Feasibility study of artificial intelligence techniques applied to the prediction of dust

by Víctor Galván Fraile

Universidad de La Laguna
June 2022

Facultad de Ciencias
Universidad de La Laguna

Feasibility study of artificial intelligence techniques applied to the prediction of dust

by
Víctor Galván Fraile

Supervised by:
Dr. Juan Pedro Díaz González (ULL)
Dr. Albano José González Fernández (ULL)

June 2022

"On the 16th of January (1833), when the Beagle was ten miles off the N.W. end of St. Jago (Cape Verde), some very fine dust was found adhering to the under side of the horizontal wind-vane at the mast-head; it appeared to have been filtered by the gauze from the air, as the ship lay inclined to the wind. The wind had been for twenty-four hours previously E.N.E., and hence, from the position of the ship, the dust probably came from the coast of Africa. The atmosphere was so hazy that the visible horizon was only one mile distant. During our stay of three weeks at St. Jago (to February 8th) the wind was N.E., as is always the case during this time of the year; the atmosphere was often hazy, and very fine dust was almost constantly falling, so that the astronomical instruments were roughened and a little injured. The dust collected on the Beagle was excessively fine-grained, and of a reddish brown colour; it does not effervesce with acids; it easily fuses under the blowpipe into a black or gray bead."

— Charles R. Darwin, *The Quarterly journal of the Geological Society of London [1846].*

Abstract

Feasibility study of artificial intelligence techniques applied to the prediction of dust

This end of degree project constitutes an introduction to the application of Machine Learning techniques on the prediction of meteorological variables, concretely, aerosols. It presents a bibliographic review of the role on atmospheric phenomena played by dust, including not only its main sources, but the fundamental production mechanisms as well. Furthermore, it presents a combination of theoretical concepts of Machine Learning algorithms, mainly based on the guidelines of the book "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems" [Gér19], and on the courses of "Machine and Deep Learning" of the University of Standford, taught online at Coursera [Ng22]. The aim of this project is, therefore, to realize a first approach to some of the basic algorithms of Machine Learning and put them into practice. Particularly, after a preprocessing phase of the data, two models with different artificial intelligence architectures were build up, training and testing them with different periods. Furthermore, a study of the input variables and the window sizes has been carried out in order to optimize the performance of the models. Finally, several analysis of the results obtained from them have been done, highlighting the strengths and weaknesses of each of them, in addition to suggesting the basis for future projects in this field. Additionally, and with the aim of increasing the transversality of this study, two dust intrusion classifying models have been made, describing not only their main characteristics, but also the results obtained and their possible improvements.

Keywords: Mineral aerosol, dust source, Machine Learning, Deep Learning, dust forecast.

Resumen

Estudio de la viabilidad del uso de técnicas de inteligencia artificial para la predicción de calima

El presente trabajo de fin de grado constituye una introducción a la aplicación de técnicas de aprendizaje automático para la predicción de variables meteorológicas, concretamente, aerosoles. En él, se presenta una revisión bibliográfica acerca del papel desempeñado por el polvo en procesos atmosféricos, sus principales fuentes y los mecanismos de producción fundamentales. Además, se presenta una combinación de conceptos teóricos sobre los que se fundamentan los algoritmos de inteligencia artificial, basados principalmente en las directrices del libro "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems" [Gér19], así como en los cursos de "Machine y Deep Learning" de la Universidad de Standford, realizados online a través de Coursera [Ng22]. El objetivo del proyecto es, por tanto, realizar una primera aproximación a alguno de los algoritmos básicos de aprendizaje automático y ponerlos en práctica. En particular, tras realizar un preprocesamiento de los datos, se ha procedido a la creación de dos modelos aplicando dos tipos de algoritmos diferentes, que han sido entrenados y testeados en diferentes periodos de tiempo. Asimismo, se ha realizado un estudio de las variables de entrada y de las ventanas de datos, con el objetivo de optimizar el rendimiento de los modelos. Finalmente, se ha realizado un análisis de los resultados obtenidos, destacando las bondades y defectos de cada uno de los modelos aplicados, además de sugerir las bases de futuros proyectos en este campo. Adicionalmente, y con objeto de aumentar la transversalidad de este estudio, se ha procedido a crear dos modelos clasificadores de las intrusiones de polvo, describiendo sus principales características, los resultados obtenidos así como las posibles mejoras de los mismos.

Palabras clave: Aerosol mineral, fuentes de polvo, aprendizaje automático, aprendizaje profundo, predicción de polvo.

Preface

This work carried out by the undergraduate student Víctor Galván Fraile constitutes his Final Degree Project at the Physics degree of the University of La Laguna. It has mainly been developed at the Faculty of Sciences under the framework of a collaboration scholarship of the MECD (Ministry of Education, Culture and Sport) carried out in the Earth and Atmosphere Observation Group (GOTA) of the mentioned institution, under the mentoring of Dr. Juan Pedro Díaz González and Dr. Albano José González Fernández.

This project constitutes an approach to the application of Machine Learning techniques in the prediction of meteorological phenomena. Particularly, it presents the key role played by dust on the global climate system and its impact on different fields such as meteorology, biology and energy production, among many others. Once the real concern of effectively predicting dust is introduced, a review of the state of the art of the main dust sources and its production mechanisms is presented, along with the ability of current models to predict its presence. Then, a revision of the basic concepts behind Machine Learning is shown, in addition to the main algorithms and techniques used, and finishing this part with a brief explanation of the most common evaluation metrics. Consecutively, a description of the meteorological variables used in this study is presented, as well as a preprocessing of them. Finally, different Machine Learning models (fully connected neuronal network and convolutional neuronal network) were trained and tested, varying their input variables, window sizes and training and testing periods. Therefore, the aim of this project is to settle the basis for the application of artificial intelligence techniques to the prediction of dust intrusions over susceptible lands, with special focus on the Canary Islands.

Acknowledgements

I would like to express my sincere gratitude to professors Juan Pedro Díaz González and Albano José González Fernández, because beyond their constant support and help throughout this adventure that began with a meeting more than a year ago, I have always felt appreciated within the excellent research group of which they are a part. Not to mention the long Tuesday afternoon meetings we have had, where we have gone through each and every one of the exciting results we have obtained through this project, and which always fly by. Thank you for making this little student a better scientist and, above all, a better person.

To my dear friends, fellows of happiness, frustration and lots of fun, my most sincere thanks for having contributed to my personal growth along these four years. Specially, I would like to thank Laura, Hugo, Ana, Elisa, Jorge and Leila, wish you the best for the time ahead.

Finally, to my mother, father and brother because they have always been there when I needed them, supporting and accompanying me throughout this wonderful life, thank you so much.

Table of Contents

List of Figures

List of Tables

Chapter 1

Global overview & state of the art

Resumen

En este capítulo se proporciona una breve introducción al concepto de polvo mineral, incluyendo las principales zonas productoras así como los procesos de formación de dichas partículas. A continuación, se describen tanto los mecanismos de levantamiento de polvo como los fenómenos meteorológicos que lo propician. Finalmente, se detallan los tipos de modelos numéricos utilizados actualmente para la predicción de polvo, sus principales características e inconvenientes.

1.1 Introduction

Mineral dust is a highly abundant atmospheric aerosol, accounting for about 35% of the total aerosol mass with diameter smaller than 10 μm [Van17]. It plays a key role in climate forcing by altering the overall radiation balance in the atmosphere through the scattering and absorption of radiation at both, solar (short-wave) and terrestrial (long-wave) portions of the electromagnetic spectrum. Mineral dust could also affect climate and meteorology by changing clouds formation and lifetime as well as precipitation processes, acting as droplet or ice condensation nuclei. Dust has also important implications regarding air quality and public health, causing respiratory, cardiovascular and infectious diseases [Gri01]. It is also a great source of iron (Fe), which has an effect on terrestrial and marine ecosystems. Several studies have highlighted the link of the Amazon rainforest productivity with the atmospheric deposition of dust emitted from the Saharan desert ([Lov10], [Swa92]). Therefore, efficient forecasting of the presence of dust is essential not only to alert the population of its danger, but also, for instance, to take into account its impact on the production of energy (i.e., photovoltaic energy).

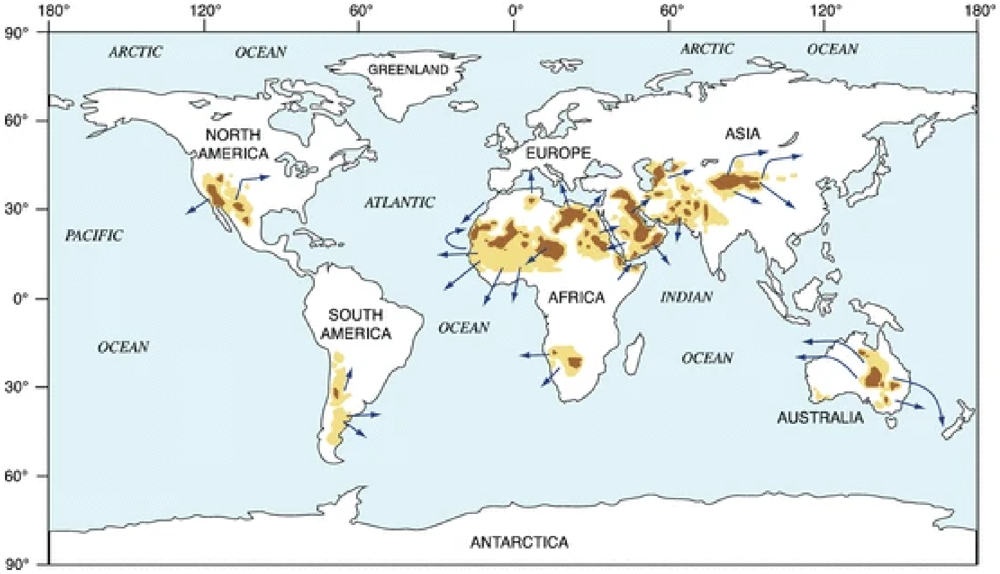


Figure 1.1: Map of global dust sources, based on multiple years of satellite imagery, derived from monthly mean frequency of occurrence (number of days per month) where the TOMS absorbing aerosol index (AAI) is greater than 0.7, using those months which best illustrate the configuration of specific dust sources. Dark brown indicate a range of 21-31 days and yellow indicate between 7 and 21 days. Additionally, blue arrows show typical dust transport pathways, based on interpretation of MODIS imagery from Terra and Aqua satellites [Kni14].

The World's most important dust sources are located around the so called "global dust belt", reaching from northern Africa across the Middle East and central Asia to northern India [Pro02] (see Figure 1.1). Although mineral aerosols are in essence natural particles emitted through natural mechanisms, part of their emissions are due to anthropogenic activities. Natural dust sources globally account for 75% of emissions, while anthropogenic sources yield the leftover. It has been estimated that the Saharan desert accounts for 55% of the global dust emissions [Feu18]. However, only 8% of them has anthropogenic origins, of which, the vast majority comes from the Sahel region [Gin12]. Specifically, the Bodélé Depression in Chad has been identified as the most active dust source in the Sahara, producing almost half of the mineral aerosols emitted from North Africa [WI09].

Due to the fact that physical and chemical characteristics of mineral dust are determined by their surface provenance, a huge effort has been made to identify dust source regions, resulting into different identification techniques.

The study of physical properties of dust deposits has been a great tool to identify aeolian sources for many decades, specifically, the study of loess, which is a sediment mainly formed by aeolian silt and clay [Muh14]. Other technique broadly used is the mineralogy of the clay, smaller than 2 μm, fraction. It has given great results when applied to the characterization of dust coming from the African continent, due to its geographic origin, with high illite contents at extratropical latitudes and high kaolinite contents in tropical latitudes [Muh14]. Linked to this technique appears the geochemical methods. However, both suffer significantly from temporal and spatial variability, which means that results may differ when measures are taken at different times or at different places. Another complementary technique to the previous ones is the study of the isotopic composition, mainly Neodymium (Nd), Strontium (Sr) and Lead (Pb), which suffers from the same drawbacks as the previous techniques. Advances in back-trajectory analyses, which basically means to analyze the atmospheric trajectories of parcels of air, have improved the results of the aforementioned dust source characterization techniques. Last but not least, the use of high-resolution Earth-orbiting satellite imagery appears, which has played a key role in the identification of contemporary dust sources. All these different approaches have concluded that desiccated or ephemeral lakes, high and low relief alluvial and aeolian systems, are the geomorphic settings most favourable to become dust sources [Pro02].

