50 < r_{corr} \leq 80$	$2,1 < r_{corr} \leq 4,2$	$1,3 < r_{corr} \leq 2,8$	—
C5	g/(m²·a)	$650 < r_{corr} \leq 1\,500$	$30 < r_{corr} \leq 60$	$25 < r_{corr} \leq 50$	$5 < r_{corr} \leq 10$
	µm/a	$80 < r_{corr} \leq 200$	$4,2 < r_{corr} \leq 8,4$	$2,8 < r_{corr} \leq 5,6$	—
CX	g/(m²·a)	$1\,500 < r_{corr} \leq 5\,500$	$60 < r_{corr} \leq 180$	$50 < r_{corr} \leq 90$	$r_{corr} > 10$
	µm/a	$200 < r_{corr} \leq 700$	$8,4 < r_{corr} \leq 25$	$5,6 < r_{corr} \leq 10$	—

Figure 20 – Conversion Table of Corrosion to Corrosiveness for Aluminum, Carbon Steel, Copper, and Zinc. The "a" in the corrosion units represents year.

Source: Source: (9223, 2012)

With all values categorized, Atmospheric Corrosion Maps were generated for each material, visualized in Figures 21 and 22. The atmospheric corrosion Models make predictions in µm/year, a unit of measure for atmospheric corrosion as indicated by ISO 9223. However, when converting Aluminum corrosion to corrosivity, the values are not expressed in µm/year, but in g/(m²·year), another unit of measure for atmospheric corrosion. To use the corrosion classifications, the atmospheric corrosion values for Aluminum were transformed from µm/year to g/(m²·year), as described in (COVINO, 2005), considering the density of Aluminum of 2.7 g/cm³.

Analyzing the Atmospheric Corrosion Maps, it is possible to observe the significant influence of each variable in the corrosion process. All materials tend to exhibit more severe corrosion in coastal areas due to the high levels of Cl^- and proximity to the sea. Another predominant factor for all materials is the low corrosivity in the Andes region, characterized by high Elevation, low Temperatures, and low Relative Humidity, resulting in a less aggressive corrosion process. These factors are well-grounded in theory, demonstrating how the Models can provide values consistent with reality.

The corrosive agents act differently on each material, as observed in the maps. Figure 21 presents the Corrosion Map for Aluminum, where the only region with C5 corrosivity is located

