Instituto Tecnológico y de Estudios Superiores de Occidente

Recognition of official validity of higher level studies according to secretarial agreement 15018, published in the Official Gazette of the Federation on November 29, 1976.

Department of Electronics, Systems and Informatics

Master in Computer Systems

DEVELOPMENT OF A SUPERVISED MODEL OF REGRESSION WITH MACHINE LEARNING FOR PARTICULATE MATTER IN THE METROPOLITAN AREA OF GUADALAJARA

RECEPTIONAL WORK to obtain the **TITLE** of **MASTER IN COMPUTER SYSTEMS**

Presented by: Francisco Alonso Alavez Sosa

Director: Gloria Elena Faus Landeros, M.S.

Co-Director: Iván Esteban Villalón Turrubiates, Ph.D.

External Advisor: Edward A. Celarier, Ph.D.

Tlaquepaque, Jalisco. June 12, 2024

ACKNOWLEDGEMENTS

The author wishes to thank the Instituto Tecnológico de Estudios Superiores de Occidente (ITESO), in Mexico, for the resources and support to develop this project under the study plan and subjects offered throughout the course, which were fundamental to the continued progress of this project.

The author would like to give special thanks to the project advisor, Professor Gloria Elena Faus Landeros, who guided the research with her extensive knowledge on the topic of particulate matter.

The author would like to thank Giovanni Nasa Web Page, for collecting historical satellite information from different sources and concentrating it in such a way that it could be accessed easily and quickly through its web platform, this information being the main reason and axis on which this analysis is developed.

Last but not least, the author would like to thank Oracle Mexico Development Center for the scholarship awarded to study the Master's Degree in Systems.

DEDICATION

The author dedicates this thesis to his family, partner and close friends. Those who supported the author at all times emotionally and through advice that helped the author face the challenges that the master's degree represents.

SUMMARY

The Metropolitan Area of Guadalajara (MAG) is the second most contaminated area in Mexico. The effects of the change could be seen in the region's climate system on the ecosystem and the health of its population, caused mainly by anthropogenic activities that produce atmospheric pollutants. A greater understanding of the atmosphere, the phenomena and effects of pollutants on the air quality of the MAG is needed. particulate matter (PM) affects directly air quality, human health and the environment, mainly. [1] The MAG does not have measurements of PM 2.5 micrometers (anthropogenic) of its entire surface, only with PM 10 micrometers. [2]

In this project, using Python, data from the ZMG and its surroundings were extracted from Giovanni Nasa Web Page files, with the netCDF4 library. Three measurements of interest (Ångström Exponent, Aerosol Optical Depth and Mass Concentration) were integrated into a single dataset. From this, descriptive statistics were obtained with Seaborn, Matplotlib and Tableau. Seasonal patterns stood out and motivated the search for a machine learning regressor that would "learn" and predict the size of the particles by month. Due to the distribution of the patterns, it was decided to use a "boosting" model. Three ways were proposed to "train" the model. The first, with the complete dataset. The second and third with "feature engineering" and a "sliding window", which considers trends followed each month, using data from the MAG and data from the MAG and its surroundings, respectively. The final model makes accurate quantitative and qualitative forecasts of particle size in a given month, useful for taking preventive measures.

Contents

List of Figures

List of Tables

LIST OF ACRONYMS AND ABBREVIATIONS

AE Ångström Exponent. 3, 6, 7, 9, 10, 13, 27, 28, 35–37, 41, 44–46, 51, 53, 56, 57

AOD Aerosol Optical Depth. 3, 5–7, 9, 10, 13, 18–20, 24, 28, 35, 36, 41, 47, 48, 56

GOCI Geostationary Ocean Color Imager. 5, 6, 18, 27

HDF5 Hierarchical Data Format 5. 36

ITESO Instituto Tecnológico de Estudios Superiores de Occidente. 1

MAG Metropolitan Area of Guadalajara. 3, 6, 8, 9, 13–15, 19, 28, 35, 36, 44, 47, 49, 52–54, 57

MC Mass Concentration. 3, 6, 7, 9, 10, 28, 36, 37, 41, 49, 50, 56

MERRA2 Modern-Era Retrospective analysis for Research and Applications, Version 2. 36

MODIS Moderate Resolution Imaging Spectroradiometer. 5, 6, 13, 18–20, 26, 27, 35, 44, 47

NASA National Aeronautics and Space Administration. 12, 13, 15, 20, 56

PM particulate matter. 3, 5, 12–14, 18–22, 35, 36, 54, 56, 57

VIIRS Visible Infrared Imaging Radiometer Suite. 5, 6, 18, 27

WHO World Health Organization. 18

1. INTRODUCTION

Summary: *particulate matter of 2.5 μm represents a danger to human health. The investigation and identification of this particulate matter is complicated and expensive as more certainty is required. However, with satellite measurements published by the National Aeronautics and Space Administration (NASA), forecasts can be obtained economically and from a certain region of interest.*

1.1 Background

There is existing research that aims to analyze and improve satellite data to reach interpretations and knowledge of the behaviors and effects of PM in the atmosphere. This information aims to have a clear picture of what happens in certain periods of time in a certain region and to reach conclusions that allow us to know and understand the processes of aerosols and their observable effects in: climate systems, chemical processes when transported in the atmosphere from one region to another, risks of morbidity and mortality of the exposed population, alterations in ecosystems, among others. Some of the articles that have used observations and analyzed satellite data to obtain greater knowledge and understanding of the role of particulate matter in a given region are mentioned, such as:

- Global monitoring of air pollution over land from the Earth Observing System-Terra Moderate Resolution Imaging Spectroradiometer (MODIS) [3]
- Aerosol Optical Depth (AOD) Retrieval Method using MODIS. [4]
- Satellite Aerosols Retrieval over Land Surfaces using the Structure Functions [5]
- Techniques of Global Validation of Aerosol Retrievals from MODIS [6]
- Satellite Measurements of the Ångström Exponent (AE) using an Innovative Mathematical Method to Identify Seasonal Aerosols [7]

1.2 Justification

There is information about some of the chemical, physical and optical properties of PM 2.5μm done by researchers in different cities around the world[8], however, there is no quantitative and qualitative study on these variables applied specifically in the MAG. It is necessary to have measurements of the physical, optical and chemical properties of the particulate matter of interest, generated by industrial and vehicular activities. These features of interest are detected by different satellite sensors. Giovanni NASA Web Page is responsible for compiling this wide variety of measurements from different sources and resolutions and, through a graphical interface, facilitates obtaining the required information in the geographical area of interest. The processing and analysis of the mentioned data will help to identify the variables of study and main characteristics of anthropogenic contamination. Particulate matter has direct effects on air quality and damage to human health. It has been proven that PM 2.5 μm is the main cause of respiratory diseases and some types of cancer. This is because it is able to cross the lung barrier and enter the blood system. Because of this, chronic exposure to PM 2.5 μm increases the risk of cardiovascular, respiratory and even respiratory diseases and even death.[9]

This project is meant to obtain quantitative and qualitative knowledge of the concentration, size, and optical depth of PM. This will enhance the knowledge about their concentration and seasonal patterns of behavior, which in turn will help to identify areas and seasons of risk to the health of MAG population. Likewise, it seeks to understand the atmospheric alterations due to the concentration of the aerosol in question and consequently, the climate and energy balance of the studied region.

1.3 Problem

A way to analyze satellite data of certain aerosol measurements is needed to help and reach valuable conclusions for decision making and implementation of preventive and/or corrective actions in a given geographic area. Aerosols and their different behaviors require an in-depth study that involves geographic, atmospheric, and demographic data, among others. In order to achieve a complete analysis of the observable phenomena, different sources and even specialized measurement instruments are required, such as a solar photometer. The complexity increases with respect to the level of certainty that is desired to be obtained in the study, so a method is needed that can reach accurate conclusions and forecasts to make decisions and necessary measures in a specific geographical area, saving resources and time. Based on historical satellite data, statistical analysis, correlation and regression models, it is intended to reach this forecast.

1.4 Hypothesis

Through the processing of satellite measurements of particulate matter, for a period greater than 20 years in the Metropolitan Area of Guadalajara of certain aerosol measurements, seasonal patterns of PM 2.5 μm can be quantified and identified. The seasonal patterns of the PM behavior identified can be used as a basis for the development of prediction and forecast models of its quantitative and qualitative characteristics.

1.5 Goals

1.5.1 General Goal

Group aerosol characteristics from publicly accessible satellite measurements of the MAG and reach conclusions that can be correlated with observable effects on the population and the environment where they prevail. Through the descriptive statistical analysis of the historical information of the MAG, the aim is to generate a regression model that

allows the making of preventive and corrective decisions.

1.5.2 Specific Goals

1. Identify, select, process, prepare and clean aerosol satellite data of interest.
2. Build a database with the aerosol measurements of interest in a determined geographical area (MAG).
3. Find seasonal patterns and generate regression models using machine learning that allow obtaining an accurate forecast.

1.6 Scientific, technological novelty or contribution

Through the regression model developed in this project for making preventive and corrective decisions, resources can be saved, including the time it takes to acquire, install and train higher precision instruments, such as a solar photometer. This research is a contribution to the state of Jalisco, using public information supported and validated by NASA. It is also a technological proposal that seeks to implement machine learning algorithms and software tools designed for data analysis and extraction of valuable information, which can be considered for future decision-making at the city or state level. A scientific contribution obtained is the seasonal patterns of behavior of particulate matter in the (MAG).

2. STATE OF THE ART OR OF THE TECHNIQUE

Summary: *Over the years, different ways have been proposed to obtain satellite measurements of aerosols and there are also techniques that study them from terrestrial means. Studies have been done to observe the behaviors of particulate matter with different sensors and conditions, helping to understand the atmospheric phenomena.*

2.1 Machine Learning

Machine learning is a subset of artificial intelligence that allows a system to "learn" and improve autonomously using tools that can improve a statistical model, without having to be programmed manually, through data ingestion. Machine learning allows computer systems to adapt and obtain more accurate results, to "train." In this way, continuous improvement is sought as more "experiences" are accumulated. The more data they consume, the more accurate the results could get. The main goal of this tool is to identify patterns and make decisions with minimal human intervention. There are different types of machine learning models:

- Supervised Learning

  These models use "labeled" training data (structured data), where a specific feature is assigned to a label. The result is known a priori and the model is trained with data from the known result. In other words, to come up with an algorithm that recognizes images of apples, it has to be fed images labeled as apples.

- Unsupervised Learning

  This class of models uses unlabeled data (unstructured data) to learn patterns. Unlike supervised learning, the "correctness" of the result is not known a priori. The algorithm learns from the data without human intervention (unsupervised) and classifies it into groups based on its characteristics. For example, if the algorithm is given images of apples and bananas, it will distinguish which image is an apple and which is a banana. Unsupervised learning is good for descriptive models and pattern distinction.

- Reinforcement Learning

  These models can be described as "learning on the fly" through trial and error. An "agent" learns a defined task with a feedback loop until its performance is within the expected range. The agent is "rewarded" when he performs the task well and "punished" when he does it poorly. An example of reinforcement learning is when Google researchers taught a reinforcement learning algorithm for playing the game Go. The model, who had no prior knowledge of the rules of Go, simply moved pieces at random and "learned" the best moves to make. The algorithm was trained using positive and negative reinforcement to the point that the machine learning model could beat a human player.

[10]

2.2 Aerosol properties in cloudy environments from remote sensing observations: a review of the current state of knowledge

There are a variety of factors that can modify the behavior of particulate matter (PM). The different effects that clouds have on PM have been observed. The different characteristics of aerosols have been taken into account, from different ranges in their measurement and positions above or below the clouds. Certain behavioral patterns and causes were found by which aerosols change their physical, optical and chemical properties. [11] Due to the sensitivity of aerosols to environmental variables, greater knowledge is necessary about their properties, behavior and effects in the atmosphere of the region of interest.

2.3 Large global variations in measured airborne metal concentrations driven by anthropogenic sources

The specific study of PM2.5 μm is essential to understand health impacts, prioritize pollution mitigation strategies and enable the development of a global chemical transport model. Among the chemical composition of the PM 2.5 that were found, metal concentrations that come from anthropogenic sources stand out, such as lead, arsenic, chromium and zinc, which exceeded the recommended health particle concentration index or limit in different areas. places. Levels of heavy metals present in the air exceed health guidelines established by the World Health Organization (WHO) and other standards in multiple regions. For example, the Dhaka and Kanpur areas exceeded the US National Ambient Air Quality Standard of 3 months for lead (150 ng m^{-3}). Kanpur, Hanoi, Beijing and Dhaka had mean annual arsenic concentrations that approached or exceeded the WHO risk level of 6.6 ng m^{-3}. [[12]

2.4 Evaluation of VIIRS, GOCI, and MODIS Collection 6 AOD retrievals against ground sunphotometer observations over East Asia

In this article, measurements obtained by three satellite sensors are compared and correlated: VIIRS, GOCI and MODIS. Information on the optical depth of the particulate matter was collected. The results of emerging methods used to obtain aerosol measurements and observations were compared with terrestrial source information, through a CIMEL solar photometer, consolidated and validated. This will serve as a reference point to verify the certainty of observations of emerging products. It is shown that the measure-

ments can change completely depending on the geographical area. Through this study, it can be seen that measurements with the solar photometer are the way to validate the data and measure the errors that result in satellite images. [13]

2.5 A Physically Based PM 2.5 Estimation Method Using AERONET Data in Beijing Area

This article provides information related to particulate matter (PM) and its effects on health and the environment in the Beijing area, China. According to this research article, particulate matter contributes to a decrease in atmospheric visibility in the urban area of this region, but in addition to that, it also increases mortality and morbidity with respect to diseases of the respiratory system, which is why it becomes a problem, both environmental and health, of utmost importance in today's society. The particulate matter studied in this article includes PM 10 and PM 2.5, particulate matter with a size equal to 10 micrometers or less and equal to 2.5 micrometers or less, respectively. This makes use of data collected by satellites, as well as observations at the level of the Earth's surface, and provides with a guide on how to use both types of data to validate the certainty of both measurements through correlations and validation of these. [14] It should be noted that this study is similar to the development of the software proposed in this project, which will allow the generation of a risk map according to the seasons of the year in the Metropolitan Area of Guadalajara, with corresponding objectives and methodology.