The formation of dust-sized mineral particles can be classified into four processes: volcanogenic, inheritance from fine-grained rocks, physical and chemical mechanisms of coarse particle reduction [Muh14]. Nevertheless, the two last processes are by far the most important ones. There is no doubt that wind erosion only occurs in areas where there is a sufficient supply of sand and dust. The formation of these sources is not only determined by aeolian transport, but by weathering and fluvial processes as well [Sha08]. It has been stated that almost all major present-day dust sources are located in arid or semiarid geomorphical features, particularly centered over topographic lows or lands adjacent to topographic highs [Pro02]. Fluvial processes had played a key role on the formation of present-day dust sources, because they are an effective producer of fine particles by pulling them apart from the soil matrix and leading to its deposition in basins or alluvial plains [Sha08]. Indeed, most dust sources are characterized by the presence of ephemeral rivers and streams, alluvial fans, playas and salt lakes [Pro02]. As mentioned above, the Bodélé Depression has been identified as the most persistent dust source of the Earth. Its origin, as most dust sources, comes from the Pleistocene (between 2 million to 10.000 years ago) when they were flooded and thick layers of sediments were deposited during the pluvial phases. Concretely, lake Mega-Chad occupied the region spanning from actual lake Chad to the northern edge of the Bodélé Depression. As the waters receded, the silts

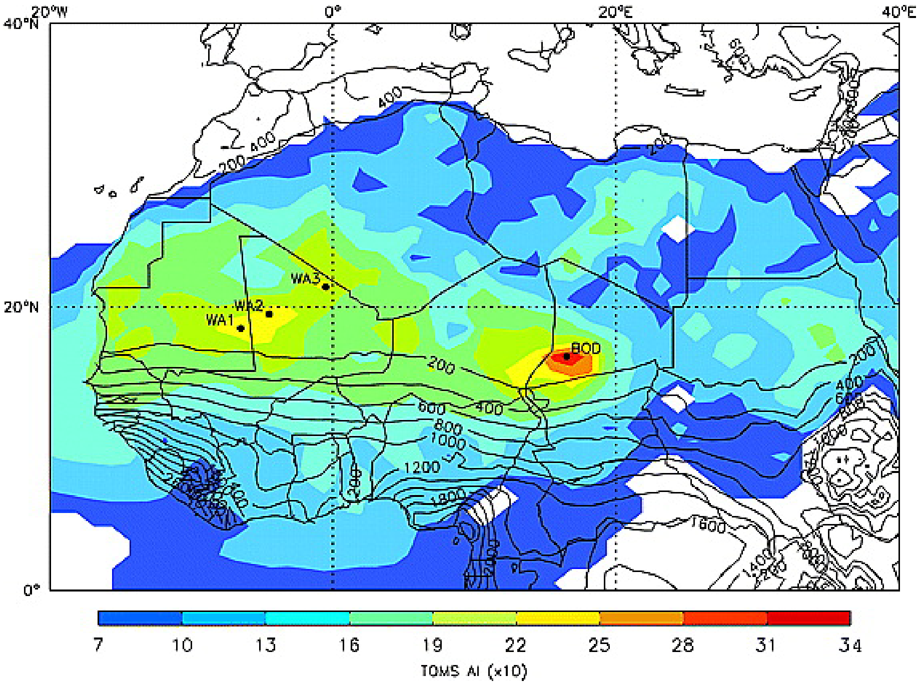


Figure 1.2: Long-term mean TOMS AAI (x10) over Africa north of the equator (filled contours) as well as long-term mean precipitation (black isohyets). West African major dust hot spots are indicated as WA1, WA2 and WA3; and the Bodélé Depression as BOD [Eng07].

and sediments resting on the lakebed were left to dry under the African sun [bod22]. These small grains of the silty sand are the main responsible for the constant dust production in the region. Another key characteristic of present-day dust sources is that they are almost all located in regions with annual rainfall under 200-250 mm [Pro02]. In Figure 1.2, the mean TOMS Aerosol Index, see [NAS22b] for more information about NASA's Total Ozone Mapping Spectrometer (TOMS), is represented as well as the isohyets over North Africa, highlighting the imaginary border of the 200 isohyet. The results below this line are not accurate enough due to the noise introduced by the biomass burning that occurs in the Sahel region in the winter months. However, as the most intense dust sources (called "hot spots") are located in regions of the Sahara desert where rainfall is very low, it is very likely that the annual dust cycle in these regions is, to a large degree, controlled by changes in near-surface winds [Eng07].

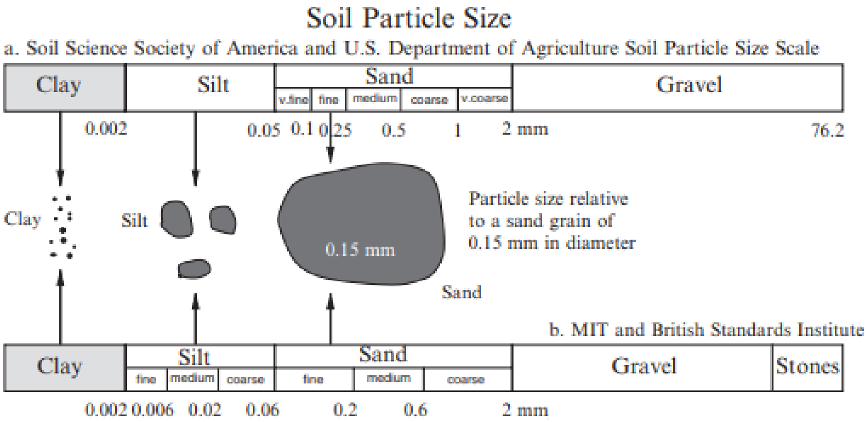


Figure 1.3: Comparison of different systems of particle-size definition [Sha08].

1.2 Dust production mechanisms

Having studied the processes of dust-sized mineral particles formation, let us continue by analyzing the mechanisms of dust production. But first, what is the difference between sand and dust? Well, both refer to solid inorganic particles that are derived from the weathering of rocks. While sand is defined as mineral particles with diameters between 62.5 and 2000 μm, dust are the particles with diameters smaller than 62.5 μm. The main difference between them is that dust particles can be readily suspended by wind, whereas sand particles are rarely suspended and are usually forming sand dunes and ripples [Kok12]. They are collectively called bedforms, which basically means that they form at the bottom of a basin at the contact between the sediment and the air. A schematic view of soil-size distribution is presented in Figure 1.3, where dust is formed by silt and clay particles.

Aeolian erosion is directly proportional to wind velocity. However, it only occurs when a certain threshold is exceeded, which is due to the fact that soil particles are roughly subjected to two forces that hold them to the surface: their weight and the interparticle cohesion forces. The last one is by far the most difficult to quantify, because it does not only depend on the type of particles (Van der Waals and electrostatic forces), but also on the presence of soil moisture (increasing the cohesive forces by increasing the presence of water). The existence of non-erodible roughness elements that can partially or totally cover the surface, could also increase the erosion threshold by absorbing a fraction of the wind momentum.

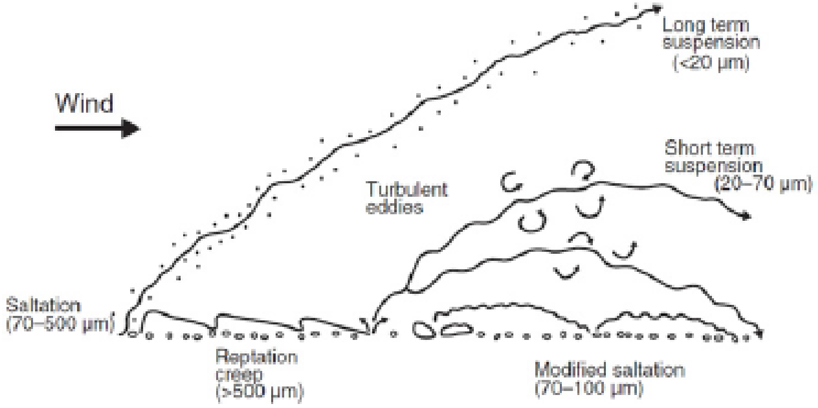

Wind
Long term suspension (<20 μm)
Short term suspension (20–70 μm)
Turbulent eddies
Saltation (70–500 μm)
Reptation creep (>500 μm)
Modified saltation (70–100 μm)

Figure 1.4: Scheme of the main modes of aeolian transport [Kok12].

Once this threshold is surpassed, the particles start to move. Their transport by the wind depend predominantly on particle size and wind speed, being controlled basically by the balance between their terminal fall and vertical air velocities. As wind speed increases, the first particles to be moved by the fluid drag are the ones with diameters near 100 μm. When lifted over the surface, these particles hop along it, in a process called saltation. It is crucial for aeolian transport because the impact of these saltating particles over the surface is strong enough to overcome the binding forces acting upon dust particles, leading to dust emission in a mode named sandblasting. The ejected particles, which can also be originated in a process called auto-abrasion (i.e., the disaggregation of the saltating soil grains), are susceptible to turbulent fluctuations and generally get involved in short-term (diameters between 20 and 70 μm) and long-term (diameters smaller than 20 μm) suspension [Kok12]. These long-term particles suspension are the responsible for dust storms that can travel hundreds or even thousands of kilometers away from their source regions. The saltating particles can mobilize particles with diameters bigger than 500 μm, however, due to their inertia, these particles do not saltate. Instead, they often do small hops before settling down in the soil, in a mode of transport called reptation. Alternatively, these big particles can roll along the surface pushed by the saltating particles as well as the wind drag forces. This last transport mode is named as creeping. See Figure 1.4 for a schematic view of the above described transport modes.

Having seen the main actual sources of dust as well as the processes of dust emission, we have concluded that wind plays a key role in the way that its speed must surpass a certain threshold to initiate dust ejection. Therefore, let us continue by describing briefly the atmospheric phenomena which generates these strong wind speeds, covering a wide variety of scales, including local, regional, synoptic and mesoscale. Dust devils stand out on the local phenomena, which are formed by turbulent circulations in the dry convective daytime planetary boundary layer (PBL) over deserts, specially during summertime [Kni14]. They can cause considerable dust emission taking into account their reduced size, varying from 10 to 100 meters of diameter. These characteristics make them very difficult to forecast and, therefore, to include them on dust weather models. At the regional scale, the main dust emitting phenomenon is the so called nocturnal low-level jets (NLLJs). They are horizontal winds produced by the build-up and decay of the PBL. With low surface winds taking place during the night, the near-surface air layers are well stratified and, consequently, there is absence of turbulence. Under these conditions, the air layers above the near-surface are frictionally decoupled from it and therefore, high wind speeds may take place. However, this stable stratified situation of the PBL is broken by convective turbulence produced by the solar heating of the surface, which takes place after dawn. With these air currents, the previous decoupled layer, becomes frictionally coupled to the surface and mixes down low-level jet momentum, reaching its most intense moment at midmorning (Figure 1.5 a). The breakdown of nocturnal low-level jets has been estimated to produce almost 60 % of total dust activity over the Sahara [Kok12], typically occurring under clear skies and low surface wind speed conditions with northeasterly trade winds (Harmattan flow). Cyclones govern on the synoptic-scale, and the corresponding mechanisms can be divided into first, the mobilization of dust by strong momentum from cyclonic surface winds and, second, its upward mixture to high altitudes by strong turbulence associated with cyclonic dynamics [Kok12]. This mechanism of dust uplifting is mainly observed over the Sahara desert in spring months, due to the presence of the highest temperature gradients between the North African coast and the Mediterranean Sea. Other relevant synoptic-scale phenomena are the so called African Easterly Waves (AEWs) as well as the lee-effect on the mountains (Atlas and Hoggar Mountains). Finally, on the mesoscale, the main dust emission is produced by deep moist convection and its associated downdrafts of cold and humid air (Figure 1.5 b). These events are mainly observed over Northwestern Sahara during the summer monsoon season, and they are commonly called haboobs. Orography plays a key role in the formation of this deep moist convection situation. The evaporative cooling of cloud particles and precipitation can form large scale density currents, which may be accompanied by high surface wind speeds and commonly, leading to dust emission [Kok12].