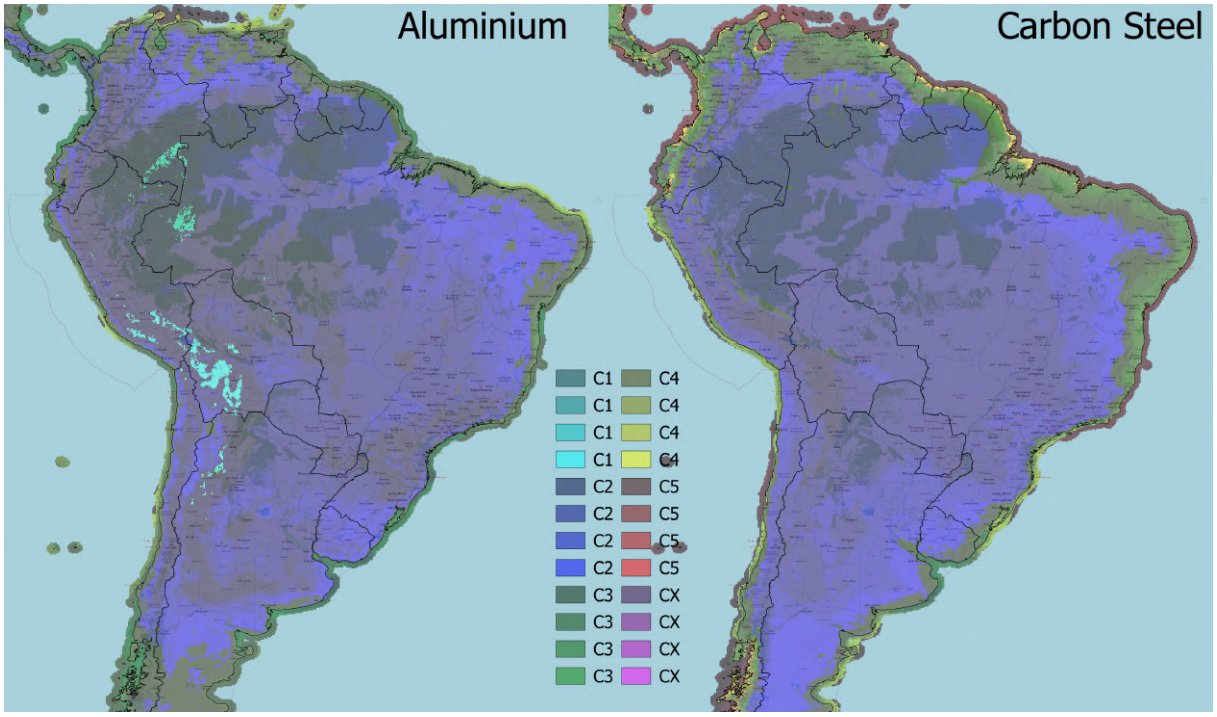

Figure 21 – Atmospheric Corrosion Map for Aluminum and Carbon Steel, with colors indicating the corrosiveness of each area.

Source: Developed by the author

in the northern part of South America, an area with high concentrations of SO_2 and strong winds.

The same Figure 21 shows the Atmospheric Corrosion Map for Carbon Steel, which has distinct properties compared to Aluminum. The continental part is classified as C2, while the coastal region exhibits a gradient of corrosivity, ranging from C3 to C5. The C5 classification for the coastal areas is predominant in the northern region of South America, which can be explained by the temperature difference, as warmer regions tend to have higher corrosivity.

Figure 22 presents the Atmospheric Corrosion Maps for Copper and Zinc. In both maps, the continental corrosivity is classified as C3, while in the Andean regions, it is classified as C2. One factor that stands out more than in the other maps is relative humidity. In the Amazon region, which has high temperatures and an extremely humid environment, the corrosivity is classified as C4, a behavior not observed in the previous maps.

Observing the SHAP plot in Figure 16, the variable Wind Speed has a high impact on the Model's decision for Zinc, whereas it is not as significant for the Copper Model. This impact is evident in the Zinc map, with the extreme northern region of the continent showing C5 corrosivity, corresponding to high Wind Speed values. According to SHAP, the main influencing factors in the Copper Model are Distance from the Sea and Temperature, which are visible by the C5 category in coastal regions with higher Temperatures.

The analysis of atmospheric corrosion maps in South America for different metals reveals distinct corrosion patterns influenced by environmental and geographical factors. Coastal and humid regions are particularly susceptible to high corrosion rates due to the combination of sea