2.6 Evaluation of MODIS C6 Combined Aerosol Product at Global Scale

This study allows the understanding on the use of data collected by the AQUA and TERRA satellites that use the Moderate Resolution Imaging Spectroradiometer remote sensing instrument. The same will be the provider of measurements that feed the software that will be developed as part of this project. Likewise, it provides useful information on the most recent collection (C6 database) obtained by the algorithms (Dark Target and Deep Blue, with their implementations for land and ocean) used to carry out remote satellite measurements (specifically, the AOD) of PM. Another important characteristic of this project is that in addition to making use of MODIS data, this data is validated against ground observations from the AERONET network to validate the certainty of these satellite measurements. It is important to use both since satellite measurements are not completely reliable because they cover large geographic areas. Depending on the resolution with which the information is captured, its certainty will vary. The result delivered is an image with averages of a certain measurement for each pixel. If the image has a lower resolution, each pixel will be translated to a larger geographic area and

therefore the average of everything observed there. If the resolution is higher, each pixel will cover less space and its average will not have as much noise from other geographic areas. [15]

2.7 Algorithm for Remote Sensing of Tropospheric Aerosol from MODIS: Collection 5

This article, obtained directly from the MODIS site on the corresponding NASA page, describes the implementation of the algorithms used by the MODIS sensor to capture PM (or aerosol) data, in this case, for the collection C5, one before the more recent C6 collection. However, it is a useful article since, between collections, the changes made in the algorithms represent corrections that are intended to improve or reduce the errors seen in previous editions of these, as exemplified in this case for collection C4 and collection C5 (the most recent at that time). It should be noted that by collection we refer to a series of observations or databases obtained by a particular version of the algorithms. Despite not being a recent article (5 years or less) and it is clear that we must look for a similar one for the C6 collection to have the newest information, the algorithms and implementations of the mathematical equations used are still essentially the same: the first for land surfaces and the second, for marine surfaces. From the C4 to C5 collection the changes were minimal and as mentioned before, the objective and result was to improve the quality of the measurements of particulate matter, such as Aerosol Optical Depth. [16]

3. Theoretical/Conceptual Frame

Summary: *This chapter presents the theoretical and conceptual bases on aerosols and particulate matter. The physical and chemical processes of interest for the scope of this project. Likewise, the computer algorithms and programming techniques explored in the development of the predictive model are described.*

3.1 Aerosol

Particles suspended in the atmosphere that can be solid or liquid. This matter has its own characteristics and reacts differently to the atmospheric and geographical phenomena of the studied region. In this study, the concept of aerosol and particulate matter will be used in the same way.

3.1.0.1 Coarse Particles

Airborne particles of relatively large size and produced mainly by the disintegration of even larger particles by mechanical processes. Some examples are: dust, pollen, spores, ashes and fragments of plants and insects. It can be seen that they are all of natural origin. [17]

3.1.0.2 Fine Particles

Suspended, airborne particles smaller than coarse particles. They have an aerodynamic diameter less than or equal to 2.5 micrometers. They are formed mostly from gases and anthropogenic activities. [18]

3.1.1 Continental Dust

Dust particles, often called particulate matter (PM), can be made up of hundreds of different chemicals. Some can be traced to a specific source, such as construction sites, unpaved roads, fields, smokestacks, or fires. However, most particles form in the atmosphere as a result of complex reactions of chemicals, such as sulfur dioxide and nitrogen oxides. These are pollutants emitted by power plants, industries and cars. Both size and chemical composition vary widely in relation to the nature of the source and the history of the particles. [19]

3.1.2 Aerosols from Nearby Sources

The main sources of aerosols include urban and industrial emissions, smoke from burning biomass, secondary formation of gaseous aerosol precursors, sea salt and dust. The pending problems seek to discover the natural sources of the aerosols and the organic part of them. It is important to be clear that aerosols from anthropogenic sources are those of interest in this study. The aerosols of natural sources do not put human health at risk and do not contribute to the Earth's energy imbalance. [20]

3.1.3 Aerosols from Distant Sources

Every year, winds blow huge amounts of soil dust from Africa, across the Atlantic and to the Caribbean. No other ocean region is so extensively and persistently affected by such high concentrations of dust, a region that extends more than 7,000 km from the coast of Africa to the Caribbean and the bordering continental coasts of the Americas. The Caribbean Basin can be considered the "main recipient" but not the only one. The Saharan dust "source" accounts for more than half of global dust emissions. It is generally recognized that, on a global scale, mineral dust can affect many aspects of climate, marine biogeochemical processes, soil fertility, air quality and human health. However, it is difficult to assess the impact in the Caribbean Basin due to the paucity of regional studies. [21]

3.1.4 Black Carbon Aerosols

Dominant form of light-absorbing particles in the atmosphere. Black carbon is emitted by incomplete combustion processes, both human (diesel engines) and natural (forest fires). Its ability to absorb visible and infrared radiation means black carbon can warm the atmosphere and darken surfaces, specifically snow and ice. [22]

3.1.5 Smoke Aerosols

The burning of biomass releases a significant amount of smoke aerosols and gases into the atmosphere that contribute to the carbon footprint. Carried through the atmosphere, smoke aerosols are the largest mass source of primary fine carbonaceous particles and degrade regional air quality, reduce visibility, influence weather and climate, and threaten public health. [23]

3.1.6 Sea Salt Sprays

Sea salt aerosol is the main source of tropospheric reactive chlorine (Cly) and bromine (Bry). The effects of other sea salts on atmospheric chemistry have not been explored. These aerosols come from physical processes where breaking waves drag air, resulting in bubbles bursting at the air-sea interface, launching jets and film droplets into the atmosphere, which form the aerosol. A subset of these catalyze cloud formation by acting as cloud condensation nuclei (CCN) or ice nucleating particles (INP). [24]

3.2 Aerosol Transport

Atmospheric chemical transport models predict air pollutant concentrations taking into account the transformation and chemical reactions of the pollutants. They model the quantitative and qualitative changes of aerosols as they are transported from one region to another. [25]

3.2.1 Convection in Hot Areas

Convection works by heating or cooling areas of a liquid or gas larger than their surroundings, causing temperature differences. These temperature differences cause areas to move. The hottest and less dense areas rise and the coldest and most dense areas sink. Convection within the atmosphere can often be observed in weather. For example, as the sun heats the Earth's surface, the air above it warms and rises. If conditions allow, this air can continue to rise, cooling as it does so, forming cumulus clouds. Stronger convection can lead to the formation of much larger clouds as the air rises before cooling, sometimes producing Cumulonimbus clouds and even thunderstorms. [26]

3.2.2 Deposition during Transport

It is the sedimentation of particles or sediments on a surface. The particles can come from a vapor, a solution, a suspension or a mixture. Deposition also refers to the phase change from gas to solid and can be wet or dry. Both wet and dry deposition can be transported by wind, sometimes over extremely long distances. [25]

3.2.3 Wet Deposition

Type of atmospheric deposition in which chemicals and atmospheric particles are incorporated into small droplets and transferred to the Earth's surface in the form of rain, fog or snow. [27]

3.3 Aerosol Optical Depth (AOD)

It is a measure of light extinction by aerosols in the atmospheric column. This satellite measurement of aerosols, also called optical thickness, is based on the fact that particles change the way the atmosphere reflects, absorbs and scatters visible and infrared light.

An optical thickness less than 0.1 indicates a crystal clear sky with maximum visibility, while a value of 1 indicates very foggy conditions. [28]

3.4 Artificial satellites

Man-made artifacts that orbit the planet and help collect information about Earth and the universe. There are thousands of these orbiting the world, some taking photos of the planet that help meteorologists predict the weather and track hurricanes. Others take photos of other planets, the sun, black holes or distant galaxies. They are important because they can collect more data, faster than ground-based instruments. [29]

3.4.1 Terra

It is a satellite the size of a small school bus. It has 5 instruments that take measurements of the Earth system. [30]

3.4.1.1 Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER)

It obtains high-resolution images of the Earth (15 to 90 square meters per pixel) at 14 different wavelengths of the electromagnetic spectrum. The range is from visible to thermal infrared light. Scientists use ASTER information to create detailed maps of the Earth's surface. They can obtain its temperature, emissivity, reflectance and elevation. [31]

3.4.1.2 Clouds and Earth's Radiant Energy System (CERES)

There are 2 identical instruments on Terra that measure the balance of Earth's total radiation, as well as compile estimates of cloud properties that allow scientists to describe the roles of clouds in radiative fluxes from the surface to the highest point of the atmosphere. [32]

3.4.1.3 Multi-angle Imaging SpectroRadiometer (MISR)

It is a new type of instrument designed to know the amount of sunlight that is scattered in different directions under natural conditions. It has a view of the Earth with cameras that point at 9 different angles. One camera points towards nadir and the others point

back and forth to the Earth's surface. They are angles of 26.1°, 45.6°, 60.0° and 70.5°. [33]

3.4.1.4 Measurement of Pollution in the Troposphere (MOPITT)

Instrument designed to improve knowledge of the lower atmosphere and observe how it interacts with the terrestrial and marine biospheres. The special focus of this sensor is the distribution, transport, sources and sinks of carbon monoxide in the troposphere, which is expelled by factories, cars and forest fires. This aerosol hinders the atmosphere's natural ability to remove harmful pollutants. [34]

3.4.1.5 Moderate Resolution Imaging Spectroradiometer (MODIS)

This sensor observes all points of the world every one or two days in 36 discrete spectral bands. For this reason, MODIS tracks a broad collection of the Earth's "vital signs" more than any other sensor. This capability allows MODIS, together with MISR and CERES, to obtain information that determines the impact of clouds and aerosols on the Earth's energy balance. [35]

3.4.2 Aqua

It is a satellite that is synchronized with the sun. Its orbit is very close to the polar orbit. It was launched on May 4, 2002 and has 6 instruments on board:

1. Atmospheric Infrared Sounder (AIRS)
2. Advanced Microwave Sounding Unit (AMSU-A)
3. Humidity Sounder for Brazil (HSB)
4. Advanced Microwave Scanning Radiometer for EOS (AMSR-E)
5. Moderate-Resolution Imaging Spectroradiometer (MODIS)
6. Clouds and the Earth's Radiant Energy System (CERES)

[36]

3.5 Visible Infrared Imaging Radiometer Suite (VIIRS)

This sensor is aboard the "NASA/NOAA Suomi National Polar-orbiting Part-nership (Suomi NPP)" and "NOAA-20 satellites". It is designed to collect visible and infrared images, along with terrestrial, atmospheric, cryosphere and ocean observations. VIIRS complements records collected by similar instruments aboard previously orbiting satellites, such as MODIS. [37]

3.6 Geostationary Ocean Color Imager (GOCI)

It is one of the 3 instruments on board the "Communication, Ocean and Meteorological Satellite (COMS)". GOCI obtains multispectral images in 8 bands (6 visible and 2 near infrared) with a spatial resolution of approximately 500 meters over the Korean Sea. [38]

3.7 Deep Blue

Algorithm that uses measurements made by satellite instruments to determine the amount of aerosols in the atmosphere and their properties. Such aerosols include all particles suspended in the atmosphere, including desert dust, smoke, volcanic ash, industrial smog and sea spray. [39]

3.8 Modern-Era Retrospective analysis for Research and Applications (MERRA), Version 2

Analysis model that combines computational algorithms with satellite observations of the Earth's surface, atmosphere and ocean. It has data since the early 1980s. It was introduced to replace MERRA because of advances made in understanding hyperspectral radiation and microwave observations, along with GPS-Radio Occulation data sets. It has a spatial resolution of approximately 50 km in latitudinal direction. [40]

3.9 Ångström Exponent (AE)

The Ångström Exponent provides information about the size distribution of particles and is defined by:

$$\alpha = \frac{\ln(\frac{\tau_2}{\tau_1})}{\ln(\frac{\lambda_1}{\lambda_2})}$$

Where α is the Ångström Exponent, τ is the Aerosol Optical Depth (AOD) and λ is the wavelength of the incident light. [7]

3.10 Mass Concentration (MC)

It is a measure of the density of aerosols. Columnar aerosol mass concentration (μgm/cm2) is the total mass of aerosols in a vertical column of atmosphere.