Emission of dust from North Africa follows a seasonal cycle, controlled by changes in the atmospheric phenomena discussed above, and the consequently changes in wind regimes. In winter (from November to

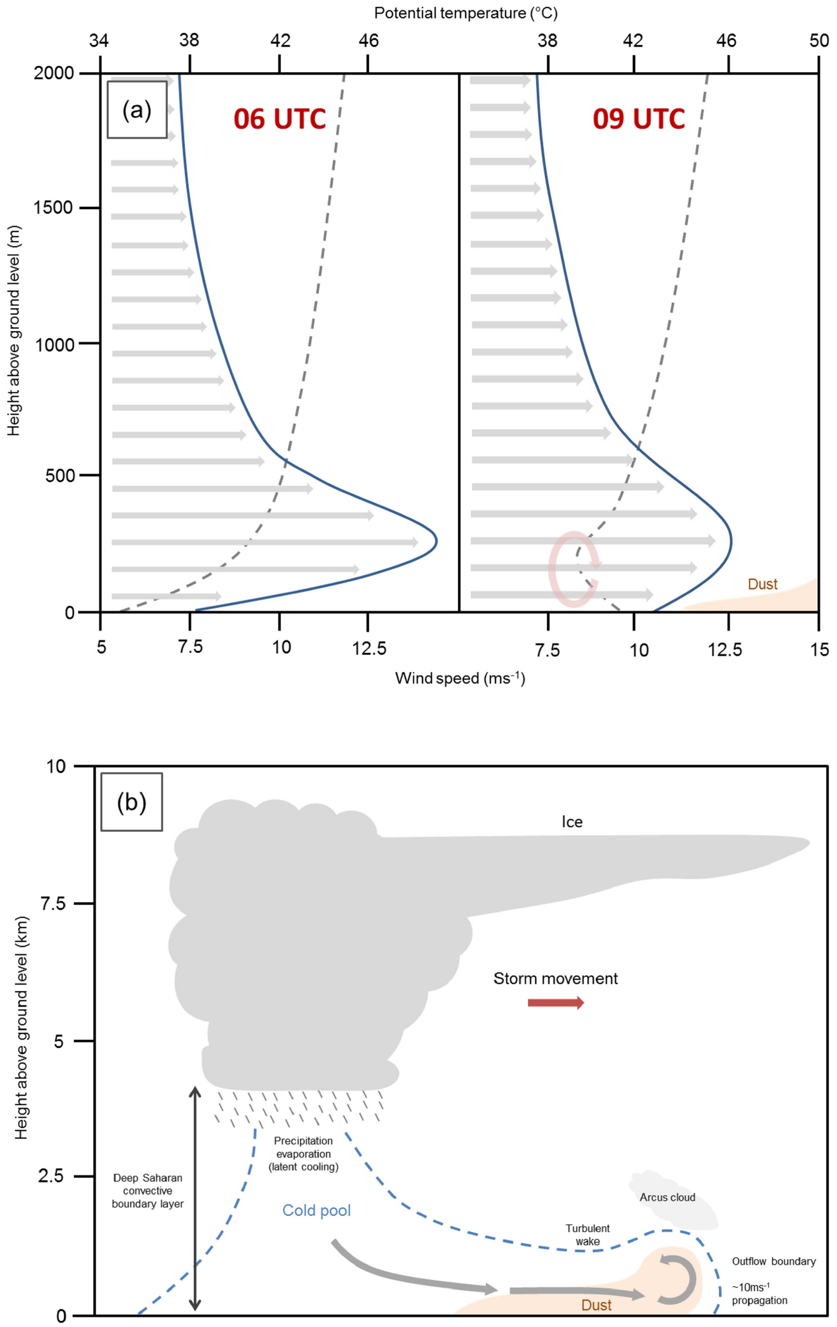


Figure 1.5: (a) This diagram shows the nocturnal low-level jet, including the typical summertime central Saharan wind (solid blue) and potential temperature (dashed gray) profiles at two different moments of the morning. (b) This illustration shows the mechanics of a cold pool outflow associated with a convective system [CHS19].

February) dust is mainly transported by the northeasterly trade winds from the north Saharan desert towards the Gulf of Guinea. This wintry transport often occurs at low-level altitudes (between 1.1 and 3.7 km on average [BA09]). Meanwhile, in summer, dust production becomes very active in western Sahara, covering central Mali, Mauritania and southern Algeria. During this season, dust is carried at higher altitudes (between 1.5 and 5.1 km on average [BA09]). In Figure 1.6, the mean TOMS aerosol index is depicted for each month, which highlights the seasonal pattern previously described.

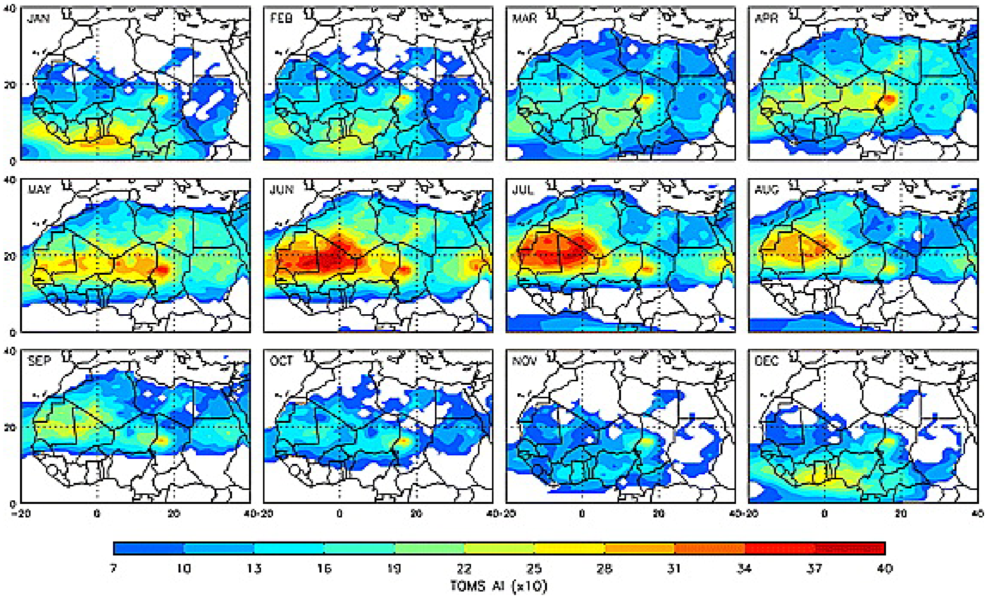


Figure 1.6: Long-term monthly average TOMS AAI (x10) over North Africa [Eng07].

1.3 Numerical dust models

Models of dust production, transport and deposition are not only used to forecast the presence of this aerosol in a certain place at a certain time, but also to study its effect on other meteorological and atmospheric phenomena like cloud formation and radiative forcing. These models can be divided into global models, which are commonly used to study the large-scale patterns of atmospheric dust loads, or regional models, which have a finer spatial resolution that enables them to capture better the synoptic and even regional scale dust emission processes discussed above. However, these models have troubles on representing small-scale features like moist convective events, what forces to make a parametrization to include them. Although they have much higher spatial resolution than global models, they are still unable to detect local events such as dust devils. Consequently, the emission of dust may be underestimated in both models. They can be improved by including soil surface information like surface roughness, sediment availability, vegetation cover, etc. Due to the remoteness of dust source areas, these data is mainly taken from satellite imagery, which highly depends on the spatio-temporal resolution of the satellite as well as the cloud cover of the study areas.

Chapter 2

Introduction to Machine Learning

Resumen

En este capítulo se muestra una sucinta introducción al aprendizaje automático, incluyendo los tres principales tipos de modelos y describiendo sus características. Seguidamente, se introducen tanto los árboles de decisión como las redes neuronales, explicando su estructura y sus aplicaciones. A continuación, se detalla la técnica del aprendizaje conjunto, distinguiendo los diferentes tipos y sus propiedades. Finalmente, se exponen las diferentes métricas utilizadas para evaluar el desempeño de cada modelo.

2.1 Machine Learning models

Machine Learning is defined as the field of study that gives computers the ability to learn without being explicitly programmed [Sam59]. The global scheme of a Machine Learning system is summarized in Figure 2.1. There are many different types of Machine Learning systems, which are commonly classified into three categories based on their supervision:

1. **Supervised learning**: The algorithm is fed with the features and their corresponding labels in a process called training. During this process, the algorithm gradually determines the relationship between features and their labels. This relationship is called the model, which is no more than mathematical functions that express the patterns between data and labels, that are acquired during training and will be afterwards used to make predictions on new data. Some examples of supervised learning algorithms are: Linear Regression, Logistic Regression, Support Vector Machines, Decision Trees, Random Forests and Neural Networks.

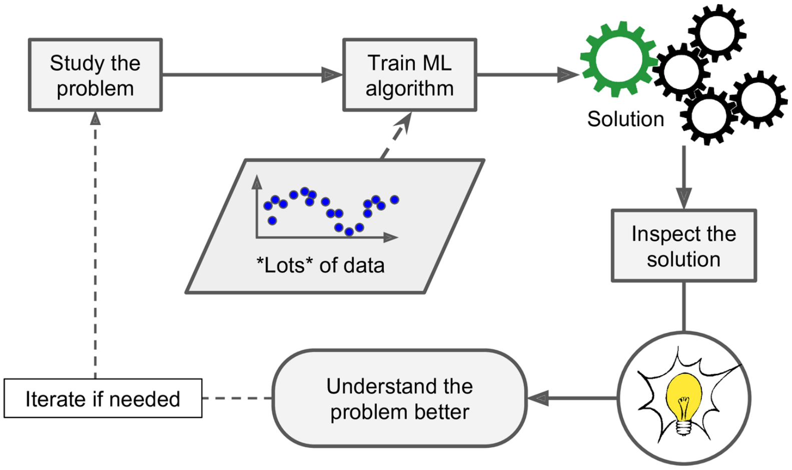


Figure 2.1: Outline of the Machine Learning approach [Gér19].

2. **Unsupervised learning**: The algorithm is fed with unlabeled data and its goal is to identify meaningful patterns on it. Some examples of unsupervised learning algorithms are: k-means and Principal Component Analysis.

3. **Reinforcement learning**: The learning system, called agent, observes the environment, select and perform actions, and get rewards or penalties in return. The goal is to learn by itself the best strategy, called policy, to get the most reward over time [Gér19]. They can outperform humans in playing games or even in driving.

There is a wide variety of Machine Learning applications, covering from agriculture and crops management [GF20], economics [Nym20] to meteorology and air pollution ([dG13],[Cho20], [GC21]), among many other fields.

2.1.1 Decision Trees

A decision tree is a flowchart structure in which each internal node represents a test on an attribute, each branch represents the outcome of the test and each leaf node represents a class label. In Figure 2.5, various decision trees inside another model, called Random Forest, are represented, whose structure will be later explained . Tree models not only work with discrete sets of target values (called classification trees), but also with continuous target values (called regression trees). They are commonly used due to their simplicity, flexibility and easy interpretable decisions. According to this last characteristic, they are called white box models. However, their simpleness could make them little useful for complex tasks.

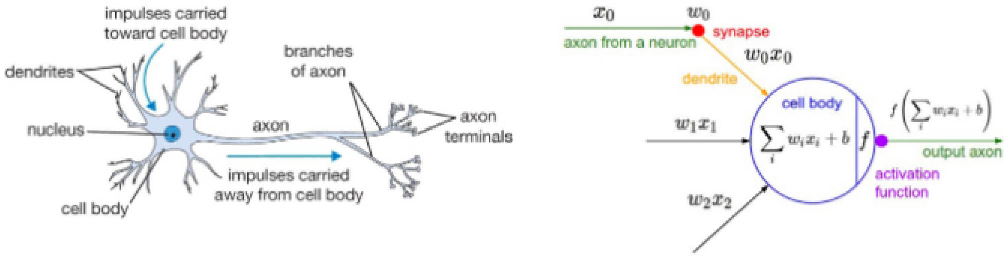


Figure 2.2: Schematic comparison of a human brain neuron and an artificial neuron [DP19].