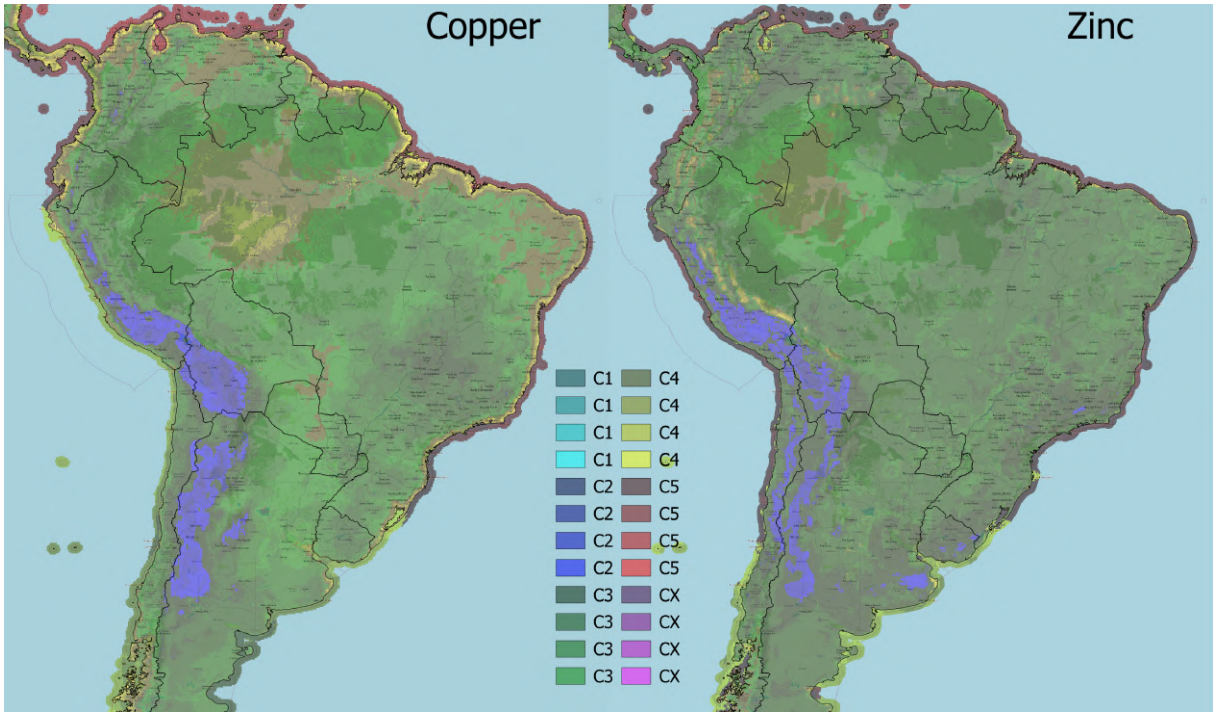

Figure 22 – Atmospheric Corrosion Map for Copper and Zinc, with colors indicating the corrosiveness of each area.

Source: Developed by the author

salt, high humidity, and intense precipitation. Each metal exhibits varying resistance to corrosion, with Aluminum proving to be relatively more resistant compared to Copper, Zinc, and Carbon Steel.

Among all classifications, the CX category does not appear frequently, being observed only in a few points. The maps have a resolution of 1 km², which allows for the presentation of a region's macro corrosivity but does not capture specific nuances. As ISO 9223 itself explains, CX corrosivity is specific to extreme regions, such as measurements near an industrial chimney or locations with extremely high salt content.

6 CONCLUSIONS AND FUTURE RESEARCH DIRECTION

Atmospheric corrosion is a phenomenon of metal degradation that occurs due to exposure to environmental factors such as humidity, atmospheric pollutants, and varying climatic conditions. This process, in addition to impacting the durability and safety of metal structures, represents a significant economic challenge, generating billions in maintenance and material replacement costs. The importance of predicting and mitigating the effects of atmospheric corrosion becomes evident when considering the substantial financial impact it causes in various industries, including construction, transportation, and infrastructure.

This study developed Machine Learning Models to predict the atmospheric corrosion rate of four materials: aluminum, carbon steel, copper, and zinc. These Models utilized reanalysis data, encompassing variables such as temperature, relative humidity, sulfur dioxide deposition, chloride ions, precipitation, and wind speed, as well as geographical variables like distance from the sea and elevation. The development of these Models allowed for the creation of atmospheric corrosion maps specific to the South American region, as well as the use of the Model for studies in specific locations.

In the development of the Models, the ExtraTrees Algorithm was used. The methodology involved collecting and processing environmental data along with field data on atmospheric corrosion, followed by training the Machine Learning Models. To verify the accuracy of the Models, the prediction results were compared with DRFs. The Machine Learning Models outperformed the DRFs, demonstrating greater accuracy in corrosion predictions when validated against real-world data.

The corrosion maps generated from the Models provide a detailed visualization of the areas most affected by atmospheric corrosion. These maps can supply engineers and industry managers with valuable information, enabling them to plan new construction projects more efficiently. As a result, it is possible to significantly reduce operational costs and increase the durability of the structures.

In addition to its practical applications, this study contributes to the state of the art by demonstrating the effectiveness of Machine Learning in predicting atmospheric corrosion. Integrating reanalysis data with atmospheric corrosion data to train Machine Learning Models opens new possibilities for future research.