3.11 Metropolitan Area of Guadalajara (MAG)

It is conformed by the municipalities of San Pedro Tlaquepaque, Tonalá, Zapopan, Tlajomulco de Zúñiga, El Salto, Juanacatlán, Ixtlahuacán de los Membrillos, Acatlán de Juárez, Zapotlanejo and Guadalajara itself, which together share a constant conurbation. The National Institute of Statistics and Geography (INEGI) indicates that the MAG is the second most populated in the Mexican Republic and is only surpassed by the Metropolitan Area of the Valley of Mexico. [41]

3.12 HDF5 Format

Stands for "Hierarchical Data Format" version 5. It is a contribution format open source that supports extensive, complex and heterogeneous data. It uses a structure similar to a file directory, which organizes the data within the file in many structured ways, such as in a computer file system. This format also allows embedding of metadata that helps describe the data. [42]

3.13 Python

Interpreted, interactive and object-oriented programming language. It incorporates modules, exceptions, dynamic code, very high-level and dynamic data types and classes. It supports different programming paradigms beyond object-oriented, such as procedural

and functional programming. Python combines outstanding power with easy, clear syntax. It has interfaces to many system calls and libraries, such as windowing systems and support for C or C++. It is useful as an extension language for applications that need a programmable interface. Finally, Python is portable, capable of running on many Unix variants, including Linux, macOS, and Windows. [43]

3.13.1 Numpy

Fundamental library for scientific computing in Python. Contains multidimensional array objects, various derived objects such as masked arrays and arrays. It also offers a variety of routines for fast operations on arrays, including mathematics, logic, shape manipulation, sorting, selection, input and output, Fourier transforms, basic linear algebra, basic statistics, simulation of randomness and much more. [44]

3.13.2 Pandas

Python package that offers fast, flexible and expressive data structures. These structures are designed to work with related or labeled data in an easy and intuitive way. It is intended to be the fundamental, high-level building block for doing practical, real-world analysis. Its advantages stand out in many different types of data such as tabular with heterogeneous data, such as an Excel table or an SQL table. You can use ordered or unordered data from time series and observational and statistical data sets. [45]

3.13.3 netCDF4

Python interface to the C netCDF library. Version 4 has new features, implemented on the HDF5 format. This module can read and write files in the version 3 or 4 format and can create files readable on HDF5 clients. It has multiple and unlimited dimensions, groups and data compression. All new numeric data types are implemented, such as unsigned integer or 64 bit. This version is composite or structured, it can have variable length and "enum" data types. However, it does not support "opaque" data types. Mixtures of structures, variable lengths, and enumerables are not supported. [46]

3.13.4 SciKit-Learn

Machine Learning library with open source code that supports supervised and unsupervised learning. It also contains several tools for training machine learning models, data preprocessing, model selection, model evaluation and many other utilities. [47]

3.13.4.1 Pipeline

In Scikit-learn, it is a sequence of data transformers with the option to have a predictor at the end. The purpose of this tool is to assemble several steps that can be cross-validated together while configuring different parameters. For this, parameters of the different steps within the process are established, using their names and the specified parameter. [48]

3.13.4.2 Simple Imputer

Scikit-learn tool useful for imputing univariately to fill in missing values with simple strategies like aconstant value or measures of central tendency: mean, median or mode. [49]

3.13.4.3 Standard Scaler

Data standardization is a common requirement for many machine learning estimators. This tool standardizes the characteristics, removing the mean from each observation and scaling it to unit variance. The standard score for each sample in the data set is calculated with:

$$z = \frac{x - u}{s}$$

Where u is the mean of all observations in the dataset or zero if specified. s is the standard deviation of the training set, or 1 if specified. [50]

3.13.4.4 Grid Search

Scikit-learn tool that does an exhaustive search for the values of a specified parameter of an estimator. The parameters of the estimator used to apply these methods are optimized with cross validation. [51]

3.13.4.5 Gradient Boosting Regressor

Regressor implemented in Scikit-learn with the Gradient Boosting method. This estimator builds an additive model incrementally in stages. It allows the optimization of arbitrary and differentiable loss functions. In each phase, a regression tree is trained on the negative gradient of the specified loss function. [52]

3.14 Cross Validation

Machine learning technique that evaluates the variability of the data and confidence level of the evaluated model. It consists of testing the performance of a machine learning model, seeking to improve the performance of the model through the random decomposition of the data series. This is so that the model can be trained with all the data and validated in the same way, with data from the entire series and not just one nature. [53]

3.15 Evaluation Metrics

When training machine learning models, it is necessary to compare their performance with that of existing references. The evaluation has 2 purposes: discarding methods that do not help the model and optimizing those methods that help the model. Depending on the model task, there are different evaluation metrics that can be used to observe the performance of machine learning models. [54]

3.15.1 F1 Score

It can be interpreted as the harmonic mean of precision and recall, where the F1 score reaches the maximum at 1 and the worst at 0. The relative contribution of precision and recall are equal. The formula for this score is:

$$F1 = \frac{2 - TP}{2 * TP + FP + FN}$$

Where TP is the number of true positives, FN is the number of false negatives and FP is the number of false positives. F1 is calculated by default as 0.0, when there are no true positives, false negatives, or false positives. [55]

3.15.2 Minimum Square Error (MSE)

Measures the amount of error in statistical models. It provides information on the difference between the square mean of the observed values and those predicted by the developed model. When a model has no error, the MSE is 0. The more the error increases, the value of the MSE also increases. It is calculated with the following formula:

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

Where y_i is the ith observed value, $\hat{y}_i$ is the corresponding predicted value and n is the number of observations. [56]

3.15.3 Minimum Absolute Error (MAE)

Amount of error in measurements, using the absolute value of the difference. This is necessary because sometimes the observed measurement will be less than the predicted one, resulting in a negative value. When trying to find an average of all these differences, all the errors have to be added, so negative values would alter the total difference and therefore the average. The formula for this metric is given by:

$$MAE = \frac{\sum |y_i - \hat{y}_i|}{n}$$

[57]

3.16 Hyperparameters

Hyperparameters are external configuration variables used to manage the training of machine learning models. They are configured manually before training a model. Normal parameters are internal model elements derived automatically during the learning process and that are not configured externally. Hyperparameters determine key characteristics such as model architecture, learning rate, and model complexity. [58]

3.17 Sliding Window

This algorithm is a technique applied in data transfer and computer networks, among other areas. Another use of this technique involves identifying value patterns in time series or sequential data. This technique is powerful because it allows you to detect patterns faster than other techniques. [59] [60]

3.18 Feature Engineering

The characteristics of a machine learning model are the inputs used during training and inference to make predictions. The accuracy of the model depends on an accurate set and composition of features. Creating features may require significant engineering effort. Feature engineering involves extracting and transforming variables from raw data so that

the features can be used for training and prediction. The necessary steps within a feature engineering process include extracting and cleaning the data, and then creating and storing the features. [61]

3.19 Spearman Correlation

It's an equivalent to Pearson correlation, but it handles non-linear relationships and non-normal distributions. It is not affected by outliers either. [62]

4. Project Development

Summary: *This chapter explains the methodology that was followed to develop the project. The results of each phase of the study and how they contributed to the fulfillment of the general and specific objectives.*

4.1 Management Method

In this project it was decided to use an iterative methodology. This form of development is based on the division of work into small iterations or cycles. Each iteration represented a planning, design, implementation and testing process that was carried out in a specific period of time. At the end of each iteration, new capabilities and improvements were introduced to the software, while it could be tested and evaluated. In this way, errors and problems were also identified and corrected in the early stages of the project. Each iteration came alongside with a review, which helped to keep the goal set on the proposed objectives.

4.1.0.1 First Stage

Research was made using related works and the state of the art, it was useful for reference and guidance of the project. Likewise, the scope of the research and the objectives, both general and specific, were defined. At this stage it was important to understand the fundamental concepts and definitions of particulate matter, the phenomena that are already known and those that need to be investigated. The challenges presented by exploring aerosols and understanding the influence they have on the environment, flora, fauna and the health of the inhabitants of a certain geographical area.

4.1.0.2 Second Stage

Once the scope of the research was defined and its context understood, this iteration focused on the collection of satellite information and the generation of a structured database. The search and extraction of data involved its cleaning and standardization for subsequent analysis and generation of descriptive statistics. There were already a priori hypotheses about particulate matter in the MAG, however, the analysis of the collected information confirmed these hypotheses and generated new ones, which remained as part of the next iteration.

At this stage, the satellite data to be consulted were delimited, five significant characteristics of the aerosols and their respective chronology were defined:

- Aerosol Optical Depth of 550μm: obtained with the Deep Blue algorithm, only for land. With the EOS-Terra MODIS instrument, in version MOD08_M3 v6.1.
- Ångström Exponent at 0.412 to 0.47 microns: obtained with the Deep Blue algorithm, only for land. With the EOS-Terra MODIS instrument, in version MOD08_M3 v6.1.

- Mass Concentration of particulate matter 2.5 μm: was obtained using the MERRA2 version M2T1NXAER v5.12.4.
- Month in which the measurement was taken
- Season of the year in which the measurement was taken
- Location where the measurement was taken (MAG, North, South, East, West, Northeast, Northwest, Southeast, Southwest of the MAG)

The coordinates considered to encompass the ZMG were 105W–100W, and 18N–22N. Once the measurements and geographical area of interest were defined, the satellite data was searched in predefined time intervals, for which the seasons of the year were used, from March 21, 2000 (beginning of spring) to March 20, 2023 (end of winter). The seasonal intervals were defined as:

- Spring: starts on March 20 and ends on June 21.
- Summer: starts on June 21 and ends on September 23.
- Autumn: starts on September 23 and ends on December 21.
- Winter: starts on December 21 and ends on March 20.

The data was downloaded in netCDF4 format, which houses data in HDF5 format. AOD and AE measurements were obtained day by day, while MC measurement was generated hour by hour. Because of this, the web platform may or may not process an entire month. If this was not possible, the measurements were divided, obtaining 2 netCDF4 files per month if necessary and thus speed up data collection. NASA's Giovanni platform calculates an average of the selected measurement in the selected time interval, so the result of each coordinate already encompasses the chosen elapsed days with the specified period. With the previous specifications, 3312 observations were obtained.

In Figure 4.1, we can observe the documentation of the code where the mentioned parameters are set, Locating the data of interest within the netCDF4 files. [63]

Figure 4.2 shows the routine used to extract the data out of the netCDF4 files to later gather them all into a single dataset. [63]

Once all the data was extracted and gathered into a single dataset, we could make a quick a priori statistical analysis. The goal was to understand the data distribution and nature. Figure 4.3 shows a Spearman correlation matrix between all the extracted variables.There are variables which show dependency with the 3 variables selected. Mass Concentration has a strong negative correlation with the season, having a correlation index of -0.66. This means that they are inversely correlated. Aerosol Optical Depth is also negatively