2.1.2 Artificial Neural Networks

The initial idea behind Artificial Neural Networks (ANNs) was to design an algorithm able to mimic the nervous system of biological organisms. They are based on a collection of connected units, called artificial neurons. Each of these connections, called edges, can transmit a signal to other neurons, like synaptic connections in a biological brain. An artificial neuron receives a signal, process it and, likely, emits its own signal to other neurons connected to it. This signal is no more than a real number, and the output of each neuron is commonly computed by some non-linear function of its inputs. In Figure 2.2, a comparison between biological and artificial neurons is depicted. Neurons and edges have weights that are adjusted during training, and they vary the strength of the signal on a certain connection. Neurons are typically build up into layers, where the signal travels from the first one (called input layer) to the last one (called output layer). If the neural network has more layers than the input and output ones, which are named hidden layers, the neural network is then called a multi-layer neural network. In Figure 2.3, a diagram of a 1 hidden layer neural network is shown, including its forward propagation steps.

Basic unit

Considering a certain neuron, which has $\mathbf{x} \in \mathscr{R}^n$ as input values and $\mathbf{w} \in \mathscr{R}^n$ as weights, its output is given by

$$\mathbf{a} = f(\mathbf{wx} + b) \tag{2.1}$$

being $\mathbf{b} \in \mathscr{R}$ the bias value and f a non-linear activation function.

Activation functions

Selecting an appropriate activation function is one of the key parts when building an ANN. The power besides the use of non-linear activation functions is that a two-layer neural network can be used as an universal function approximator. The most common activation functions are:

- **Sigmoid**: The sigmoid activation function is useful for binary classification tasks, as it outputs a real value $f(\mathbf{z}) \in [0{,}1]$ following the relationship
$$f(\mathbf{z}) = \frac{1}{1+e^{-\mathbf{z}}} \tag{2.2}$$

- **ReLU**: The Rectified Linear Unit (ReLU) activation function has the advantage of being fast to compute, which is essential in training deep neural networks. Its expression is given by
$$f(\mathbf{z}) = \max(0,\mathbf{z}) \tag{2.3}$$

There are more complex activation functions, such as the hyperbolic tangent and the SoftMax (which is no more than a generalization of the sigmoid function used in the output layer of classification tasks), among many others.

Loss function and backward propagation

Having seen how the output of the ANN is computed (see Figure 2.3), let us continue by evaluating how well is this prediction. For this purpose, it is defined a loss function (also called cost function) over the training set and update the weights of the different layers by minimizing it. This process of update, called backpropagation, is based on the derivatives of the loss function. As with the activation, there are different types of loss functions and, depending on the particular task, one would be more appropriate than others. The most common ones are the Mean Squared Error (MSE) for regression tasks and the Cross-Entropy for classification tasks.

Let us sum up the whole training process: for each training instance, the algorithm first makes a prediction (forward pass), measures the error, then goes through each layer in reverse to measure the error contribution from each connection (backward pass), and finally slightly tweaks the connection weights to reduce the error [Gér19].

Dropout

Dropout is a commonly used regularization technique for deep neural networks, which aims to reduce the overfitting of the training set. It is based

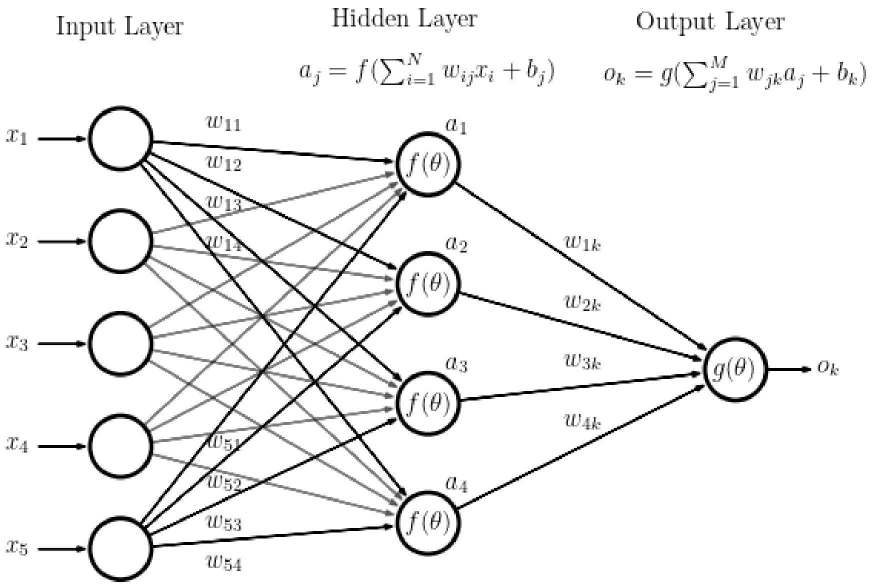


Figure 2.3: Architecture of a 1 hidden layer artificial neural network. The non-linear activation functions between layers are shown which, in a whole, represent the feedforward process [DP19].

on "dropping out" some neurons at every training step with a probability given by the hyperparameter p, which is called the dropout rate [Gér19]. It can omit units from both, input and hidden layers, meaning that those units will be ignored during the training step. However, when training has ended, neurons do not get dropped anymore.

2.1.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of neural network that emerged from the study of the visual cortex of the brain, which have been widely used not only in image recognition, but in natural language processing and time series analysis as well. They are multi-layer neural networks with, at least, one convolutional layer. In Figure 2.4, a basic CNN architecture is shown. The key idea behind this type of layers is that each neuron is not connected to every single neuron in the previous layer (like on usual ANNs), but only to those in its receptive fields. This architecture allows the network to focus on low-level features at each layer, taking into account the spatial relationships between separate features.

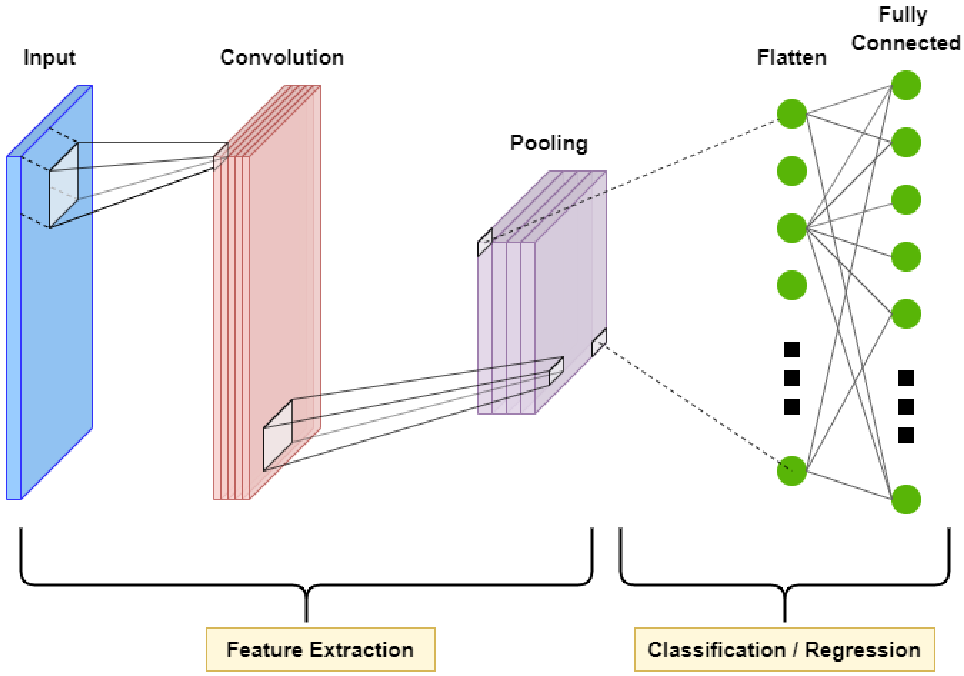


Figure 2.4: Schematic illustration of a basic Convolutional Neural Network architecture.

2.2 Ensemble Learning

Ensemble methods use multiple learning algorithms to obtain a better predictive performance than the one that would be obtained from any of its constituent learning algorithms alone. There are roughly three classes of ensemble learning techniques:

2.2.1 Bagging & Random Forests

Bagging, which is the short of bootstrap aggregating, is an ensemble learning technique that trains the same algorithm for every predictor, almost always decision trees, training them on different random subsets of the training set. The outputs made by each predictor are then combined using statistics, such as voting or averaging. The general structure of this type of algorithm is depicted in Figure 2.5. To generate the different random subsets from the same training set, there are two main approaches: bagging and pasting. The difference between them is just that with bagging, the sampling is performed with replacement. Thus, in this case, each training instance can appear several times for the same predictor. As mentioned above, a Random Forest is an ensemble of Decision Trees, generally trained via the

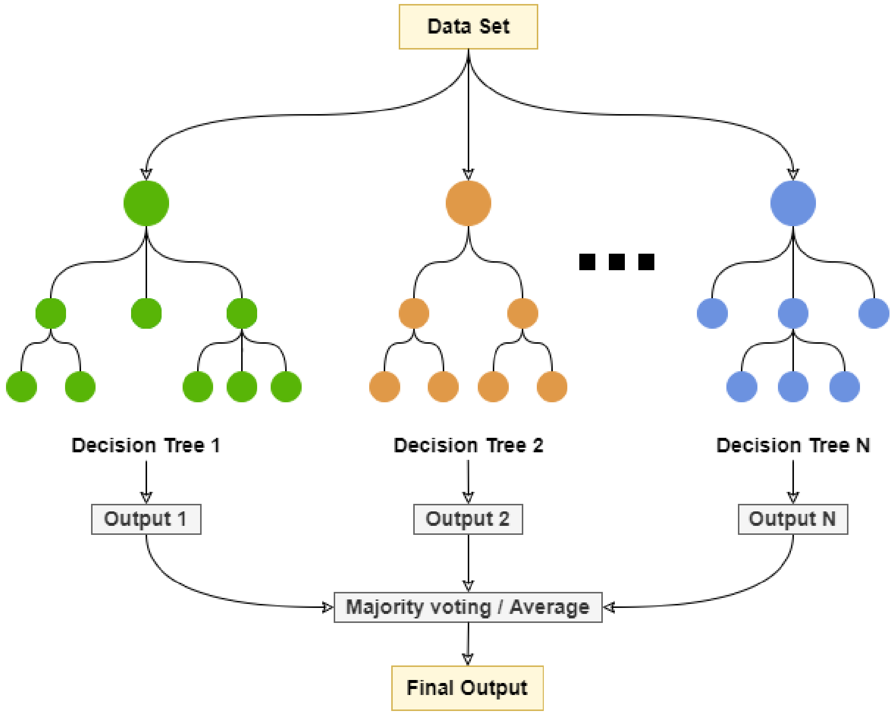


Figure 2.5: Diagram of the general structure of a Random Forest model.

bagging method. They outperform Decision Tree algorithms because they are less likely to overfit the training set. Although their increased complexity in comparison with their constituents, they maintain the capacity to measure the relative importance of each feature.

2.2.2 Stacking

Stacked Generalization is an ensemble method that combines the predictions from multiple Machine Learning models using the same dataset. At first, the constituents algorithms are trained, then a combiner algorithm is trained to make the final prediction using all the predictions made from the other algorithms as inputs of it.

2.2.3 Boosting

Boosting is an ensemble method that, like the ones before, combines several weak learners into a strong one. The general idea of it is to train predictors sequentially, each one trying to correct the performance of its predecessor. The most common ones are AdaBoost and Gradient Boosting.

2.3 Performance metrics

Every Machine Learning task requires an evaluation metric in order to assess the performance of the model. Due to the wide variety of algorithms, the best performance measure may be different for each of them. Therefore, it can be split into regression and classification tasks.

2.3.1 Regression metrics

Root Mean Square Error (RMSE)

It is the typical performance measure of a regression problem and gives an idea of how much error the model makes in its predictions, emphasizing with higher weight for large errors. It corresponds to the *Euclidean norm* (l_2 norm) and the mathematical formula to compute it is given by

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - \hat{\mathbf{y}}^{(i)}\right)^2} \tag{2.4}$$

where m is the number of instances in the dataset, $\hat{\mathbf{y}}^{(i)}$ is the predicted value for the i^{th} instance in the dataset and $y^{(i)}$ is its corresponding label.