The results highlight the importance of considering multiple environmental variables in the development of atmospheric corrosion Models. Variables such as wind speed and precipitation, which are not included in the DRFs, have proven essential for increasing the precision of Machine Learning Models. This study demonstrated that including these variables allows the Models to better understand the environment, resulting in more reliable predictions.

6.1 FUTURE WORKS

In the future, the primary goal is to enhance the accuracy of the models by incorporating additional local data, such as proximity to industrial areas, to improve the estimation of SO_2 and Cl^-, which are currently predicted using Machine Learning models based on high-resolution reanalysis data but still lack precision. Another key advancement in this project is applying these improved models to case studies that are more relevant to Brazil, ensuring a more contextually appropriate validation of the study's results. Once the models achieve greater accuracy, efforts will be directed toward increasing the resolution of the corrosion map to 0.09 km², providing more detailed information for specialists. Additionally, it is crucial to analyze how corrosion models can interpret an environment's microclimate, allowing for more precise predictions in the presence of local corrosive agents.

6.2 SCIENTIFIC CONTRIBUTIONS

During this project, a national article was published. In "Prediction of Atmospheric Corrosion in Metallic Materials Using Machine Learning" (GEREMIAS et al., 2023), it is demonstrated how Machine Learning Models can deliver better results compared to DRFs for predicting the atmospheric corrosion rate. Additionally, an international article was written explaining the application of the SO_2 model described in this master's thesis. This article is currently in the publication process.

BIBLIOGRAPHY

9223, I. *Corrosion of metals and alloys - Corrosivity of atmospheres -Classification, determination and estimation*. 2012. Cited 4 times in pages 15, 20, 22, and 62.

9224, I. *Corrosion of metals and alloys — Corrosivity of atmospheres — Guiding values for the corrosivity categories*. 2012. Cited on page 20.

9225, I. *Corrosion of metals and alloys — Corrosivity of atmospheres — Measurement of environmental parameters affecting corrosivity of atmospheres*. 2012. Cited 2 times in pages 20 and 21.

9226, I. *Corrosion of metals and alloys — Corrosivity of atmospheres — Determination of corrosion rate of standard specimens for the evaluation of corrosivity*. 2012. Cited on page 20.

ALJAMAAN, Hamoud; ALAZBA, Amal. Software defect prediction using tree-based ensembles. In: **Proceedings of the 16th ACM International Conference on Predictive Models and Data Analytics in Software Engineering**. New York, NY, USA: Association for Computing Machinery, 2020. (PROMISE 2020), p. 1–10. ISBN 9781450381277. Disponível em: <https://doi.org/10.1145/3416508.3417114>. Cited on page 30.

ALLAN, Richard; PEREIRA, L.; SMITH, Martin. **Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56**. [S.l.: s.n.], 1998. v. 56. Cited on page 49.

ALMEIDA, M. Morcillo E.; ROSALES, B. Atmospheric corrosion of zinc part 2: Marine atmospheres. **British Corrosion Journal**, Taylor & Francis, v. 35, n. 4, p. 289–296, 2000. Disponível em: <https://doi.org/10.1179/000705900101501362>. Cited on page 41.

ALMEIDA, Neusvaldo Lira de; PANOSSIAN, Zehbour. **Corrosão Atmosférica 17 anos**. São Paulo: IPT, 1999. Cited on page 21.

ASIF, Daniyal et al. Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization. **Algorithms**, v. 16, n. 6, 2023. ISSN 1999-4893. Disponível em: <https://www.mdpi.com/1999-4893/16/6/308>. Cited on page 29.

BERRAR, Daniel. Cross-validation. In: RANGANATHAN, Shoba et al. (Ed.). **Encyclopedia of Bioinformatics and Computational Biology**. Oxford: Academic Press, 2019. p. 542–545. ISBN 978-0-12-811432-2. Disponível em: <https://www.sciencedirect.com/science/article/pii/B978012809633820349X>. Cited on page 31.