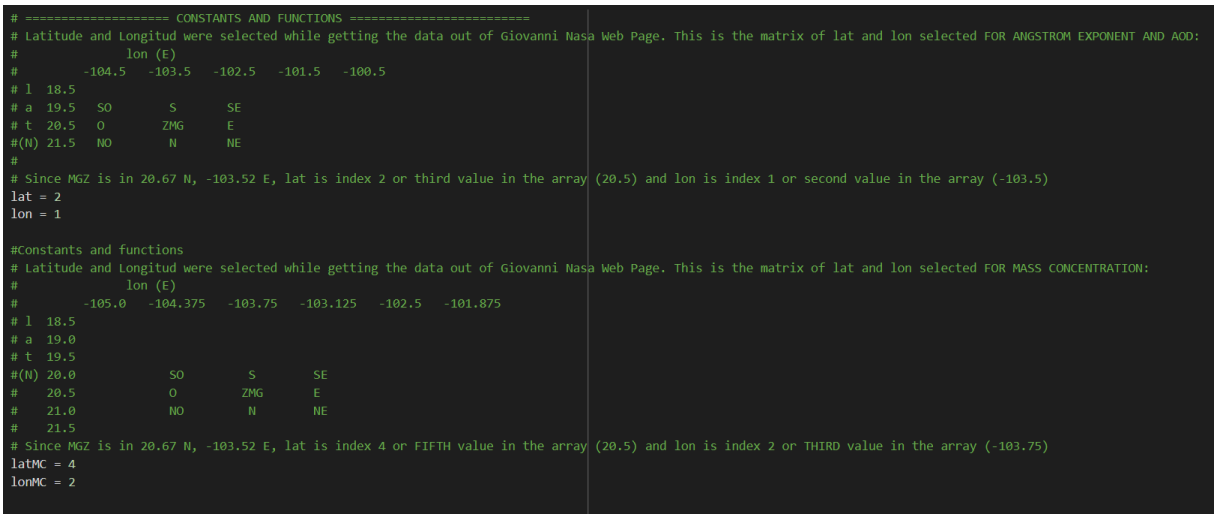


Figure 4.1: Code documentation showing the netCDF4 structure

correlated with the location, having an index of -0.44. Finally, Ångström Exponent and Mass Concentration show a weak, negative correlation, with -0.25. On another quick analysis, Figure 4.4 shows a scatter plot matrix between all the extracted variables. It shows observations concentration in certain range and also a clear distinction between seasons. These quick observations encouraged the further research and development of the machine learning model that could describe these behaviors and get a prediction. [63]

```python
def getMeasurementDataframe(measureName, sensorName, auxName, path, latitude, longitude):
    measureArr = []
    startDateArr = []
    endDateArr = []
    seasonArr = []
    yearArr = []
    coordArr = []
    monthArr = []
    locArr = []
    dfFinal = pd.DataFrame()

    isMC = (measureName == 'Mass Concentration')
    fullpath = path + sensorName + '*.nc'


    print(f'Getting data frame for {measureName}.')


    for filename in glob.glob(fullpath):
        #getting the netCDF file
        ds = nc.Dataset(filename)
        validFile = True
        #Check if we have the actual sensor or we had to use an auxiliary one (given missing satellite data for certain period of time).
        if variableExists(ds, auxName):
            sensorName, auxName = auxName, sensorName
        #Assert that the current file is a valid and known sensor.
        if not variableExists(ds, sensorName):
            print(f"UNKNOWN SENSOR DATA FOR FILE {filename}")
            validFile = False
        #print(filename)
        if validFile:
            startDate, endDate = getFileDates(filename)
            season = getSeasonName(startDate)

            for key in adjustedCoordDict.keys():
                currentLocation = key
                currentLat = latitude
                currentLon = longitude

                currentLat += adjustedCoordDict[currentLocation][0]
                currentLon += adjustedCoordDict[currentLocation][1]
                measure = float(ds[sensorName][currentLat,currentLon].data)

                #print(f"measure:{measure} lat: {currentLat} lon:{currentLon} loc:{currentLocation}")
                measureArr.append(measure)
                startDateArr.append(startDate)
                endDateArr.append(endDate)
                seasonArr.append(season)
                yearArr.append(startDate.year)
                monthArr.append(startDate.month)
                locArr.append(currentLocation)

                if (isMC and len(measureArr) > rowsPerYear * extractedYears):
                    measureArr, startDateArr, endDateArr, seasonArr, monthArr, yearArr = reduceArray(measureArr, startDateArr, endDateArr, seasonArr)

            #once we have all data into their corresponding arrays, we can proceed to create the dataframe
            data = {
                measureName: measureArr,
                'Start Date': startDateArr,
                'End Date': endDateArr,
                'Month' : monthArr,
                'Season': seasonArr,
                'Year' : yearArr,
                'Location' : locArr
            }
            dfMeasure = pd.DataFrame(data)

            #store the dataframe of the 9 locations in the final one
            dfFinal = pd.concat([dfFinal, dfMeasure], ignore_index=True)
```

Figure 4.2: Code documentation showing the data extraction from the netCDF4 files

Figure 4.3: Spearman Correlation Matrix Between All the Extracted Variables

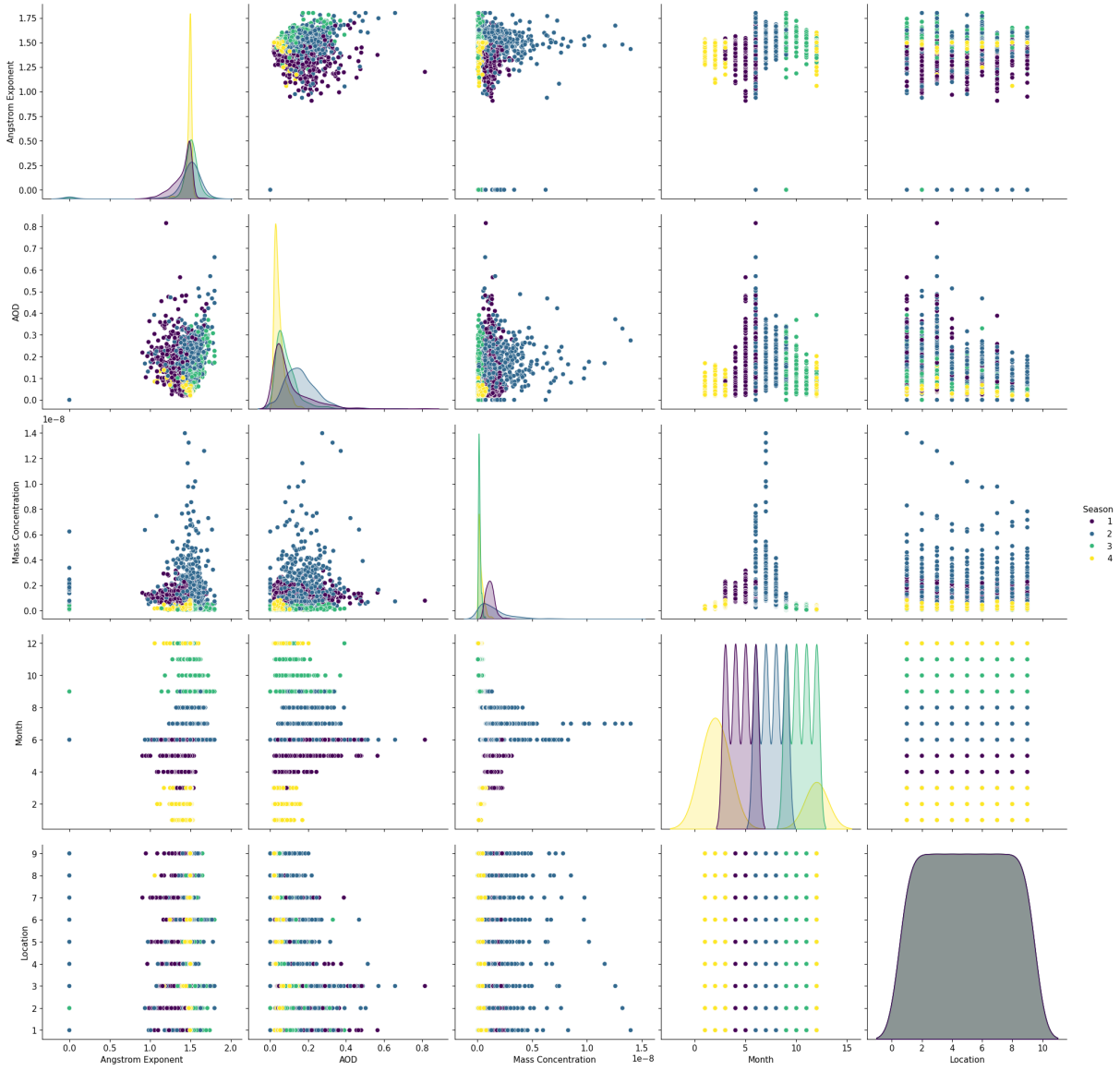


Figure 4.4: Scatter Plot Matrix Between All Extracted Variables. (1: Spring. 2: Summer. 3: Autumn. 4: Winter)

4.1.0.3 Third Stage

Once the structured and labeled database was generated, a regression model was proposed, focused only on the AE, which provides information on the size of the particle and therefore a measure of how small the particles present in the system are during certain season of the year. [63]

Dataset Design. The proposed regression model allows the prediction of the particle size according to the AOD, MC, the month, the season of the current year and the geographical location of the measurement. It is expected to predict the AE due to its direct relationship with the size of the particle. By knowing its size, you can know if they are "thin" and tend to be of anthropogenic origin, or if they are "thick" and are of natural origin. It was decided to consider all the dimensions of the observations and with feature engineering, to enrich the data set so that the model had more information in the characteristics matrix. The dataset design takes advantage of all the data collected and did not disregard information.

Regression Model. Once the dimensions were treated, it was decided to use a regression model called "Gradient Boosting". This was preferred due to its virtues in finding non-linear relationships. It also has the great advantage that the algorithm does not assume that the trend is normal. This helps in the project because, although there is a normal trend between some characteristics of the dataset, we must not forget that they are meteorological data so they can vary due to multiple environmental factors. The versatility of the regressor in the way of constructing weak learners and evaluating with different error functions allows us to better refine the strong learner and therefore better adjust the predictor to the context of the data. With gradient boosting, we seek to expand the scope of learning, taking into account the context of the observations, where the technique with which the dataset was designed, the sliding window, also helps.

Libraries Used for Development. To extract the satellite information, the netCDF4 library was used, which facilitated the extraction and manipulation of the data to organize them into a final data set. Once a way to extract the information from the HDF5 satellite files was found, "SciKit Learn" was mostly used with tools such as NumPy and Pandas for the organization of the data and the design of the database, including the sliding window algorithm. As per the representation and analysis of the information, Seaborn and Tableau were used. Both tools have several color palettes and clear ways to graph information. In Seaborn's case, methods such as "pairplot" were used, which crosses all the dimensions in pairs in scatter plots to graphically understand how one behaves with respect to the other. In the case of Tableau, the rapid interaction with the graphs to resolve doubts and propose hypotheses.

To divide the dataset into training and testing fragments, "train_test_split" was used

as the first attempt; however, since this is an analysis over time, a function was designed to split the dataset in an expected and orderly manner, that is, certain consecutive years of training and certain consecutive years of testing, in a sequential manner (Figure 4.5). Finally, to build the model, the tools "Pipeline", "SimpleImputer", "StandardScaler", "cross_val_score", "GridSearchCV" and "GradientBoostingRegressor" were used. All were tools to improve the entry and adaptation of the data to the regression model. The other tools served as a structure for the exploration and evaluation of the model. A pipeline was made with an imputator, scaler and the model as such. This pipeline was evaluated in a cross search with Grid Search, using some hyperparameters of the regressor such as: learning rate, number of estimators, max depth, loss and criterion (Figure 4.6). [63]

```python
def train_test_timeseries_split(X, y, test_size):
  featuresArr = X.values
  resArr = y.values
  test_size = round(len(featuresArr) * test_size)
  return featuresArr[0:len(X) - test_size], X[len(X) - test_size : ], resArr[0:len(y) - test_size], resArr[len(y) - test_size :]
```

Figure 4.5: Function to Split Time Series

```python
def setSpaceAndRegressor (regressor):
  space = dict()
  if regressor == 'GBR':
    regressor = GradientBoostingRegressor(random_state=42)
    space["log_model__learning_rate"] = [0.01, 0.001]
    space["log_model__n_estimators"] = [100, 300]
    space["log_model__max_depth"] = [4, 10]
    space["log_model__loss"] = ["squared_error", "huber"]
    space["log_model__criterion"] = ["friedman_mse", "squared_error"]
  elif regressor == 'XGBR':
    regressor = XGBRegressor(random_state=42)
    space["log_model__eta"] = [0.01, 0.001]
    space["log_model__n_estimators"] = [100, 300, 600]
    space["log_model__max_depth"] = [2, 4, 10]
    space["log_model__tree_method"] = ["hist", "exact", "approx"]
  elif regressor == 'ABR':
    regressor = AdaBoostRegressor(random_state=42)
    space["log_model__learning_rate"] = [0.01, 0.001]
    space["log_model__n_estimators"] = [100, 300, 600]
    space["log_model__max_depth"] = [2, 4, 10]
    space["log_model__loss"] = ["squared_error", "huber"]
    space["log_model__criterion"] = ["friedman_mse", "squared_error"]
  else:
    regressor = GradientBoostingRegressor(random_state=42)
    space["log_model__learning_rate"] = [0.01, 0.001]
    space["log_model__n_estimators"] = [100, 300, 600]
    space["log_model__loss"] = ["linear", "square", "exponential"]

  return regressor, space
```

Figure 4.6: Hyperparameters Set for each model

5. RESULTS AND DISCUSSION

Summary: *This chapter presents the results of each phase presented in the previous chapter and a discussion on the identification of aerosols, the proposed objectives and how they were achieved.*

5.1 Results

In this investigation, two results were obtained that alone can conclude relevant information. The first part of the project, where a set of specific satellite data of the MAG was generated, products generated by descriptive statistics were obtained. This tool allowed us to observe patterns in the behavior of the Ångström Exponent in relation to the season of the year and therefore, with certain months. The second part of the project was able to generate a prediction model for the Ångström Exponent with a certain margin of error. The results of the two phases are detailed below.

5.1.1 Descriptive Statistics Results

Descriptive statistics were made of three satellite measurements delimited for this investigation only in MAG. The behaviors of these measurements converge in a pattern directly correlated with the season of the year. The details of each statistic are set out below.

5.1.1.1 Ångström Exponent

The Ångström Exponent data collected with MODIS TERRA in the geographical area the MAG show the behavior of Figure 5.1, which consists of a two-dimensional graph of the average of the Ångström Exponent in each month of the year, separating by color the season of the year to which each month belongs. The exponent has no units of measurement because it is a coefficient.

Figure 5.1: Ångström Exponent Behavior (Average Per Month)

Table 5.1 shows the quantitative measurements obtained with the data graphed in Figure 5.1. It can be seen that August is the month with the largest average Ångström Exponent with 1.5874, while June is the smallest with 1.3291. However, it is also the month with the most variation according to its standard deviation of 0.2036. This can be interpreted as the month with diverse behaviors over 23 years. On the other hand, the month with the smallest deviation is March, with 0.0127, so it can be considered the most predictable month with respect to the Ångström Exponent.