Mean Absolute Error (MAE)

MAE is another way to measure the distance between the vector of predictions and the vector of target values. It is useful when small errors are more important than large errors and corresponds to the *Manhattan norm* (l_1 norm). The expression to compute this measure is given by

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y^{(i)} - \hat{\mathbf{y}}^{(i)}\right| \tag{2.5}$$

where the symbology is the same as for the previous metric.

2.3.2 Classification metrics

Accuracy

This metric consists in the ratio between the correct predictions and the total number of predictions made. This correlation can be expressed by

$$\text{accuracy} = \frac{\text{\# correct predictions}}{\text{Total number of predictions made}} \tag{2.6}$$

Confusion Matrix

Due to its simplicity, accuracy penalizes in the same way the errors committed when misclassifying any of the two classes, which in some cases will not be useful at all. Otherwise, a confusion matrix presents more detailed and schematic information about the model performance. Its general structure is depicted in Table 2.1.

		Truth label		
		Positive	Negative	Total
Model Prediction	Positive	TP	FP	$TP+FP$
	Negative	FN	TN	$FN+TN$
	Total	$TP+FN$	$FP+TN$	

Table 2.1: Confusion matrix diagram where TP=true positive, FP=false positive, FN=false negative and TN=true negative.

Chapter 3

Methodologies

Resumen

En el presente capítulo, se muestra la localización del estudio así como las características de las bases de datos utilizadas como entradas y validación de los distintos modelos. Se describen tanto los datos de polvo como los de las distintas variables meteorológicas utilizadas. Seguidamente, se realiza un análisis de la variabilidad y estacionalidad de los datos de polvo. Posteriormente, se muestra el mapa de erodibilidad de la zona en estudio [Gin01] así como las condiciones sinópticas propiciadoras de las invasiones de polvo en las Islas Canarias. Finalmente, se describen las técnicas utilizadas para reescalar y dividir los datos, además de una descripción de los modelos utilizados.

3.1 Study location & data

The study location is a point situated between the islands of Tenerife and Gran Canaria (Spain), concretely, at 28ºN and 16.25ºW (see Figure 3.1). The election of this location is due to its position in the middle of the Canary Islands, serving as a generalizer of dust concentrations in the whole archipelago. For the aforementioned point both, the dust load on the whole air column as well as the dust concentrations at two pressure levels (over the surface and at 750 hPa) have been taken. These dust data have been obtained from the Modern-Era Retrospective analysis for Research and Applications 2 (MERRA-2) [mer21]. The main advantages of this data set is that it provides data since 1980 and it also has a relatively high spatial (½° latitude by ⅝° longitude) and temporal (hourly) resolutions. In this study, daily mean time series have been created from these data.

Reanalysis is essential due to the fact that there is a great abundance of meteorological and geophysical data, obtained by modern observation methods (weather stations, weather balloons, aircrafts, ships, satellites, etc).

However, these observations are not evenly distributed around the globe. Therefore, reanalysis combines observations with weather models, creating a complete picture of the past weather. In this way, MERRA-2 reanalysis is produced by combining GEOS-5 (Goddard Earth Observing System v.5) atmospheric model with a 3D variational data assimilation algorithm (3DVAR) to ingest observational data. In the case of aerosols, MERRA-2 assimilates aerosol optical depth (AOD) using data obtained by different sensors on board satellites, such as, the Advanced Very High Resolution Radiometer (AVHRR), Moderate Resolution Imaging Spectroradiometer (MODIS), Multi-angle Imaging SpectroRadiometer (MISR), and also from ground-based remote sensing measurements, specifically, from Aerosol Robotic Network (AERONET) data sets. To take into account aerosol processes, GEOS-5 is radiatively coupled to the Goddard Chemistry Aerosol Radiation and Transport (GOCART) module.

Furthermore, data from ERA5 reanalysis [era21] have been taken as input features to the model. ERA5 takes data since 1979 to present in a grid with a horizontal resolution of 0.25º x 0.25º, at 37 different levels of vertical resolution and with a temporal resolution of one hour. Concretely, in this study, the variables taken from the ERA5 are:

- U-component of the wind $[m/s]$: Eastward component.
- V-component of the wind $[m/s]$: Northward component.
- Temperature $[K]$: Air temperature.
- Relative Humidity [%]: Water vapour pressure as a percentage of the value at which the air becomes saturated, which is the point at which water vapour begins to condense into liquid water or deposition into ice.
- Total Precipitation $[m]$: Accumulated liquid and frozen water, comprising rain and snow, that falls to the surface of the Earth.
- Total column vertically-integrated water vapour $[kg/m^2]$: Total amount of water vapour in a column extending from the surface of the Earth to the top of the atmosphere.

Due to data availability of the two data sets, the study period comprises from 1980 to 2020. In spite of the great amount of meteorological data provided by ERA5 in the study region, the original spatial resolution has been degraded to a 2º x 2º grid, as shown in Figure 3.1. Additionally, the 12:00 UTC value of ERA5 variables have been taken as representative for each day.

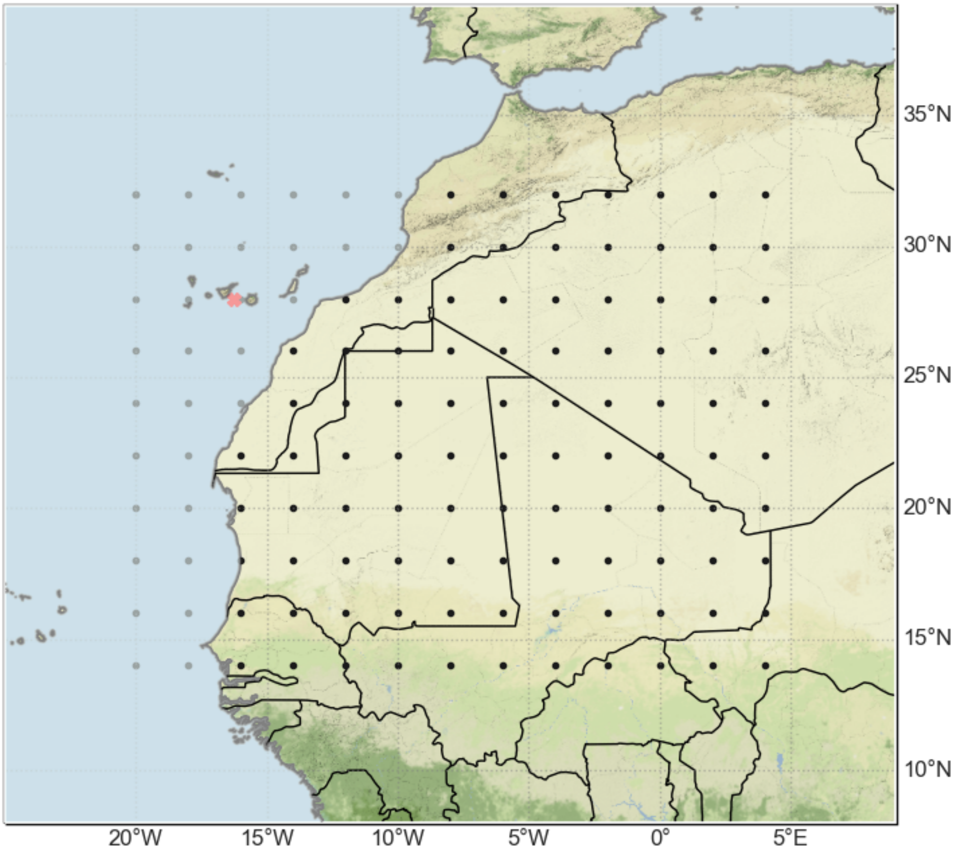


Figure 3.1: Map of the African continent with the grid of points from which ERA5 data have been used as well as the study location, marked with a red cross at 28ºN and 16.25ºW.

3.2 Data preprocessing

3.2.1 Dust measurements

As it was mentioned in Section 3.1, the dust concentrations data set consists of measurements of this variable at two pressure levels. In Figure 3.2, a representation of the dust concentrations over the year 2020 is displayed. There is a seasonal pattern of dust invasions, whilst in Winter the dust is mainly transported at low altitudes, in Summer this tendency is the opposite, carrying the dust mainly at high altitudes (over the inversion layer, Saharan Air Layer). However, these trends are very generic and could not be quite accurate, as can be seen in Figure 3.2 in some days of the Summer months. To see the aforementioned trends, the monthly mean dust concentrations over the whole data set (from 1980 to 2020) have

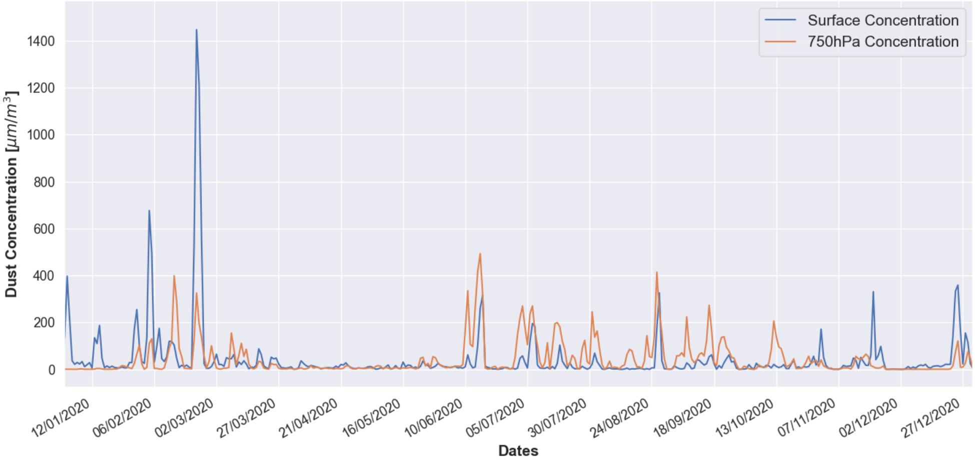


Figure 3.2: Dust concentrations at two pressure levels on 2020 for the study location.

been estimated and depicted in Figure 3.3, where this tendency is easily distinguishable. As with the dust concentrations, the monthly mean dust loads over the whole study period have been computed, getting the results represented in Figure 3.4. The dust load trend is similar to the one of the 750 hPa dust concentration in the way that both reach their maximum during the Summer months.

Huge efforts have been made to characterize the main dust sources. One of the most relevant results was obtained by [Gin01], which took into account both, topography and vegetation, to create a source function. Using it, the source map shown as the base map in Figure 3.5 has been composed.

Wind speed

A typical synoptic condition of the surface wind on a dust day (23-2-2020) is shown in Figure 3.5, where the strong easterly winds over the Saharan region produces a high surface dust concentration peak on the Canary Islands, which can be seen in Figure 3.2. Additionally, in Figure 3.6 is depicted the image taken for the same day from the WorldView-2 satellite [NAS22a], where the dust plume coming from West Sahara towards the Canary Island can be clearly seen.

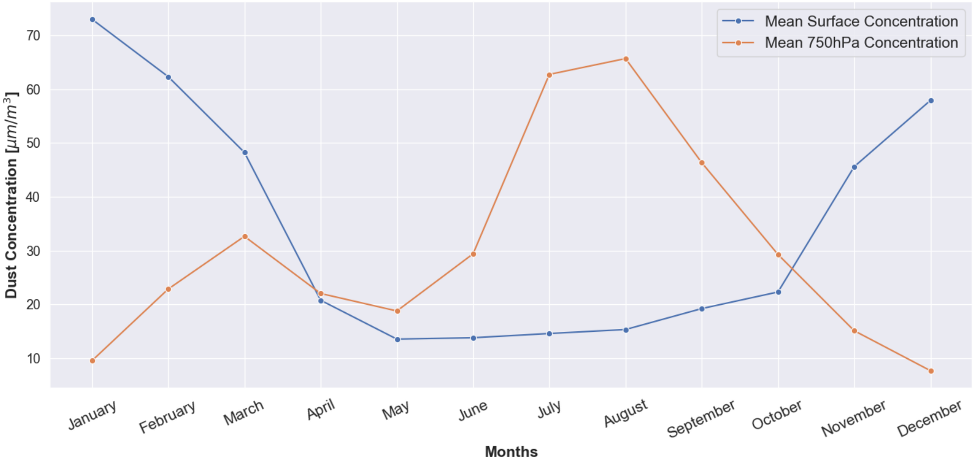


Figure 3.3: Monthly average dust concentrations at two pressure levels for the study location.