BRANDENBURG, Thiago. **Aplicação de dados de reanálise climática e aprendizado de máquina para predição de deposição seca de íons de cloreto**. 2024. Trabalho de Conclusão de Curso (Graduação) - Universidade do Estado de Santa Catarina, Centro de Ciências Tecnológicas, Curso de Ciência da Computação, Joinville, 2024. Cited on page 57.

BREIMAN, Leo. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, Oct 2001. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Cited on page 30.

BUI, Xuan-Nam; NGUYEN, Hoang; SOUKHANOUVONG, Phonepaserth. Extra trees ensemble: A machine learning model for predicting blast-induced ground vibration based on the bagging and sibling of random forest algorithm. In: VERMA, Amit Kumar et al. (Ed.). **Proceedings of Geotechnical Challenges in Mining, Tunneling and Underground**

Infrastructures. Singapore: Springer Nature Singapore, 2022. p. 643–652. ISBN 978-981-16-9770-8. Cited 2 times in pages 29 and 30.

CAI, Yikun et al. Atmospheric corrosion prediction: a review. **Corrosion Reviews**, v. 38, n. 4, p. 299–321, 2020. Disponível em: <https://doi.org/10.1515/corrrev-2019-0100>. Cited on page 20.

CHAE, Hobyung et al. Evaluation of supercritical carbon dioxide corrosion by high temperature oxidation experiments and machine learning models. **Metallurgical and Materials Transactions A**, v. 53, n. 7, p. 2614–2626, Jul 2022. ISSN 1543-1940. Disponível em: <https://doi.org/10.1007/s11661-022-06691-5>. Cited on page 36.

COELHO, Leonardo Bertolucci et al. Reviewing machine learning of corrosion prediction in a data-oriented perspective. **npj Materials Degradation**, v. 6, n. 1, p. 8, Jan 2022. ISSN 2397-2106. Disponível em: <https://doi.org/10.1038/s41529-022-00218-4>. Cited on page 34.

COLE, I. S. et al. A study of the wetting of metal surfaces in order to understand the processes controlling atmospheric corrosion. **Journal of The Electrochemical Society**, The Electrochemical Society, Inc., v. 151, n. 12, p. B627, oct 2004. Disponível em: <https://dx.doi.org/10.1149/1.1809596>. Cited on page 20.

Copernicus Climate Change Service. **Land cover classification gridded maps from 1992 to present derived from satellite observations**. ECMWF, 2019. Disponível em: <https://cds.climate.copernicus.eu/doi/10.24381/cds.006f2c9a>. Cited on page 25.

COVINO, Jr. Stephen D. Cramer; Bernard S. **Corrosion: Materials**. ASM International, 2005. ISBN 9781627081832. Disponível em: <http://dx.doi.org/10.31399/asm.hb.v13b.9781627081832>. Cited on page 62.

DEAN DAGMAR KNOTKOVA, Katerina Kreislova Sheldon W. **ISOCORRAG International Atmospheric Exposure Program: Summary of Results**. [S.l.]: ASTM, 2010. Cited 2 times in pages 23 and 47.

EMD. **Glob Cover - Wiki-WindPRO**. 2005. Disponível em: <https://help.emd.dk/mediawiki/index.php/Glob_Cover>. Cited on page 26.

GAVRYUSHINA, Andrey Marshakov Marina; PANCHENKO, Yuliya. Application of the random forest algorithm to predict the corrosion losses of carbon steel over the first year of exposure in various regions of the world. **Corrosion Engineering, Science and Technology**, Taylor Francis, v. 58, n. 3, p. 205–213, 2023. Disponível em: <https://doi.org/10.1080/1478422X.2022.2161336>. Cited 4 times in pages 15, 16, 36, and 41.

GAVRYUSHINA, M.A.; MARSHAKOV, A.I.; PANCHENKO, Yu. M. Application of the random forest algorithm of corrosion losses of aluminum for the first year of exposure in various regions of the world. **Protection of Metals and Physical Chemistry of Surfaces**, v. 59, n. 1, p. 85 – 95, 2023. Cited by: 0. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159860635&doi=10.1134%2fS2070205123700259&partnerID=40&md5=0dc55b0d4de2e35f739d1d0f07f68e68>. Cited on page 36.