Month	**Ångström Exponent Average**	**Standard Deviation**
January	1.4627	0.0411
February	1.4543	0.0584
March	1.4955	0.0127
April	1.4774	0.0233
May	1.3878	0.0652
June	1.3291	0.2036
July	1.5325	0.0552
August	1.5874	0.0595
September	1.5407	0.0723
October	1.5208	0.0251
November	1.4788	0.0284
December	1.4647	0.0766

Table 5.1: Ångström Exponent averages and standard deviation per month, from March 2000 to March 2023.

Figure 5.2 shows a graph of the seasonal average Ångström Exponent over the time period.

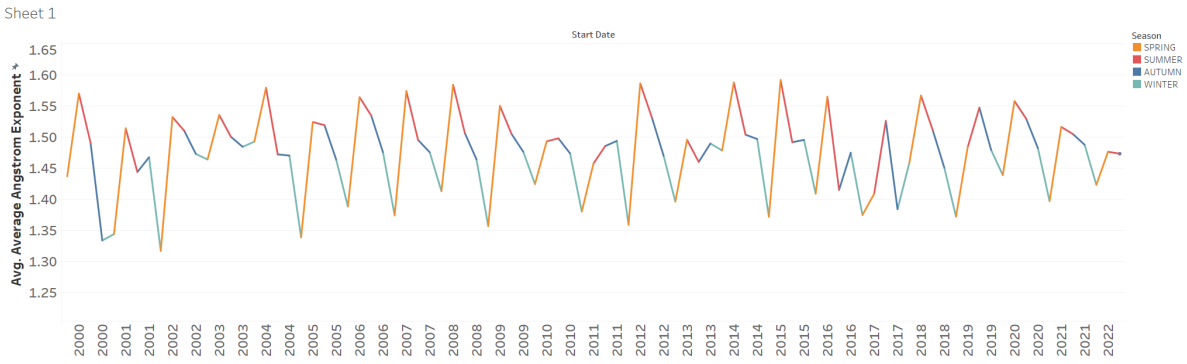


Figure 5.2: Ångström Exponent Behavior (Average per Season)

shows the quantitative information obtained with the data graphed in Figure 5.2. The seasons of the year are concentrated in this table. It can be seen that the season of the year with the largest Ångström Exponent in 23 years was summer, with 1.5349, while spring was the season with the smallest exponent, with 1.4001. The most stable and predictable

season was winter with a standard deviation of 0.0599. While the most dispersed was spring with a deviation of 0.1352.

Season	**Ångström Exponent Average**	**Standard Deviation**
Spring	1.4001	0.1352
Summer	1.5349	0.1101
Autumn	1.4982	0.0605
Winter	1.4669	0.0599

Table 5.2: Ångström Exponent averages and standard deviation per season, from beginning of spring 2000 to end of winter 2023.

5.1.1.2 Aerosol Optical Depth

The satellite data of the Aerosol Optical Depth (AOD) collected with MODIS TERRA of MAG show the behavior of Figure 5.3, which is monthly-averaged.

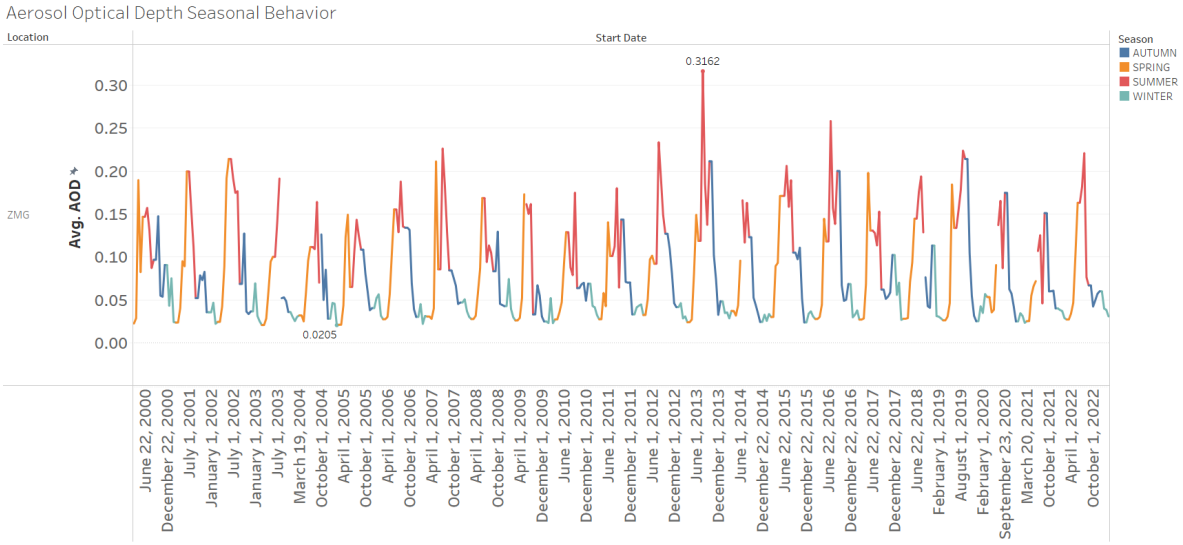


Figure 5.3: Aerosol Optical Depth Behavior (Average per Month)

Table 5.3 shows the quantitative measurements obtained with the data graphed in Figure 5.3. It can be seen that July is the month with the greatest optical depth with 0.1643, however, it is also the month with the greatest variation in the last 23 years with a standard deviation of 0.0571. It can also be observed that the months with the greatest optical depth are June, July, August and September, which are consecutive and belong to the summer season, which can be noted graphically in Figure 5.3, with the pattern of red peaks (summer) over the years. The month with the smallest average optical depth was March with 0.0293 and it was also the least dispersed month with a standard deviation of 0.0073. This can be seen graphically represented in Figure 5.3 with the lowest points on the graph, which belong to the end of winter and beginning of spring, in the month of March.

La Figura 5.4 shows a graph of the average AOD in each season of the year, represented by colors.

Table 5.4 shows the quantitative information obtained with the data graphed in Figure 4. The seasons of the year are summarized in this table. Highest average AODs are seen in the summer; lowest AODs occur in the winter, and have the smallest variability.

Month	AOD Average	Standard Deviation
January	0.0423	0.0123
February	0.0398	0.0133
March	0.0293	0.0073
April	0.0374	0.0146
May	0.0698	0.0344
June	0.1283	0.0401
July	0.1643	0.0571
August	0.1590	0.0337
September	0.1222	0.0494
October	0.0860	0.0305
November	0.0599	0.0158
December	0.0480	0.0207

Table 5.3: Aerosol Optical Depth averages and standard deviation per month, from March 2000 to March 2023.

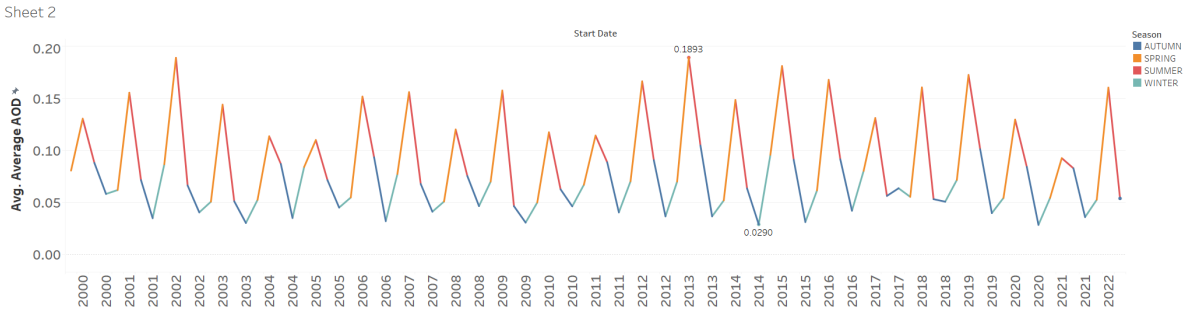


Figure 5.4: Aerosol Optical Depth Behavior (Average per Season)

Season	AOD Average	Standard Deviation
Spring	0.0655	0.0480
Summer	0.1467	0.0465
Autumn	0.0762	0.0402
Winter	0.0399	0.017

Table 5.4: Aerosol Optical Depth averages and standard deviation per season, from beginning of spring 2000 to end of winter 2023.

5.1.1.3 Mass Concentration

The satellite data of Mass Concentration (MC) collected with MERRA2 for MAG show the behavior of Figure 5.5, which consists of a two-dimensional graph of the average Mass Concentration in each month of the year. Separating by color the season of the year to which each month belongs.

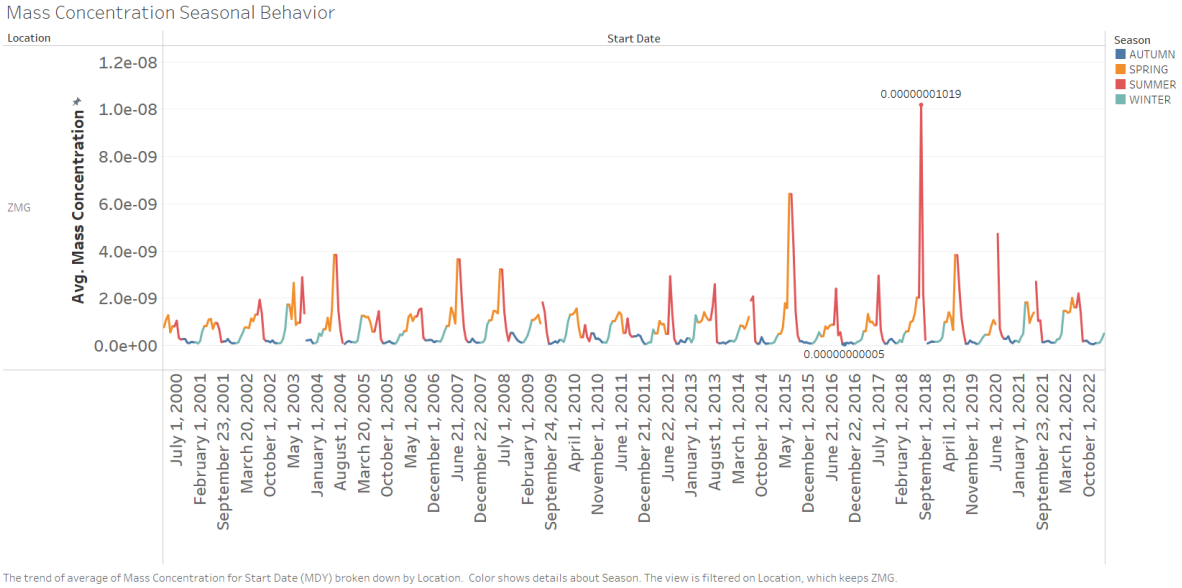


Figure 5.5: Mass Concentration Behavior (Average per Month in kg/m^3)

Table 5.5 shows the 23-year monthly averages of the same data in kg/m^3 units. It is observed that in July there is a higher mass concentration, 2.45E–09 kg/m^3. In general, the months from April to August (spring to summer) stand out with a considerably greater concentration than the rest of the year. This can be seen in Figure 5.5, where the peaks start orange (spring) and have their greatest extent in red (summer). December has smaller MC, 1.39E–10 kg/m^3. January is the least variable month, with a standard deviation of 5.07E–11 kg/m^3.

Figure 5.6 shows the seasonal-averaged mass concentration in each season of the year, represented by colors.

Table 5.6 shows the 23-year seasonal averaged MC. The seasons of the year are summarized in this table. We note that the greatest concentration occurs in summer with 1.44E–09 kg/m^3, and has the greatest dispersion, with a standard deviation of 1.52E–09 kg/m^3. On the other hand, the season with the lowest mass concentration is autumn, with 1.85E–10 kg/m^3 and, likewise, it is the season with the least dispersion over 23 years, with a standard deviation of 1.01E–10 kg/m^3.

Month	Mass Concentration Average (kg/m^3)	Standard Deviation (kg/m^3)
January	1.61E-10	5.07E-11
February	3.00E-10	1.08E-10
March	7.66E-10	3.57E-10
April	1.13E-09	2.48E-10
May	1.27E-09	4.07E-10
June	1.44E-09	1.17E-09
July	2.45E-09	1.95E-09
August	1.14E-09	6.11E-10
September	2.64E-10	1.92E-10
October	2.16E-10	1.09E-10
November	1.84E-10	7.72E-11
December	1.39E-10	5.38E-11

Table 5.5: Mass Concentration averages and standard deviation per month, from March 2000 to March 2023.