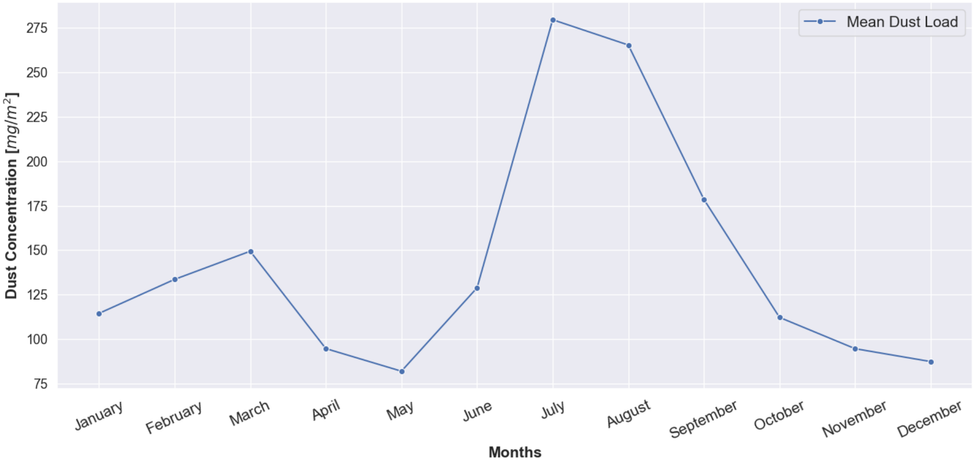


Figure 3.4: Monthly average dust load on the whole air column for the study location.

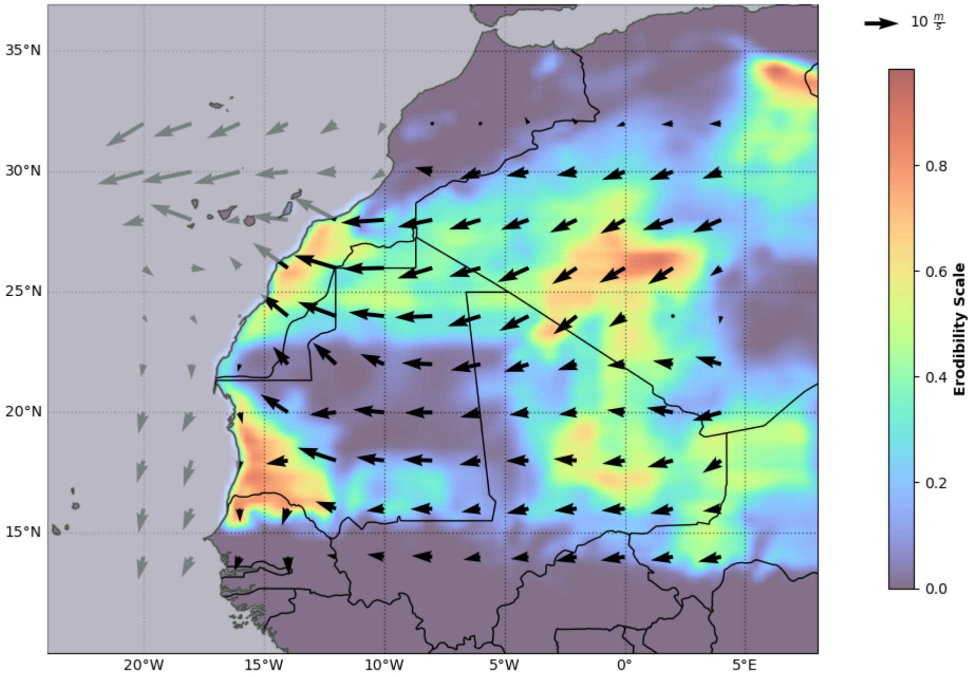


Figure 3.5: Surface wind map in the study region for the day 23-2-2020 (high dust intrusion) with the erodibility mask provided by [Gin01], where the length and orientation of the arrows indicate the wind intensity and the direction in which the wind blows, respectively.

3.2.2 Feature scaling

Most Machine Learning models are sensitive to feature scaling, specifically those algorithms that make use of gradient descent as the optimization technique. This is produced because the difference in ranges of features may cause different step sizes for each of them, which will slow down the process of converging to the minima. Normalization is a commonly scaling technique in which values are shifted and rescaled, ending up with values between 0 and 1, by using its maximum and minimum values. The mathematical expression for this process is given by

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.1}$$

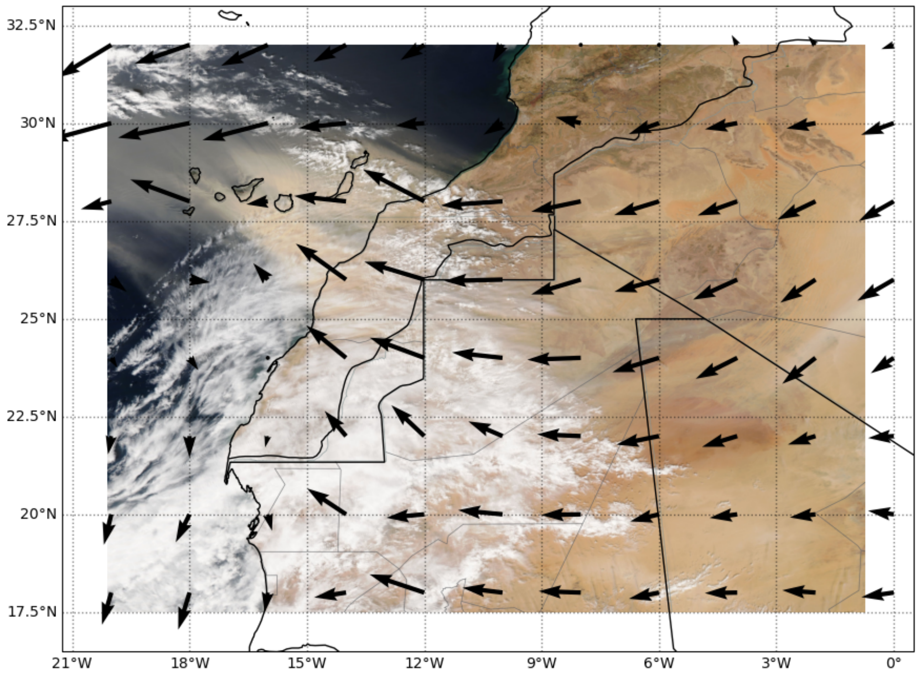


Figure 3.6: WorldView-2 image [NAS22a] with the wind map of the day 23-2-2020 over Northwest Africa, where the wind speed scale is the same as in Figure 3.5.

3.2.3 Data splitting

In Machine Learning models, the main reason to apply data splitting is to avoid overfitting, which basically means that the model fits the training data very well but fails to generalize them effectively, obtaining poor performances on the test set. The original data are divided into:

- **Training set**: Is the portion of data used to train the model. In this case, the training set ranges from the first date 1-1-1980 to 12-31-2009 (75%).
- **Validation set**: Is the data set used to tune the learning process parameters and also serves as model selection. In this work, the validation set ranges from 1-1-2010 to 31-12-2014 (10%).
- **Test set**: Is the portion of data used to test the final model. It acts as an evaluation of the final model performance. In this project, the test set ranges from 1-1-2015 to 31-12-2020 (15%).

There are other alternative split techniques such as k-fold cross-validation.

3.3 Machine Learning models

Throughout this work, different Machine Learning techniques have been applied, namely Decision Trees, Random Forests and Artificial Neural Networks. Of all of them, deep neural networks proved to be the best at predicting dust. Additionally, taking into account its relatively simple structure, the results obtained with 2 different neural networks are shown along this project. The first of them will be a classical ANN, while the other will be another ANN with two convolutional layers. At a first stage, the hyperparameters were chosen manually looking for the best performance of the models. Subsequently, an automatic search was applied with the help of the Keras-Tuner package, thereby obtaining more refined configurations.

3.3.1 ANN (Model 1)

The structure of this model is the one shown in Figure 2.3, adding to it two additional hidden layers. Its main characteristics are the following:

- *First hidden layer*. This first hidden layer is formed by 64 neurons with the ReLU activation function.
- *Second hidden layer*. The second hidden layer consists of 128 neurons with the ReLU activation function and dropout ($p^{[2]} = 0.01$).
- *Third hidden layer*. The third hidden layer composed of 64 neurons with the ReLU activation function.
- *Output layer*. The output layer presents one or two nodes, depending on if the model is intended to predict dust concentrations at the two pressure levels mentioned or if it only predicts the dust load on the whole air column.

Hence, this ANN is composed of 116,610 trainable parameters which are updated using an Adam optimizer throughout 20 epochs, which are the number of evaluations of the whole training set [1].

3.3.2 ANN with convolutionals (Model 2)

The specific structure of this model is depicted in Figure 3.7. Its main characteristics are:

[1] All the neural networks shown in this project have been programmed in the Keras-Tensorflow deep learning framework, with the use of the Scikit-Learn package as well.

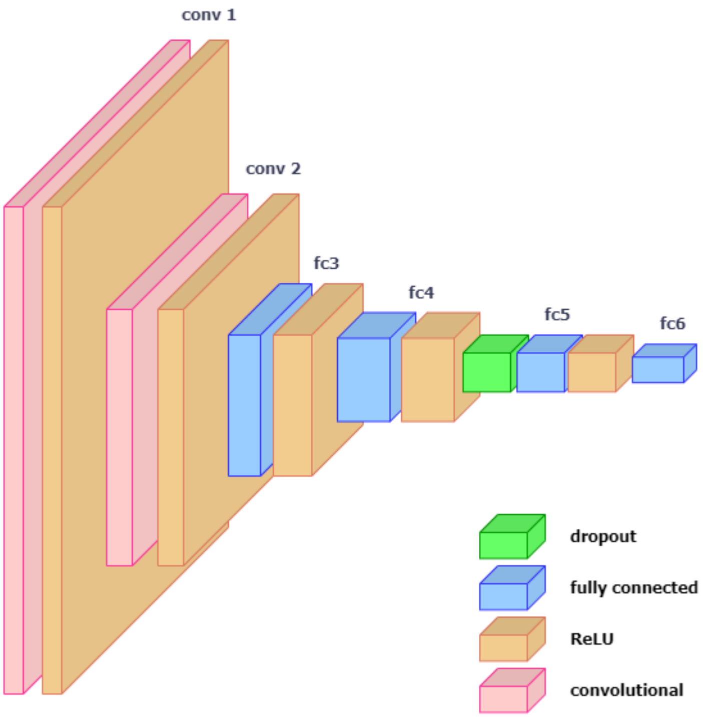


Figure 3.7: Model 2 architecture: ANN with two convolutional layers.

- *First hidden layer*. This first hidden layer is formed by a convolutional layer with 64 neurons with the ReLU activation function.
- *Second hidden layer*. The second hidden layer consists of another convolutional layer with 32 neurons with the ReLU activation function.
- *Third hidden layer*. This third hidden layer is composed of 64 neurons with the ReLU activation function.
- *Fourth hidden layer*. The fourth hidden layer is formed by 128 neurons with the ReLU activation function and dropout ($p^{[2]} = 0.01$).
- *Fifth hidden layer*. The fifth hidden layer is made up of 64 neurons with the ReLU activation function.
- *Output layer*. The output layer presents one or two nodes, depending on if the model is intended to predict dust concentrations at the two levels mentioned or if it only predicts the dust load on the whole air column.

Therefore, this ANN with convolutionals is composed of 283,362 trainable parameters which are updated using an Adam optimizer along 20 epochs.

Chapter 4

Results

Resumen

En este capítulo se compara la capacidad de predicción de los modelos descritos en la Sección 3.3 en diferentes periodos de tiempo (invierno, verano y año completo), habiendo sido entrenados también en diferentes épocas. A continuación, se varía tanto las ventanas de tiempo en las que el modelo toma variables de entrada, como diferentes magnitudes físicas, comparando los resultados de cada uno de ellos. Posteriormente, se realiza una pequeña modificación en los mencionados modelos con la finalidad de que predigan la concentración de polvo tanto en superficie como en un nivel de presión de 750 hPa. Finalmente, se comparan los errores cometidos en la predicción de estas dos concentraciones variando los periodos de entrenamiento y testeo, así como las variables de entrada.