GELARO, Ronald et al. The modern-era retrospective analysis for research and applications, version 2 (merra-2). **Journal of Climate**, American Meteorological

Society, Boston MA, USA, v. 30, n. 14, p. 5419 – 5454, 2017. Disponível em: <https://journals.ametsoc.org/view/journals/clim/30/14/jcli-d-16-0758.1.xml>. Cited 2 times in pages 16 and 25.

GENTIL, Vicente. **Corrosão. 6ª edição**. [S.l.: s.n.], 2011. ISBN 852-1618042. Cited on page 19.

GEREMIAS, Vinicius Michelon et al. Predição de corrosão atmosférica em materiais metálicos utilizando aprendizado de máquina. In: **Anais do XVI Congresso Brasileiro de Inteligência Computacional (CBIC2023)**. Salvador, Bahia: SBIC, 2023. p. 1–8. Cited 4 times in pages 16, 51, 53, and 66.

GEURTS, Pierre; ERNST, Damien; WEHENKEL, Louis. Extremely randomized trees. **Machine Learning**, v. 63, n. 1, p. 3–42, Apr 2006. ISSN 1573-0565. Disponível em: <https://doi.org/10.1007/s10994-006-6226-1>. Cited on page 29.

GERON, Aurelien. **Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems**. Sebastopol, CA: Reilly Media, 2017. ISBN 978-1491962299. Cited 2 times in pages 28 and 32.

HERSBACH, H. et al. **ERA5 hourly data on single levels from 1940 to present**. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2023. 2024-02-01. Disponível em: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview>. Cited 3 times in pages 16, 24, and 25.

HOYER, S.; HAMMAN, J. xarray: N-D labeled arrays and datasets in Python. **Journal of Open Research Software**, Ubiquity Press, v. 5, n. 1, 2017. Disponível em: <https://doi.org/10.5334/jors.148>. Cited 2 times in pages 26 and 46.

JAMES, Gareth et al. **An Introduction to Statistical Learning: with Applications in Python**. Springer International Publishing, 2023. ISSN 2197-4136. ISBN 9783031387470. Disponível em: <http://dx.doi.org/10.1007/978-3-031-38747-0>. Cited 2 times in pages 28 and 31.

LEYGRAF, Christofer et al. **Atmospheric Corrosion**. Wiley, 2016. ISBN 9781118762134. Disponível em: <http://dx.doi.org/10.1002/9781118762134>. Cited on page 19.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. **CoRR**, abs/1705.07874, 2017. Disponível em: <http://arxiv.org/abs/1705.07874>. Cited on page 32.

MARCíLIO, Wilson E.; ELER, Danilo M. From explanations to feature selection: assessing shap values as feature selection mechanism. In: **2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [S.l.: s.n.], 2020. p. 340–347. Cited on page 32.

Microsoft Corporation. **Microsoft Excel**. 2018. Disponível em: <https://office.microsoft.com/excel>. Cited on page 47.

MIKHAILOV, A. A.; STREKALOV, P. V.; PANCHENKO, Yu. M. Atmospheric corrosion in tropical and subtropical climate zones: 3. modeling corrosion and dose-response function for structural metals. **Protection of Metals**, v. 43, n. 7, p. 619–627, Nov 2007. Disponível em: <https://doi.org/10.1134/S0033173207070028>. Cited 3 times in pages 15, 23, and 41.

MORCILLO, M. et al. **Corrosion y Protección de Metales en las Atmósferas de Iberoamérica. Parte I — Mapas de Iberoamérica de Corrosividad Atmosférica (Proyecto MICAT, XV.1/CYTED)**. [S.l.]: CYTED, 1998. Cited 3 times in pages 15, 24, and 47.

NASTESKI, Vladimir. An overview of the supervised machine learning methods. **Horizons**, v. 4, p. 51–62, 2017. Disponível em: <https://api.semanticscholar.org/CorpusID:171520859>. Cited on page 27.