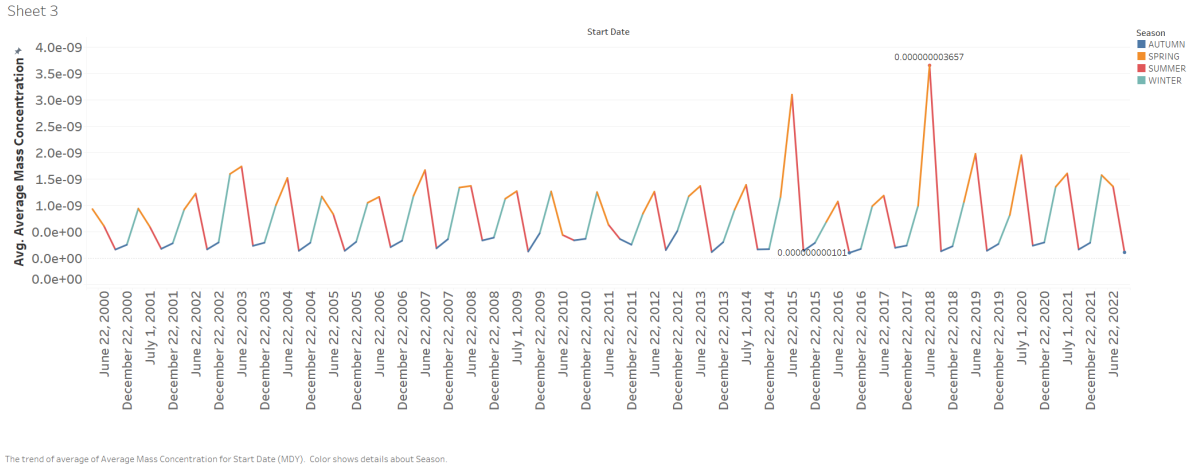


Figure 5.6: MC Behavior (Average per Season in kg/m^3)

Season	Mass Concentration Average (kg/m^3)	Standard Deviation (kg/m^3)
Spring	1.10E-09	3.69E-10
Summer	1.44E-09	1.52E-09
Autumn	1.85E-10	1.01E-10
Winter	3.01E-10	2.32E-10

Table 5.6: Mass Concentration averages and standard deviation per season, from beginning of spring 2000 to end of winter 2023.

5.1.2 Machine Learning Model Results

Three ways of approaching the problem were proposed to compare the difference and validate if the initial assumptions of the best way to organize the data set actually improve the results of the predictive model. [63]

5.1.2.1 First Proposal: Naive

In this first proposal, the dataset was taken as any Machine Learning problem, where for each set of characteristics or dimensions, a result is labeled. In this case, each observation was marked with a measurement of the Ångström Exponent. This proposal is referred as naive because it does not take into account any history of the previous measurements, to put it in a simple way, the model only learns that with certain characteristics, without taking into account what has happened before and throws an Ångström Exponent. In Figure 5.7, the data used to train the model is represented in blue. The orange color shows the Ångström Exponent actually captured by the satellite and the green color shows the forecast that the model made. It can be seen that the model follows the trend in almost all intervals, however, every certain period, there is a considerable difference, where the prediction is at least one unit away from reality. This in other words could predict that there will not be as much fine particle pollution, but in reality there will be. So it could lead to an erroneous recommendation in a practical case, a false negative. Table 5.7 shows the error calculated with this model. You can see a square and absolute error graduated in hundredths and thousandths, but since these measurements are for microscopic particles, they have a greater impact. [63]

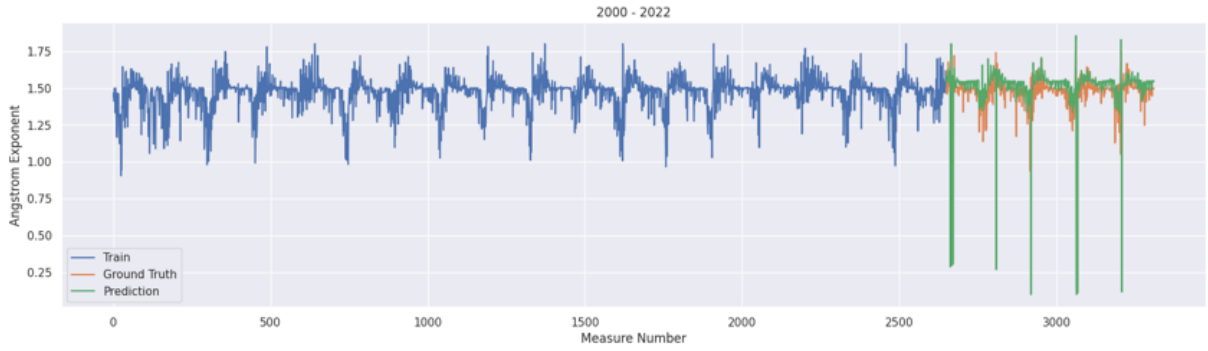


Figure 5.7: First Regressor Proposal: Naive

Score	Training Data	Testing Data
$\mathbf{R^2}$	0.86068464	0.86028673
MSE	0.00327475	0.00326522
MAE	0.03675071	0.03680163

Table 5.7: Naive Regressor Performance

5.1.2.2 Second Proposal: MAG Sliding Window

In this approach, a 16-space window was used, but only taking into account the data from the MAG. The dataset for this approach ended up having 49 dimensions (columns) and 352 observations (rows) [64]. These were expanded because each window space has a historical measurement of previous days. This model obtained the results shown in Figure 5.8, where it can be seen that there are still predictions that are lower than what actually happened. The differences were no longer marked in each time period. This is because in this model, only the measurements of the MAG and not its surroundings were taken into account. This removes the noise that certain data could introduce into the prediction, but in the same way data can be lost that can be correlated with the specific behavior of the MAG. The trend is equalized and the error obtained is observed in Table 5.8. These are very similar with the first proposal, however the changes are by ten-thousandths of a unit: 0.00000989 difference in this second model in the test data. Again, since we are talking about microscopic particles, it is considered an improvement. [63]

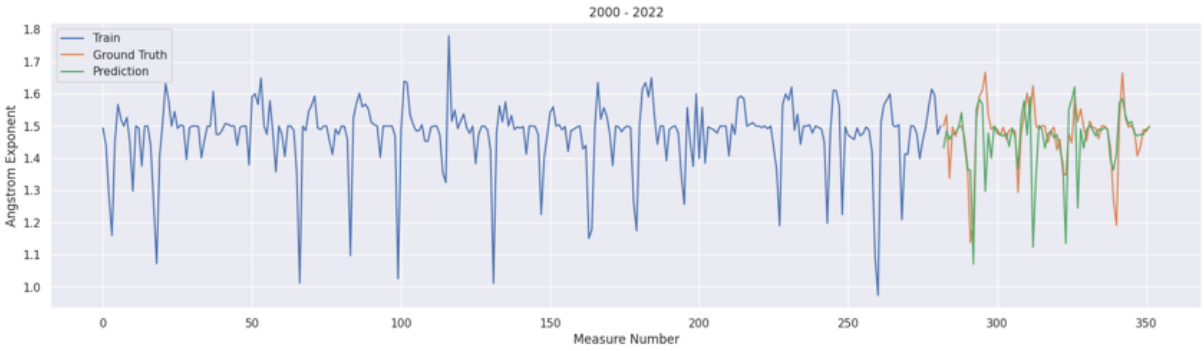


Figure 5.8: Second Regressor proposal: Feature Engineering with MAG data

Score	Training Data	Testing Data
$\mathbf{R^2}$	0.86080858	0.86081928
MSE	0.00326707	0.00327511
MAE	0.03675415	0.03671479

Table 5.8: Second Regressor Performance

5.1.2.3 Third Proposal: Sliding window of all Locations

In this third and final proposal, a form was chosen in which all the data collected over the years from all MAG locations would be included. A 16-space sliding window was also tried, resulting in a total of 351 observations. However, what must be highlighted is the number of columns, now there are 435 dimensions or characteristics because each location has its own characteristics and all were entered in a space of the sliding window. So, if the window had a single space and a single dimension, there would be at least 9 dimensions because there are 9 locations. [65]

The results are represented in Figure 5.9. It can be seen that we return to the error of the first proposal, where the regressor returns measures that are considerably lower than reality. However, it can be seen that some peaks are also better approximated. In Table 5.9 you can see the error of this model with respect to the Ångström Exponent that was actually recorded. Regarding the squared error of the previous proposal in the evaluation, this proposal worsened by 0.00000784. This could be due to the fact that the areas around the MAG change a lot and instead of providing information, they provide noise, so it would be necessary to analyze those measurements that actually provide information to the model and discard the others.

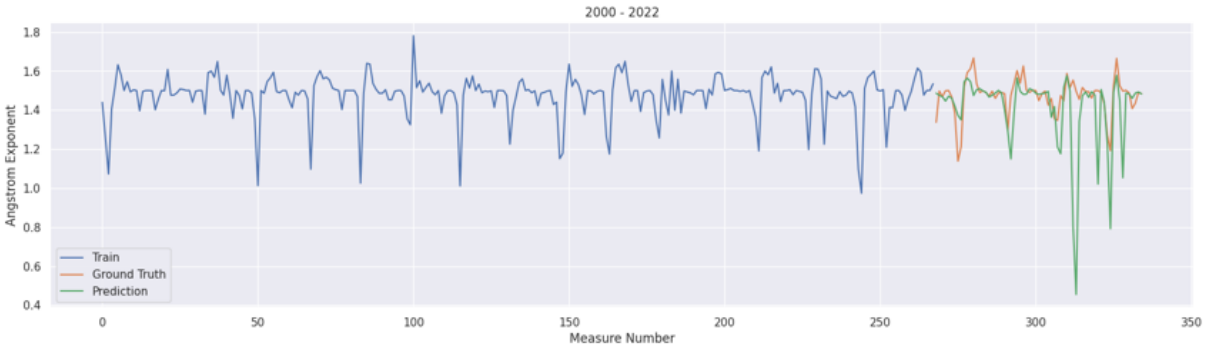


Figure 5.9: Third Regressor Proposal: MAG and its surrounding feature engineering.

Score	Training Data	Testing Data
$\mathbf{R^2}$	0.86080013	0.8601767
MSE	0.00326179	0.00326727
MAE	0.03689758	0.03682286

Table 5.9: Third Regressor Performance

5.2 Discussion

The proposed hypotheses were validated because based on the observation of MAG satellite data, a pattern was identified in the behavior of particle size. This behavior could be used in a Machine Learning regression model and a forecast of the particulate matter of the MAG was achieved.

6. CONCLUSIONS

Summary: *This chapter presents the conclusions of each phase of the project and future work in relation to the development of better machine learning models to improve the certainty of the forecasts and make them more helpful.*

6.1 General Conclusions

Similarly to the results, two specific conclusions from the project stand out. That of the first part of the descriptive statistics, which in turn served as the basis for the development of the model developed and tested in the second part of the research. Regarding the project in general, it is concluded that satellite measurements provide valuable information that needs to be collected, cleaned and observed. These studies can be achieved with the publicly accessible information that NASA publishes daily and can be the motivation for investment in projects of greater scope and with greater precision, where the final goal continues to be public health, the care of the flora and fauna and the taking of preventive and corrective measures against environmental phenomena caused by anthropogenic aerosols.

6.1.1 Descriptive Statistics Conclusion

It is concluded that summer was the season of the year with the largest Ångström Exponent, Aerosol Optical Depth and Mass Concentration during the period from spring 2000 to the beginning of spring 2023. This corresponds to fine particulate matter of anthropogenic origin. In the graphical aspect, figures 5.1 to 5.6 shown show an outstanding pattern in the behavior of aerosols from the perspective of the measurements studied. The sequence begins in spring, where the three selected characteristics increase progressively until reaching summer, where the maximum value of the current year is found. Subsequently, it is observed that the trend is to decrease in autumn and reach the lowest point of the year in winter, and then return to the midpoint in spring and repeat the pattern. This research confirms that the seasonal behavior of the ZMG follows a sequence seen for 23 years, so it is possible to take advantage of this pattern and correlate the atmospheric phenomena seen in the corresponding seasons of the year with the data presented here, trying to mitigate them. and ideally correct them. These conclusions were synthesized and submitted to the IEEE International Geoscience and Remote Sensing Symposium (IGARSS). The paper proposal was accepted and its cover is exposed in appendix B, which is going to be published on the last trimester of 2024. [66]

6.1.2 Machine Learning Regression Model Conclusion

With the proposed regressor, it was possible to approximate the trend of the observations and bring the predictions closer to a considerable error. The data extracted over 23 years were used in their entirety and patterns and trends were seen with respect to the seasons and the months of the year. The model with the best results obtained an accuracy of 86.08% in the R^2 metric, an average square error of 0.0032 and an average absolute error of 0.0367. Although the results of the errors in the three proposals are similar, as mentioned above, microscopic particles are being treated and the "small" differences become relevant.

In general, we achieved the objective of obtaining a regressor that approximated the measurements of the Ångström Exponent, certain enough to be considered as a first step towards a pollution traffic light in the MAG. Likewise, by having a tool that allows us to identify fine particles from coarse ones, campaigns can be justified that allow us to reduce the finer particles, which are more harmful to humans, because they can reach the alveoli and cause respiratory diseases, causing greater morbidity and mortality in the exposed population.