4.1 Dust Load

As a first step, the models described in Section 3.3, fully connected neural network (Model 1) and convolutional neural network (Model 2), will be used to predict the dust load in the study location, using not only different meteorological variables from the previously described data grid (see Figure 3.1), but also three different training periods: year (whole year data), winter (December, January, February and March) and summer (June, July, August and September) as well as different window sizes. To easily compare the performance of the two models, two statistical variables to quantify the errors made by each model, which are the ones explained in Section 2.3.1, will be given. Moreover, the window size refers to the number of previous days from which data is taken as input features to the model. The reason behind training the model with different periods is that, as seen in Section 3.2.1, dust intrusions follow a clearly seasonal pattern, which may affect the performance of both models. Bearing this in mind, in Figure 4.1, the forecast of both models, using only surface wind speed data for the

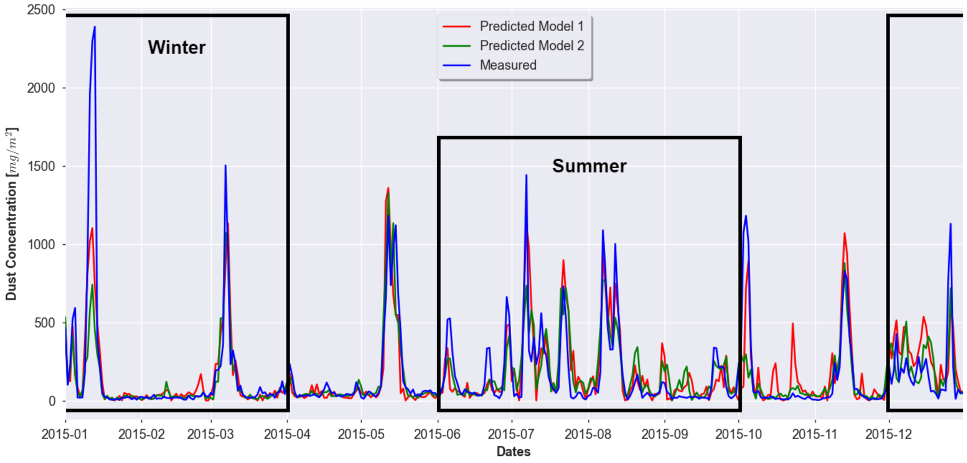


Figure 4.1: Comparison of models performance on dust load forecast for 2015, including the winter and summer periods.

three aforementioned periods, is plotted. To effectively compare the models accomplishment with different training periods, Table 4.1 has been created, obtaining that the best whole year performances are achieved when the models are trained with the whole year data, as expected. However, when the models are trained with seasonal data (say winter or summer), a slightly better performance on the same season is reached than the whole year model, but it fails to generalize well on the other season, getting much worse results. The large errors obtained when models are trained with summer data and applied to winter series are remarkable, and more when they are compared with just the opposite structure, that is, trained with winter data and tested with summer series.

Afterwards, different window sizes have been tested (see Table 4.2), obtaining that the best performances are reached when a window size of 5 days lag is taken. Finally, models prediction errors with different input variables are compared in Table 4.3, where both, surface and 750 hPa wind speeds as well as them with surface precipitation, seem to be the best entry variables of the models in order to efficiently forecast the dust loads at the study location. The worsening of the results for the summer period when column water content or temperature are included is noteworthy.

Model	Training Period	Testing Period	RMSE	MAE
Model 1	Year	Year	194.48	97.65
		Winter	164.81	85.96
		Summer	155.24	107.71
	Winter	Year	228.98	126.14
		Winter	156.54	78.90
		Summer	292.12	236.48
	Summer	Year	354.53	208.67
		Winter	542.88	384.89
		Summer	140.82	103.77
Model 2	Year	Year	183.52	89.41
		Winter	153.19	76.90
		Summer	145.48	97.78
	Winter	Year	224.14	114.76
		Winter	151.31	74.83
		Summer	249.02	180.97
	Summer	Year	290.68	161.79
		Winter	412.37	262.43
		Summer	148.08	106.45

Table 4.1: Comparison of models prediction errors on dust load forecasts on different time periods with surface wind data and a window of 5 days, where the given metrics are expressed in mg/m^2.

Model	Window Size	Testing Period	RMSE	MAE
Model 1	6	Year	198.53	98.66
		Winter	198.66	128.24
		Summer	204.02	161.74
	5	Year	194.48	97.65
		Winter	164.81	85.96
		Summer	155.24	107.71
	4	Year	191.01	99.71
		Winter	183.28	98.61
		Summer	195.40	129.97
	3	Year	251.17	140.82
		Winter	231.00	130.82
		Summer	263.37	178.30
	2	Year	224.65	127.99
		Winter	225.45	128.92
		Summer	333.72	214.37

(Continues on next page)

Model	Window Size	Testing Period	RMSE	MAE
Model 2	6	Year	188.54	94.14
		Winter	180.50	108.81
		Summer	172.94	133.48
	5	Year	183.52	89.41
		Winter	153.19	76.90
		Summer	145.48	97.78
	4	Year	189.21	96.52
		Winter	183.01	97.77
		Summer	199.78	130.27
	3	Year	249.27	135.20
		Winter	228.61	123.65
		Summer	256.97	171.02
	2	Year	229.01	124.89
		Winter	224.33	125.99
		Summer	346.91	221.25

Table 4.2: Performance of both models on dust load predictions trained with surface wind data, the whole year period and different window sizes, where the shown metrics are given in mg/m^2.

Model	Input Variables	Testing Period	RMSE	MAE
Model 1	750hPa wind speed	Year	190.27	100.19
		Winter	184.36	100.49
		Summer	230.80	130.27
	Surface and 750hPa wind speeds	Year	163.74	83.05
		Winter	148.36	75.17
		Summer	168.16	109.24
	Surface wind speed and total column water content	Year	180.32	98.94
		Winter	147.80	81.24
		Summer	237.05	150.22
	Surface and 750hPa wind speeds and surface precipitation	Year	168.00	84.13
		Winter	159.02	79.68
		Summer	148.57	86.58
	Surface and 750hPa wind speeds and temperatures	Year	191.65	108.19
		Winter	176.92	108.16
		Summer	239.32	144.76

(Continues on next page)

Model	Input Variables	Testing Period	RMSE	MAE
Model 2	750hPa wind speed	Year	184.01	93.27
		Winter	176.35	93.70
		Summer	223.98	122.32
	Surface and 750hPa wind speeds	Year	166.42	84.01
		Winter	152.65	79.16
		Summer	169.23	102.36
	Surface wind speed and total column water content	Year	172.87	89.17
		Winter	146.82	79.61
		Summer	226.74	131.42
	Surface and 750hPa wind speeds and surface precipitation	Year	168.62	84.10
		Winter	159.64	79.46
		Summer	162.86	97.25
	Surface and 750hPa wind speeds and temperatures	Year	182.89	97.62
		Winter	167.71	96.71
		Summer	228.70	131.36

Table 4.3: Models prediction accurateness on dust load forecasts trained with different input variables, the whole year period and a window size of 5 days, where the displayed performance measures are expressed in mg/m^2.

4.2 Dust Concentration

The strong seasonal pattern in dust intrusions over the Canary Islands, observed in Figure 3.3, suggests to predict the dust concentrations at two different levels: surface and 750 hPa. For this purpose, the models described in Section 3.3 have been slightly modified, by changing from 1 to 2 units in the output layer. As an example, in Figure 4.2, the output dust concentrations at the two levels for both models are depicted for the year 2019. The model ability to distinguish the seasonal pattern on dust intrusions stands out, even more when the lack of day information (for instance, month or julian day) the model has, is taken into account.

Additionally, the best performances were achieved when a window size of 5 days was taken, as shown in Table 4.2. For this reason, all dust concentration models have the aforementioned window size. Taking this into account, the performance of both models on the three testing periods and with different training periods and input variables are summarized in Table 4.4. As can be observed, the models prediction accuracy at the two levels are quite similar. The same seasonal results as with dust loads are obtained (see Section 4.1) and again, the best global performances are achieved when models are trained with the whole year. Regarding the input variables, good results are obtained with all of them, specially with surface wind speed. When the models are

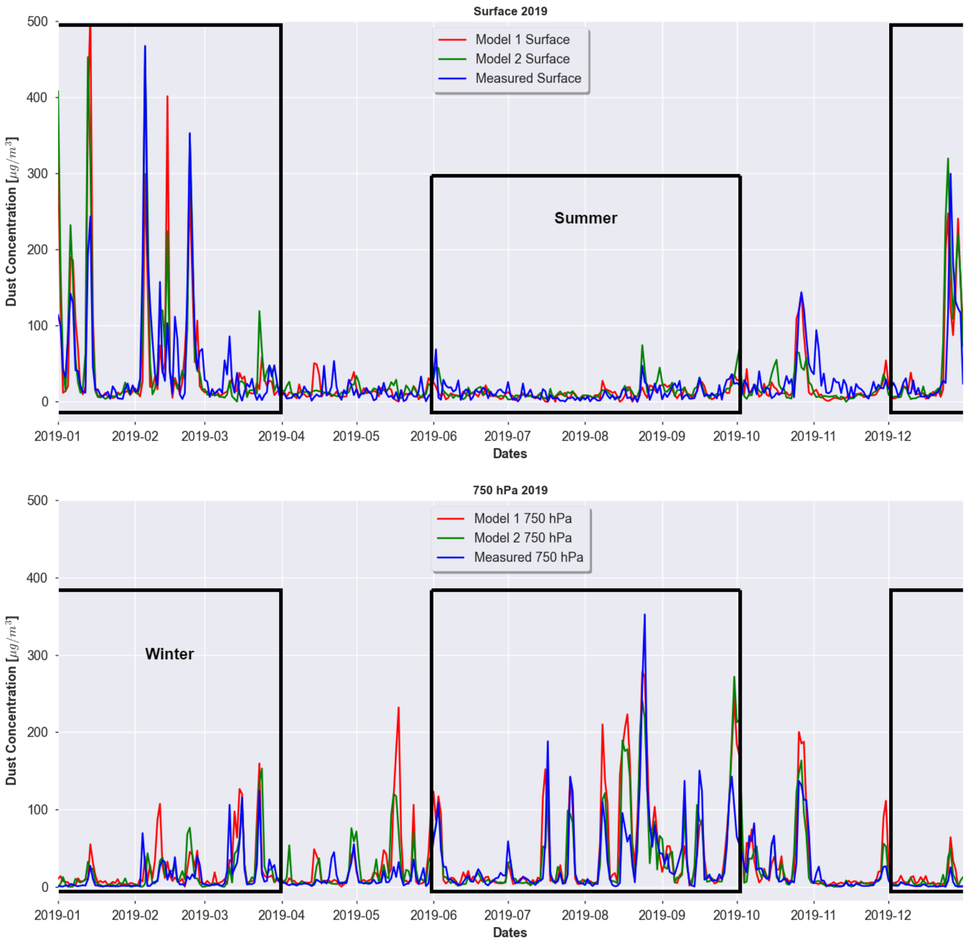


Figure 4.2: Comparison of models performance on dust concentration forecast for 2019 at two different pressure levels, including the winter and summer periods.

trained with surface wind speed or relative humidities as inputs, the results are similar. However, the simultaneous combination of both variables does not improve the predictions using the tested models. When dust concentrations at two different levels are considered, rather than the total column, the prediction results for the winter season are not so bad when the models are only trained on the summer data. The exceptional performances obtained with very different physical input variables suggests the idea to create a global model made up of specialized algorithms, applying any Ensemble Learning technique, like those described in Section 2.2.