NATIONS, Department of Economic United; DEVELOPMENT, Social Affairs Sustainable. General Assembly, **Transforming our world: the 2030 Agenda for Sustainable Development**. 2015. 16301 p. Disponível em: <https://sdgs.un.org/2030agenda>. Cited on page 17.

PANDAS. **pandas-dev/pandas: Pandas**. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.3509134>. Cited on page 46.

PANNONI, Fabio Domingos. **Manual De Construção do Aço: Projeto e Durabilidade**. [S.l.]: Instituto Aço Brasil, 2017. Cited on page 19.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Cited on page 46.

PEI, Zibo et al. Towards understanding and prediction of atmospheric corrosion of an fe/cu corrosion sensor via machine learning. **Corrosion Science**, v. 170, 2020. Cited by: 77; All Open Access, Hybrid Gold Open Access. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084403176&doi=10.1016%2fj.corsci.2020.108697&partnerID=40&md5=715b8dada8c673e3fce61f9fee129364>. Cited on page 35.

PINTOS, Salvador et al. Artificial neural network modeling of atmospheric corrosion in the micat project. **Corrosion Science**, v. 42, n. 1, p. 35–52, 2000. ISSN 0010-938X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010938X99000542>. Cited on page 41.

RIOS-ROJAS, John Fredy et al. Annual atmospheric corrosion rate and dose-response function for carbon steel in bogotá. **Atmósfera**, v. 30, n. 1, p. 53–61, 2017. ISSN 01876236. 10.20937/ATM.2017.30.01.05. Disponível em: <https://www.elsevier.es/en-revista-atmosfera-76-articulo-annual-atmospheric-corrosion-rate-dose-response-S0187623617300413>. Cited on page 23.

ROSSUM, G. Van; DRAKE, F.L. **The Python Language Reference Manual: For Python Version 3.2**. Network Theory Limited, 2011. (Python Manual). ISBN 9781906966140. Disponível em: <https://books.google.com.br/books?id=Ut4BuQAACAAJ>. Cited 2 times in pages 26 and 46.

RUSSELL, Stuart J; NORVIG, Peter. **Artificial intelligence**: a modern approach. 3. ed. New Jersey: Prentice Hall, 2010. Cited on page 27.

SANDWELL, D.T.; SMITH, W.H.F.; BECKER, J.J. **SRTM30+ Global 1-km Digital Elevation Model (DEM): Version 11: Land Surface**. 2014. Distributed by the Pacific Islands Ocean Observing System (PacIOOS). Disponível em: <http://pacioos.org/metadata/srtm30plus_v11_land.html>. Cited on page 25.

SANTANA, Juan J. et al. Shortcomings of international standard iso 9223 for the classification, determination, and estimation of atmosphere corrosivities in subtropical archipelagic conditions—the case of the canary islands (spain). **Metals**, v. 9, n. 10, 2019. ISSN 2075-4701. Disponível em: <https://www.mdpi.com/2075-4701/9/10/1105>. Cited 2 times in pages 15 and 23.

SCHWEITZER, Philip A. **Fundamentals of metallic corrosion**. Boca Raton, FL: CRC Press, 2006. (Corrosion Engineering Handbook, Second Edition). Cited on page 19.

SCHWEITZER P.E., Philip A. **Fundamentals of Metallic Corrosion: Atmospheric and Media Corrosion of Metals**. CRC Press, 2006. Disponível em: <http://dx.doi.org/10.1201/9780849382444>. Cited 2 times in pages 15 and 19.

SEGHIER, Mohamed El Amine Ben et al. On the modeling of the annual corrosion rate in main cables of suspension bridges using combined soft computing model and a novel nature-inspired algorithm. **Neural Computing and Applications**, v. 33, n. 23, p. 15969–15985, Dec 2021. ISSN 1433-3058. Disponível em: <https://doi.org/10.1007/s00521-021-06199-w>. Cited on page 36.