6.2 Future Work

It is likely that there are still more Machine Learning or even Deep Learning models that consider more factors and result in a more accurate approximation, but these models would be based on the collection and observation of the historical satellite measurements of the MAG collected in this research. The Machine Learning models proposed in this research can continue to be trained with data from subsequent years until today. Likewise, we would seek to find a model that has better precision with respect to the results of the best model obtained here. An R^2 of 86.08%, MSE of 0.0032 and MAE of 0.0367. These models would make it possible to predict the risk areas of the MAG and make decisions with less uncertainty, including taking preventive measures with greater confidence. More boosting models can be proposed to compare performance and also fine-tune the hyper parameters to try to generalize the behavior and reach a model that is generic enough to allow reliable predictions.

The seasonal patterns of particulate matter obtained in this study is knowledge that can help future research on the physical and optical properties of clouds formed by PM nuclei of anthropogenic origin that will allow us to understand changes in rainfall during the season. of rains that have been generated in the MAG.

BIBLIOGRAPHY

[1] B. N. Holben, D. Tanré, A. Smirnov, *et al.*, "An emerging ground-based aerosol climatology: Aerosol optical depth from AERONET," *Journal of Geophysical Research Atmospheres*, vol. 106, no. D11, pp. 12 067–12 097, Jun. 2001, ISSN: 01480227. DOI: 10.1029/2001JD900014.

[2] Medio Ambiente y Desarrollo Territorial, *SEMADET*. [Online]. Available: https://aire.jalisco.gob.mx/.

[3] D. A. Chu, Y. J. Kaufman, G. Zibordi, *et al.*, "Global monitoring of air pollution over land from the Earth Observing System-Terra Moderate Resolution Imaging Spectroradiometer (MODIS)," *Journal of Geophysical Research: Atmospheres*, vol. 108, no. 21, Nov. 2003, ISSN: 01480227. DOI: 10.1029/2002jd003179.

[4] K. A. Jalal, A. Asmat, and N. Ahmad, "Aerosol Optical Depth (AOD) retrieval method using MODIS," in *International Conference on Space Science and Communication, IconSpace*, vol. 2015-September, IEEE Computer Society, Sep. 2015, pp. 370–374, ISBN: 9781479919406. DOI: 10.1109/IconSpace.2015.7283802.

[5] D. Tanré, E. Vermote, B. N. Holben, and Y. J. Kaufman, "Satellite aerosols retrieval over land surfaces using the structure functions," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 2, Institute of Electrical and Electronics Engineers Inc., 1992, pp. 1474–1477, ISBN: 0780301382. DOI: 10.1109/IGARSS.1992.578487.

[6] C. Ichoku, D. A. Chu, S. Mattoo, *et al.*, "Techniques of global validation of aerosol retrievals from MODIS," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 3, 2001, pp. 1203–1205. DOI: 10.1109/igarss.2001.976792.

[7] I. E. Villalon-Turrubiates, G. E. Faus-Landeros, and E. A. Celarier, "SATELLITE MEASUREMENTS OF THE ANGSTROM EXPONENT USING AN INNOVATIVE MATHEMATICAL METHOD TO IDENTIFY SEASONAL AEROSOLS," Tech. Rep., 2009.

[8] J. Jiang, J. Liu, D. Jiao, and S. Cao, "Analysis of Aerosol Types and Sources Over Taihu Lake Based on Aeronet Data," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 2022-July, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 6698–6701, ISBN: 9781665427920. DOI: 10.1109/IGARSS46834.2022.9883423.

[9] C. Krittanawong, Y. K. Qadeer, R. B. Hayes, *et al.*, "PM2.5 and Cardiovascular Health Risks," *Current Problems in Cardiology*, vol. 48, no. 6, p. 101 670, Jun. 2023, ISSN: 0146-2806. DOI: 10.1016/J.CPCARDIOL.2023.101670.

[10] Google Cloud, *¿Qué es el aprendizaje automático?* [Online]. Available: https://cloud.google.com/learn/what-is-machine-learning?hl=es-419%20https://cloud.google.com/learn/what-is-machine-learning#section-7.

[11] A. Ackerman, A. M. da Silva, T. Eck, *et al.*, "Aerosol Properties in Cloudy Environments from Remote Sensing Observations: A Review of the Current State of Knowledge," *Bulletin of the American Meteorological Society*, vol. 102, no. 11, E2177–E2197, Nov. 2021, ISSN: 0003-0007. DOI: 10.1175/BAMS-D-20-0225.1. [Online]. Available: https://journals.ametsoc.org/view/journals/bams/102/11/BAMS-D-20-0225.1.xml.

[12] J. McNeill, G. Snider, C. L. Weagle, *et al.*, "Large global variations in measured airborne metal concentrations driven by anthropogenic sources," *Scientific Reports*, vol. 10, no. 1, Dec. 2020, ISSN: 20452322. DOI: 10.1038/s41598-020-78789-y.

[13] Q. Xiao, H. Zhang, M. Choi, *et al.*, "Evaluation of VIIRS, GOCI, and MODIS Collection 6 AOD retrievals against ground sunphotometer observations over East Asia," *Atmospheric Chemistry and Physics*, vol. 16, no. 3, pp. 1255–1269, Feb. 2016, ISSN: 16807324. DOI: 10.5194/acp-16-1255-2016.

[14] G. Chen, J. Guang, Y. Xue, Y. Li, Y. Che, and S. Gong, "A Physically Based PM2.5 Estimation Method Using AERONET Data in Beijing Area," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 6, pp. 1957–1965, Jun. 2018, ISSN: 21511535. DOI: 10.1109/JSTARS.2018.2817243.

[15] M. Bilal and Z. Qiu, *EVALUATION OF MODIS C6 COMBINEDAEROSOL PRODUCT AT GLOBAL SCALE*. IEEE, 2018, ISBN: 9781538671504.

[16] L. A. Remer, T. ! Didier, Y. J. Kaufman, R. Levy, and S. Mattoo, "ALGORITHM FOR REMOTE SENSING OF TROPOSPHERIC AEROSOL FROM MODIS: Collection 005," Tech. Rep.

[17] Green Facts, *Partículas Gruesas*, Jul. 2023. [Online]. Available: https://www.greenfacts.org/es/glosario/pqrs/particulas-gruesas.htm#:~:text=Las%20part%C3%ADculas%20gruesas%20tienen%20un,part%C3%ADculas%20ultrafinas%20(PM0.1)..

[18] Green Facts, *Partículas Finas*, Jul. 2023. [Online]. Available: https://www.greenfacts.org/es/glosario/pqrs/particulas-finas.htm.

[19] APIS, *Air Pollution Information System*, 2016. [Online]. Available: https://www.apis.ac.uk/overview/pollutants/overview_particles.htm#:~:text=Dust%20particles%2C%20often%20referred%20to,the%20history%20of%20the%20particles..

[20] World Meteorological Organization, *WMO Aerosols Bulletin focuses on fires*, May 2021. [Online]. Available: https://wmo.int/media/news/wmo-aerosols-bulletin-focuses-fires.

[21] J. M. Prospero and O. L. Mayol-Bracero, "Understanding the transport and impact of African dust on the Caribbean Basin," *Bulletin of the American Meteorological Society*, vol. 94, no. 9, pp. 1329–1337, Sep. 2013, ISSN: 00030007. DOI: 10.1175/BAMS-D-12-00142.1.

[22] The Centre for Atmospheric Science is part of the Department of Earth and Environmental Sciences, *Black Carbon Aerosol*. [Online]. Available: http://www.cas.manchester.ac.uk/resactivities/aerosol/topics/black/.

[23] X. Lu, X. Zhang, F. Li, and M. A. Cochrane, "Investigating Smoke Aerosol Emission Coefficients Using MODIS Active Fire and Aerosol Products: A Case Study in the CONUS and Indonesia," *Journal of Geophysical Research: Biogeosciences*, vol. 124, no. 6, pp. 1413–1429, Jun. 2019, ISSN: 21698961. DOI: 10.1029/2018JG004974.

[24] H. M. Horowitz, C. Holmes, A. Wright, *et al.*, "Effects of Sea Salt Aerosol Emissions for Marine Cloud Brightening on Atmospheric Chemistry: Implications for Radiative Forcing," *Geophysical Research Letters*, vol. 47, no. 4, Feb. 2020, ISSN: 19448007. DOI: 10.1029/2019GL085838.

[25] M. H. Askariyeh, H. Khreis, and S. Vallamsundar, "Air pollution monitoring and modeling," *Traffic-Related Air Pollution*, pp. 111–135, Jan. 2020. DOI: 10.1016/B978-0-12-818122-5.00005-3.

[26] Met Office, *What is convection?* [Online]. Available: https://www.metoffice.gov.uk/weather/learn-about/weather/how-weather-works/what-is-convection.

[27] NAL Agricultural Thesaurus, *Deposición húmeda*, Aug. 2006. [Online]. Available: https://agclass.nal.usda.gov/es/vocabularies/nalt/concept?uri=https%3A//lod.nal.usda.gov/nalt/68066.

[28] R. Levy and NASA, *Aerosol Optical Depth*. [Online]. Available: https://earthobservatory.nasa.gov/global-maps/MODAL2_M_AER_OD.

[29] NASA, *What is a Satellite?* Sep. 2023. [Online]. Available: https://spaceplace.nasa.gov/satellite/en/.

[30] NASA and N. Rayne, *Sobre Terra*, Apr. 2024. [Online]. Available: https://terra.nasa.gov/about.

[31] NASA and N. Rayne, *ASTER*, Apr. 2024. [Online]. Available: https://terra.nasa.gov/about/terra-instruments/aster.

[32] NASA and N. Rayne, *CERES*, Apr. 2024. [Online]. Available: https://terra.nasa.gov/about/terra-instruments/ceres.

[33] NASA and N. Rayne, *MISR*, Apr. 2024. [Online]. Available: https://terra.nasa.gov/about/terra-instruments/misr.

[34] NASA and N. Rayne, *MOPITT*, Apr. 2024. [Online]. Available: https://terra.nasa.gov/about/terra-instruments/mopitt.

[35] NASA and N. Rayne, *MODIS*, Apr. 2024. [Online]. Available: https://terra.nasa.gov/about/terra-instruments/modis.

[36] NASA, L. Boisvert, P. Przyborski, and L. Oreopoulos, *Sobre Aqua*, Mar. 2024. [Online]. Available: https://aqua.nasa.gov/content/about-aqua.

[37] NASA, *VIIRS*, 2024. [Online]. Available: https://www.earthdata.nasa.gov/sensors/viirs.

[38] NASA and S. Bailey, *GOCI*, 2024. [Online]. Available: https://oceancolor.gsfc.nasa.gov/about/missions/goci/.

[39] NASA, K. Mohr, and R. Holland, *Deep Blue*, Apr. 2024. [Online]. Available: https://earth.gsfc.nasa.gov/climate/data/deep-blue%2004/16/2024.

[40] NASA, S. Pawson, and K. Patel, *MERRA2*, Sep. 2022. [Online]. Available: https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/.

[41] Gobierno de Jalisco, *Área Metropolitana de Guadalajara*, 2024. [Online]. Available: https://www.jalisco.gob.mx/es/jalisco/guadalajara.

[42] The National Science Foundation's National Ecological Observatory Network (NEON), *Hierarchical Data Formats - What is HDF5?* 2024. [Online]. Available: https://www.neonscience.org/resources/learning-hub/tutorials/about-hdf5#:~:text=The%20Hierarchical%20Data%20Format%20version,with%20files%20on%20your%20computer..

[43] Python Software Foundation, *Python*, 2024. [Online]. Available: https://docs.python.org/3/faq/general.html#what-is-python.

[44] NumPy Developers, *NumPy documentation*, 2024. [Online]. Available: https://numpy.org/doc/stable/.

[45] Pandas, "Pandas Package Overview," 2024. [Online]. Available: https://pandas.pydata.org/docs/getting_started/overview.html.

[46] GitHub, *netCDF4*. [Online]. Available: https://unidata.github.io/netcdf4-python/.

[47] Scikit-learn developers, *Scikit-Learn*, 2024. [Online]. Available: https://scikit-learn.org/stable/getting_started.html.

[48] Scikit-learn developers, *Pipeline*, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html.

[49] Scikit-learn developers, *Imputation of missing values*, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/impute.html#impute.

[50] Scikit-learn developers, *StandardScaler*, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[51] Scikit-learn developers, *GridSearchCV*, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[52] Scikit-learn developers, *GradientBoostingRegressor*, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegr
html.

[53] N. Flores, *Cross validation: qué es y su relación con machine learning*, Nov. 2022. [Online]. Available: https://blog.maestriasydiplomados.tec.mx/cross-validation-que-es-y-su-relacion-con-machine-learning.

[54] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, no. 1, p. 6086, 2024, ISSN: 2045-2322. DOI: 10.1038/s41598-024-56706-x. [Online]. Available: https://doi.org/10.1038/s41598-024-56706-x.

[55] Scikit-learn developers, *F1 Score*, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html.

[56] J. Frost, *Mean Squared Error (MSE)*, 2024. [Online]. Available: https://statisticsbyjim.com/regression/mean-squared-error-mse/.

[57] Statistics How To, *Absolute Error & Mean Absolute Error (MAE)*, 2024. [Online]. Available: https://www.statisticshowto.com/absolute-error/.