SIEFERT, Cesar Augusto Crovador et al. Avaliação de séries de velocidade do vento de produtos de reanálises climáticas para o brasil. **Revista Brasileira de Meteorologia**, FapUNIFESP (SciELO), v. 36, n. 4, p. 689–701, dez. 2021. ISSN 0102-7786. Disponível em: <http://dx.doi.org/10.1590/0102-7786360026>. Cited on page 49.

TERRADOS-CRISTOS, Marta et al. Corrosion prediction of weathered galvanised structures using machine learning techniques. **Materials**, v. 14, n. 14, 2021. Cited by: 6; All Open Access, Gold Open Access, Green Open Access. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85111098385&doi=10.3390%2fma14143906&partnerID=40&md5=c5bb7259e1656f897607f350f747409a>. Cited on page 35.

TIDBLAD, Johan et al. Un ece icp materials: Dose-response functions on dry and wet acid deposition effects after 8 years of exposure. **Water, Air, and Soil Pollution**, v. 130, n. 1, p. 1457–1462, Aug 2001. ISSN 1573-2932. Disponível em: <https://doi.org/10.1023/A:1013965030909>. Cited on page 23.

TOWNSEND, Herbert E. **Outdoor Atmospheric Corrosion**. ASTM International, 2002. ISBN 9780803128965. Disponível em: <http://dx.doi.org/10.1520/STP1421-EB>. Cited on page 20.

TRAN, Ngoc-Long et al. A machine learning-based model for predicting atmospheric corrosion rate of carbon steel. **Advances in Materials Science and Engineering**, v. 2021, 2021. Cited by: 11; All Open Access, Gold Open Access. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85122378342&doi=10.1155%2f2021%2f6967550&partnerID=40&md5=61fdd699d88665681f2c5283ba4b6462>. Cited on page 35.

VERA, R et al. Mapa de corrosión atmosférica de chile: resultados después de un año de exposición. **Revista de la construcción**, Pontificia Universidad Catolica de Chile, v. 11, n. 2, p. 61–72, ago. 2012. ISSN 0718-915X. Disponível em: <http://dx.doi.org/10.4067/S0718-915X2012000200007>. Cited 2 times in pages 16 and 22.

YAN, Luchun; DIAO, Yupeng; GAO, Kewei. Analysis of environmental factors affecting the atmospheric corrosion rate of low-alloy steel using random forest-based models. **Materials**,

v. 13, n. 15, 2020. Cited by: 12; All Open Access, Gold Open Access, Green Open Access. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089751868&doi=10.3390%2fMA13153266&partnerID=40&md5=d276c24195799958d8b78daafbe57d36>. Cited on page 35.

YANG, Xiaojia et al. Data-mining and atmospheric corrosion resistance evaluation of sn- and sb-additional low alloy steel based on big data technology. **International Journal of Minerals, Metallurgy and Materials**, v. 29, n. 4, p. 825–835, Apr 2022. ISSN 1869-103X. Disponível em: <https://doi.org/10.1007/s12613-022-2457-9>. Cited on page 36.

ZHI, Yuanjie et al. Prediction and knowledge mining of outdoor atmospheric corrosion rates of low alloy steels based on the random forests approach. **Metals**, v. 9, n. 3, 2019. ISSN 2075-4701. Disponível em: <https://www.mdpi.com/2075-4701/9/3/383>. Cited on page 16.

ZHI, Yuanjie et al. Improving atmospheric corrosion prediction through key environmental factor identification by random forest-based model. **Corrosion Science**, v. 178, 2021. Cited by: 34. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096197529&doi=10.1016%2fj.corsci.2020.109084&partnerID=40&md5=eb3f027dd91281bc5e6b16c12b4abb8f>. Cited on page 35.

ZHI, Yuanjie; YANG, Tao; FU, Dongmei. An improved deep forest model for forecast the outdoor atmospheric corrosion rate of low-alloy steels. **Journal of Materials Science and Technology**, v. 49, p. 202 – 210, 2020. Cited by: 30. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85081126046&doi=10.1016%2fj.jmst.2020.01.044&partnerID=40&md5=b9e7ac859ae48b041321be2441d1b89d>. Cited on page 15.