[58] I. Amazon Web Services, *¿En qué consiste el ajuste de hiperparámetros?* 2023. [Online]. Available: https://aws.amazon.com/es/what-is/hyperparameter-tuning/#:~:text=de%20vista%20computacional.-,%C2%BFQu%C3%A9%20son%20los%20hiperpar%C3%A1metros%3F,antes%20de%20entrenar%20un%20modelo..

[59] A. Aysha, *Sliding Windows in Pandas*, Mar. 2023. [Online]. Available: https://towardsdatascience.com/sliding-windows-in-pandas-40b79edefa34.

[60] JavaTpoint, *Sliding Window Algorithm*, 2021. [Online]. Available: https://www.javatpoint.com/sliding-window-algorithm.

[61] I. Amazon Web Services, "¿En qué consiste la ingeniería de características?," 2023. [Online]. Available: https://aws.amazon.com/es/what-is/feature-engineering/.

[62] National University, *Statistics Resources*, May 2024. [Online]. Available: https://resources.nu.edu/statsresources/Spearmans.

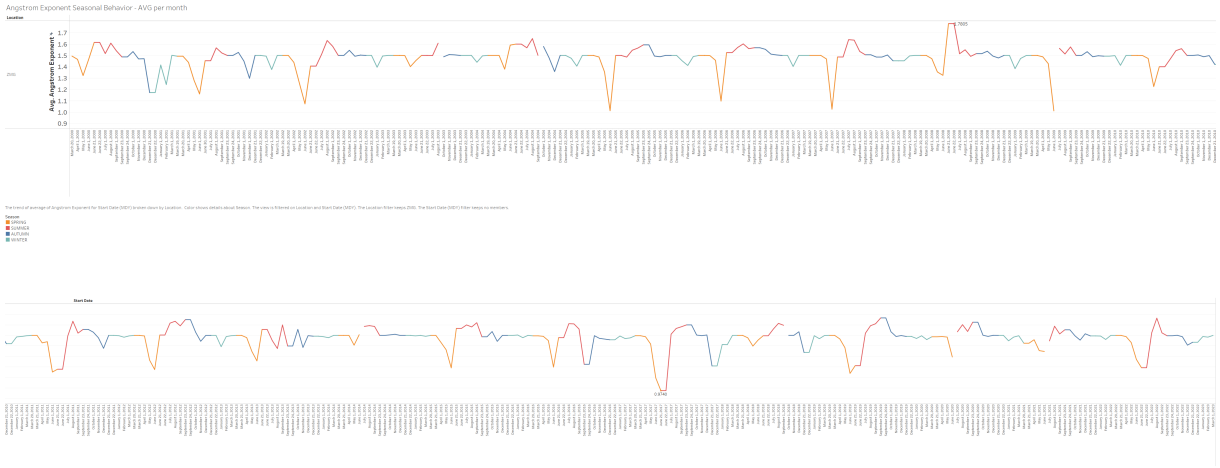
[63] F. A. Alavez Sosa, *Processing and Extraction Algorithm for Satellital Data of ZMG*, Jalisco, 2023. [Online]. Available: https://colab.research.google.com/drive/1XYZbytwTs31JNAqvjWp9NMjmu6fHk7Jh?usp=sharing.

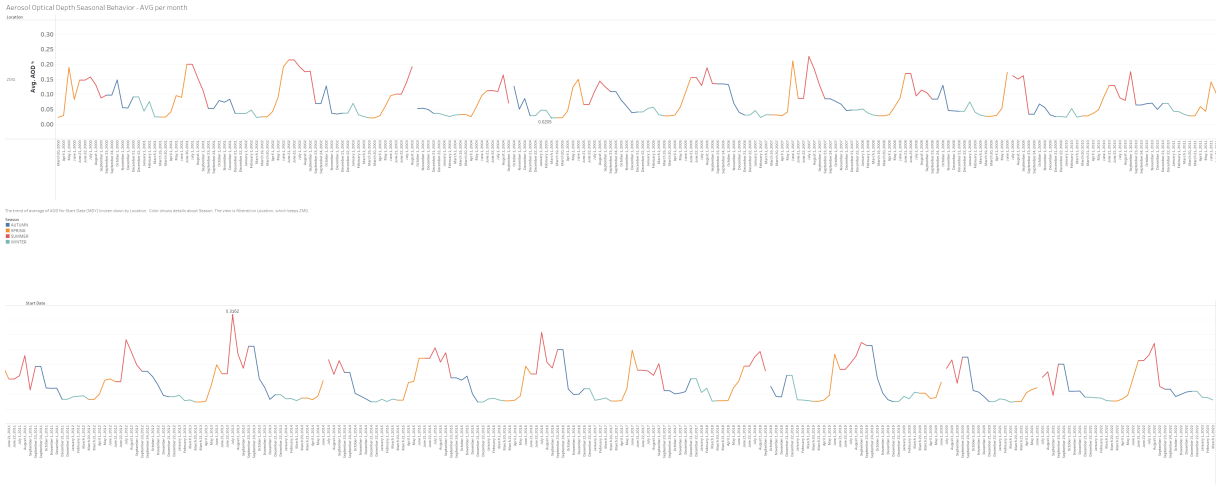
[64] F. A. Alavez Sosa, *Sliding Windows MAG*, Jun. 2024. [Online]. Available: https://colab.research.google.com/drive/1_Yz5DcqdjIj9lwyuia-3PFKeofHEpF3B#scrollTo=E9Ew1X7yw4zy&line=1&uniqifier=1.

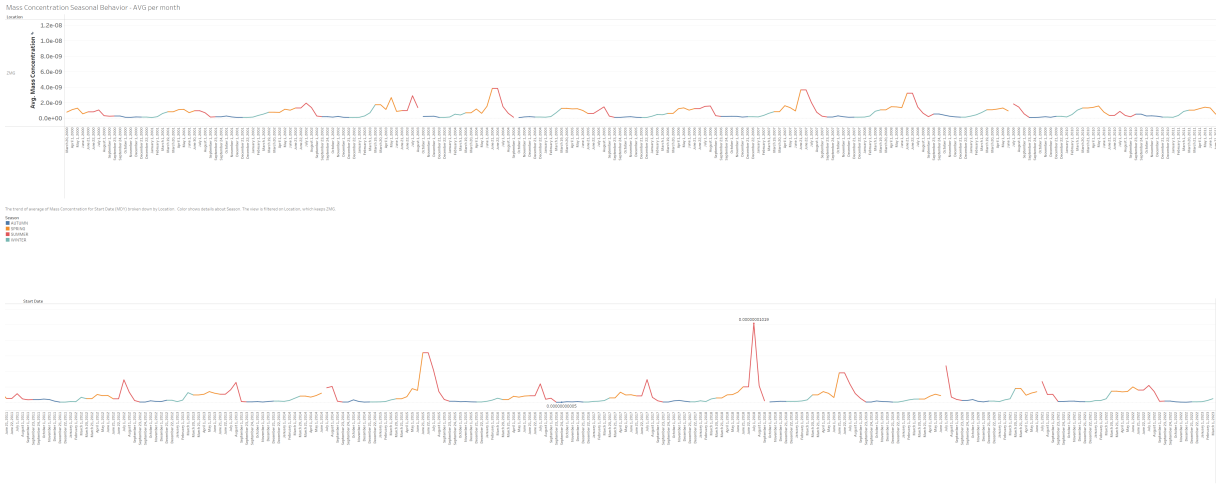
[65] F. A. Alavez Sosa, *Sliding Window All Locations (MAG and surroundings)*, 2024. [Online]. Available: https://colab.research.google.com/drive/1_Yz5DcqdjIj9lwyuia-3PFKeofHEpF3B#scrollTo=zwc9swDrslxt&line=1&uniqifier=1.

[66] IEEE and GRSS, *2024 IEEE International Geoscience and Remote Sensing Symposium*, 2024. [Online]. Available: https://2024.ieeeigarss.org/.

APPENDIX A. COMPLETE GRAPHICS OF THE BEHAVIOR OF THE MEASUREMENTS OBTAINED ON AVERAGE PER MONTH




Aerosol Optical Depth Seasonal Behavior - AVG per month
Location
Avg. AOD
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0295
The trend of average of AOD for Start Date (MDY) broken down by Location. Color shows details about Season. The view is filtered on Location, which keeps ZIAD.
Season
AUTUMN
SPRING
SUMMER
WINTER
Start Date
0.2162


Mass Concentration Seasonal Behavior - AVG per month
Location
Avg. Mass Concentration
1.2e-08
1.0e-08
8.0e-09
6.0e-09
4.0e-09
2.0e-09
0.0e+00
ZMG
The trend of average of Mass Concentration for Start Date (MDY) broken down by Location. Color shows details about Season. The view is filtered on Location, which keeps ZMG.
Season
AUTUMN
SPRING
SUMMER
WINTER

Start Date
0.000000010239
0.000000010035

APPENDIX B. PAPER SUBMITTED TO IEEE IGARSS (COVER ONLY).

MODELING THE BEHAVIOR OF SEASONAL AEROSOLS WITH 23-YEAR SATELLITE DATA IN THE GUADALAJARA METROPOLITAN AREA, JALISCO, MEXICO.

F. Alonso Alavez-Sosa[1], Gloria E. Faus-Landeros[2], *Senior Member IEEE,* Ivan E. Villalon-Turrubiates[3], *Senior Member IEEE,* Edward A. Celarier[4]

1 Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO), Mexico, francisco.alavez@iteso.mx
2 Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO), Mexico, faus.gloria@gmail.com
3 Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO), Mexico, villalon@ieee.org
4 Garud Technology Services, Inc. Columbia, MD, USA, edward.celarier@gmail.com

ABSTRACT

Satellite images have been used for more than 60 years to better understand the atmosphere, land, and water of our planet. In this study, a historical analysis was carried out for a period of 23 years of satellite images and specific measurements of 2.5 μm particulate matter of anthropogenic origin obtained through the NASA Giovanni website [1]. Atmospheric aerosols have a direct effect on the energy balance, cloud formation, and an indirect effect on rainfall, and consequently on flora and fauna regionally and globally [2]. In addition, they have a harmful impact on air quality, affecting the health of the population. These effects indicate the need for a quantitative study of this type of aerosols, especially in the cities of the world, which is where most sources of anthropogenic aerosols are located, mainly emitted by industries and vehicles. The Guadalajara Metropolitan Zone (GMZ), the third-largest metropolis in Mexico [3], The GMZ does not currently have ground-based measurements of particulate matter for the entire city. In this quantitative study, the seasonal behavior of aerosols was modeled. This allowed us to quantify and discover seasonal patterns in the geographic region of interest. This knowledge may contribute to future research related to the effects of aerosols in the GMZ and to the generation of an approximate forecast that allows preventive and corrective actions to be taken to reduce their harmful effects that affect the city and its inhabitants.

Keywords—particulate matter 2.5, Guadalajara Metropolitan Zone, satellite measurements, seasonal behavior of aerosols.

I. INTRODUCTION

Aerosols are particles suspended in the atmosphere, which can be solid or liquid. The complexity of its study lies in the fact that each particle has its own characteristics and reacts differently to the atmospheric and geographical phenomena of certain regions. In particular, the study of PM2.5 is essential to understand health impacts, prioritize pollution mitigation strategies, and provide input for global chemical transport modeling. Among the contaminants that have been found, metal species that come from anthropogenic sources stand out, such as lead, arsenic, chromium, and zinc, which often exceed the recommended health index in metropolitan areas. [4] For the observation and modeling of aerosol behavior, three aerosol measurements were selected: optical depth, Ångström exponent and mass concentration. Historical measurement data were obtained from March 2000 to March 2023 using the NASA's Giovanni web site. All the information was extracted, cleaned, analyzed, and concentrated through software developed in Python, which is meant to consume the daily satellite measurements and store it in such a way that facilitates their study [5].

II. DATA SOURCES AND SPECIFICATIONS

For this study, five significant features of the aerosols and their chronology were defined. The aerosol optical depth at 550 nm and the Ångström exponent at 0.412 to 0.47 microns were obtained with the Deep Blue algorithm, only for land, with the EOS-Terra MODIS instrument, in version MOD08_M3 v6.1. The concentration of the mass of particulate matter 2.5 microns was obtained using the MERRA2 product, version M2T1NXAER v5.12.4. Chronological characteristics are the month, year, and season of the year in which the measurement was taken. The geographic region considered to encompass the ZMG were 105W–100W, and 18N–22N. Satellite information was searched in time intervals predefined by the seasons of the year, from March 21, 2000 (beginning of spring) to March 20, 2023 (end of winter). Finally, data were downloaded in netCDF4 format, which stores data in HDF5 format. NASA's Giovanni platform calculates an average of the measurement over the selected time interval. Using these specifications, 368 individual MODIS observations were obtained.

III. ANGSTRÖM EXPONENT

The satellite-estimated Ångström exponents show the behavior of Figure 1, which presents the monthly average Ångström exponent throughout the year, separating by color the season of the year to which each month belongs.

Table 1 shows the quantitative measurements presented in Figure 1, in both, the exponent has no units of measurement because it is a coefficient. August is the month with the highest average Ångström exponent, 1.5874, while June has the lowest, 1.3291. However, it is also the month with the most variation according to its standard deviation of 0.2036. This can be interpreted as the month with diverse behaviors over 23 years. On the other hand, the month with the lowest standard deviation