Model	Input Variables	Training	Testing	Surface		750 hPa	
		Period	Period	RMSE	MAE	RMSE	MAE
Model 1	Surface wind speed	Year	Year	55.77	22.94	53.53	23.02
			Winter	83.99	35.19	80.69	32.47
			Summer	45.39	22.86	46.26	24.37
	Surface and 750hPa wind speeds	Year	Year	57.85	23.06	57.06	22.67
			Winter	83.73	33.79	81.36	31.65
			Summer	42.45	21.17	43.14	22.05
		Winter	Year	69.08	31.88	70.52	33.57
			Winter	84.24	34.59	82.36	32.46
			Summer	72.75	42.32	79.37	49.84
		Summer	Year	75.98	35.33	74.99	35.35
			Winter	117.49	61.32	115.17	59.66
			Summer	43.05	21.92	43.49	22.97
	Surface and 750hPa relative humidities	Year	Year	60.54	24.55	59.89	24.75
			Winter	88.90	35.21	86.99	33.51
			Summer	45.73	23.16	46.28	24.90
	Surface and 750hPa wind speeds and relative humidities	Year	Year	56.14	22.40	55.29	22.45
			Winter	88.81	33.95	85.05	33.25
			Summer	47.97	24.57	47.88	25.51
Model 2	Surface wind speed	Year	Year	54.56	22.97	52.67	22.98
			Winter	85.45	36.35	76.73	29.52
			Summer	43.44	20.97	48.16	26.24
	Surface and 750hPa wind speeds	Year	Year	59.02	23.12	55.58	22.35
			Winter	86.06	35.32	77.63	28.96
			Summer	41.26	19.84	44.53	23.70
		Winter	Year	68.66	30.58	68.57	32.12
			Winter	87.19	36.30	78.40	29.75
			Summer	67.89	37.23	78.45	48.39
		Summer	Year	76.56	33.88	74.14	36.86
			Winter	119.60	59.31	112.52	61.75
			Summer	41.46	20.48	45.26	24.84
	Surface and 750hPa relative humidities	Year	Year	61.73	24.84	58.52	24.47
			Winter	92.51	37.61	82.50	30.74
			Summer	43.83	21.72	48.52	26.78
	Surface and 750hPa wind speeds and relative humidities	Year	Year	57.22	22.92	53.93	22.16
			Winter	90.36	37.33	81.02	30.41
			Summer	45.93	22.65	51.10	28.01

Table 4.4: Performance of the models on dust concentration predictions at surface and 750 hPa levels with different input variables, training periods and a window size of 5 days, where the given metrics are expressed in $\mu g/m^3$.

Chapter 5

Conclusions & Future Work

The goal of this project, based on meteorological data obtained from reanalysis, was to develop an algorithm capable of predicting the dust concentration or the dust load over the place of study, located between the islands of Tenerife and Gran Canaria, in Spain. As a first step, a global overview of the main dust sources around the world is given, also presenting the key role that it plays in the Earth lithosphere, atmosphere and biosphere. A succinct analysis of the dust production mechanisms was given by reviewing the related literature, paying special attention to those occurring in northwestern Africa. With this is mind, the fundamental concepts behind Machine Learning were reviewed, describing the key components of some of the algorithms as well as the metrics most used to easily compare their performances on different tasks. Afterwards, the data used to train the models were described, considering its spatial and temporal resolution, in addition to the study location position and its importance. Once this step was completed, it was followed by a preprocessing phase, where the seasonal pattern of dust intrusions was analyzed. Hereunder, the techniques applied to the data set were explained as well as the main characteristics of the models used. Finally, different Machine Learning models were not only trained with different periods, window sizes and input variables, but also tested on various seasonal periods. The results obtained were then compared between the models, in order to establish the best one either to forecast the dust concentration at the two pressure levels studied or to predict the dust load. Furthermore, a brief analysis of the use of neural networks acting as classifiers, rather than regressors, is provided in the Appendix A.

The results obtained throughout this project led to the following conclusions:

1. Despite the fact of having few meteorological variables as model input data and no information about orography or lithology, the results

obtained are remarkable. Specially, the ability of neural networks to detect patterns in data seems quite impressive, and may lead to a better understanding of dust events.

2. Surface wind speed has demonstrated to be the best meteorological input variable to the models in order to predict dust loads.

3. Additionally for dust load, a window size of 5 days lag has resulted of being the best option to achieve better performances.

4. Models trained with whole year data have turned out to be the ones that have obtained the best global results. The models trained with each of the seasons separately provide better results for that period, but significantly worse results for the other one.

5. When it comes to predicting dust concentrations, both surface wind speed and relative humidities at the two pressure levels, became the more convenient input variables, when used independently, to obtain the finest performances.

At the time of this work, neural networks have been broadly applied to pollution aerosol forecasts ([dG13], [Cho20], [Pak20]), but few studies exist on the application of these techniques to the prediction of dust ([GC21], [Kan19]). However, the results obtained reveal the potential of these algorithms on these tasks.

Future work of this project will be based on applying other Machine Learning techniques like ensemble learning or recurrent neural networks, as well as expanding the data grid to a better spatial and temporal resolution. Having seen the good performance of the models in predicting dust over the Canary Islands, it seems very promising that these models will also achieve good performances in other places with very different meteorological phenomena. In conclusion, the encouraging results obtained throughout this project along with the chance of great future improvements on Machine Learning algorithms, enable the possibility of creating new projects in this line of research.

Bibliography

[BA09] Koren-Ilan Altaratz Ben-Ami, Yuval. Patterns of north african dust transport over the atlantic: winter vs. summer, based on calipso first year data. *Atmospheric Chemistry and Physics*, 9(20):7867–7875, 2009.

[bod22] Nasa earth observatory. https://earthobservatory.nasa.gov/images/12734/dust-storms-from-africas-bodele-depression, 2022.

[Cho20] Abdolshahnejad Mahsa-Moradi Ehsan-Querol Xavier Mosavi-Amir Shamshirband Shahaboddin Ghamisi Pedram Choubin, Bahram. Spatial hazard assessment of the pm10 using machine learning models in barcelona, spain. *Science of The Total Environment*, 701:134474, 2020.

[CHS19] Washington-Richard Engelstaedter Caton Harrison, Thomas and Sebastian. A 14-year climatology of saharan dust emission mechanisms inferred from automatically tracked plumes. *Journal of Geophysical Research: Atmospheres*, 124(16):9665–9690, 2019.

[dG13] Trizio-Livia Di Gilio Alessia Pey-Jorge Pérez Noemi Cusack Michael Alastuey-Andrés Querol Xavier de Gennaro, Gianluigi. Neural network model for the prediction of pm10 daily concentrations in two sites in the western mediterranean. *Science of the Total Environment*, 463:875–883, 2013.

[DP19] Atria Dika Puspita. Get to know how artificial neural network formed in computer science. https://medium.com/@atriadplt/, 2019.

[Eng07] Washington Richard Engelstaedter, Sebastian. Atmospheric controls on the annual cycle of north african dust. *Journal of Geophysical Research: Atmospheres*, 112(D3), 2007.

[era21] Era5 hourly data on pressure levels from 1979 to present. https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=overview, 2021.

[Feu18] Schepanski Kerstin Feuerstein, Stefanie. Identification of dust sources in a saharan dust hot-spot and their implementation in a dust-emission model. *Remote Sensing*, 11(1):4, 2018.

[GC21] Aguilar-RM Criado-Hernandez C Gonzalez-Mendoza LA Gonzalez-Calvo, D. Multivariate influence through neural networks ensemble: Study of saharan dust intrusion in the canary islands. *Applied Soft Computing*, 107:107497, 2021.

[Gér19] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc.", 2019.

[GF20] Javier Galván Fraile. Machine learning for remote sensing of xylella fastidiosa. 2020.

[Gin01] Chin Mian-Tegen Ina-Prospero Joseph M-Holben Brent Dubovik Oleg Lin Shian-Jiann Ginoux, Paul. Sources and distributions of dust aerosols simulated with the gocart model. *Journal of Geophysical Research: Atmospheres*, 106(D17):20255–20273, 2001.

[Gin12] Prospero Joseph M Gill-Thomas E Hsu-N Christina Zhao Ming Ginoux, Paul. Global-scale attribution of anthropogenic and natural dust sources and their emission rates based on modis deep blue aerosol products. *Reviews of Geophysics*, 50(3), 2012.

[Gri01] Kellogg Christina A-Shinn Eugene A Griffin, Dale W. Dust in the wind: long range transport of dust in the atmosphere and its implications for global public and ecosystem health. *Global Change and Human Health*, 2(1):20–33, 2001.

[Kan19] Kim Namgi Lee Byoung-Dai Kang, Sunwon. Fine dust forecast based on recurrent neural networks. In *2019 21st International Conference on Advanced Communication Technology (ICACT)*, pages 456–459. IEEE, 2019.

[Kni14] Stuut Jan-Berend W Knippertz, Peter. Mineral dust. *Mineral dust—A key player in the Earth system*, pages 121–147, 2014.

[Kok12] Parteli-Eric JR-Michaels Timothy I-Karam Diana Bou Kok, Jasper F. The physics of wind-blown sand and dust. *Reports on progress in Physics*, 75(10):106901, 2012.

[Lov10] Richard Lovett. African dust keeps amazon blooming, 2010.

[mer21] Nasa global modeling and assimilation office. https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/, 2021.

[Muh14] Prospero-Joseph M-Baddock Matthew C-Gill Thomas E Muhs, Daniel R. Identifying sources of aeolian mineral dust: Present and past. In *Mineral Dust*, pages 51–74. Springer, 2014.

[NAS22a] NASA. Eosdis worldview. https://worldview.earthdata.nasa.gov/, 2022.

[NAS22b] NASA. Toms absorbing aerosol index. https://earthobservatory.nasa.gov/images/1043/toms-aerosol-index, 2022.

[Ng22] Andrew Ng. Machine Learning. Coursera. Stanford University, 2022.

[Nym20] Ormerod Paul Nyman, Rickard. Understanding the great recession using machine learning algorithms. 2020.

[OA22] World Meteorological Organization and AEMET. Barcelona dust regional center. https://dust.aemet.es/, 2022.

[Pak20] Ma Jun-Ryu Unsok-Ryom Kwangchol Juhyok-U Pak Kyongsok Pak Chanil Pak, Unjin. Deep learning-based pm2. 5 prediction considering the spatiotemporal correlations: A case study of beijing, china. *Science of The Total Environment*, 699:133561, 2020.

[Pro02] Ginoux-Paul Torres-Omar Nicholson Sharon E Gill Thomas E Prospero, Joseph M. Environmental characterization of global sources of atmospheric soil dust identified with the nimbus 7 total ozone mapping spectrometer (toms) absorbing aerosol product. *Reviews of geophysics*, 40(1):2–1, 2002.

[Sam59] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3.3:210–229, 1959.

[Sha08] Yaping Shao. *Physics and modelling of wind erosion*. Springer, 2008.

[Swa92] Garstang Michael Greco S Talbot R Kållberg P Swap, Robert. Saharan dust in the amazon basin. *Tellus B*, 44(2):133–149, 1992.

[Van17] De Mazière Martine Vandenbussche, Sophie. African mineral dust sources: a combined analysis based on 3d dust aerosols distributions, winds and surface parameters. *Atmospheric Chemistry and Physics Discussions*, pages 1–37, 2017.

[WI09] Bouet Christel Cautenet Guy Mackenzie Elisabeth Ashpole Ian Engelstaedter Sebastian Lizcano Gil Henderson-Gideon M Schepanski Kerstin Tegen Washington, Richard and Ina. Dust as a tipping element: the bodélé depression, chad. *Proceedings of the National Academy of Sciences*, 106(49):20564–20571, 2009.

Appendices

A Dust Events Classifier

Resumen

En este apéndice se muestra una breve descripción de uno de los centros regionales, pertenecientes a la Organización Meteorológica Mundial (WMO), dedicados a la predicción de aerosoles, concretamente, de polvo. A continuación, se muestra la predicción realizada por uno de los algoritmos utilizados en dicha institución. Seguidamente, se describe la estructura de los modelos usados en el presente proyecto, además de mostrar sus predicciones y los errores cometidos en la clasificación de los diferentes eventos de polvo.

A.1 General purpose

As it was briefly discussed in Section 1.3, nowadays dust numerical models are not completely accurate, which may be understood by taking into account the importance of small-scale features on dust production and the difficulty to quantify them on global models. One of the leading centers on dust prediction is the Barcelona Dust Forecast Center, located in the mentioned city, in Spain. It forms part of the World Meteorological Organization (WMO) by being a Regional Center specialized on Atmospheric Sand and Dust Forecast. Furthermore, it produces dust predictions for Northern Africa, Middle East and Europe, which includes dust surface concentration and dust load, among many other variables. In Figure A.1, one example of a dust load forecast made by the Multiscale Online Nonhydrostatic AtmospheRe CHemistry (MONARCH) model is depicted. Its predicted dust load values are presented as 9 discrete ranges. Concretely, in Table A.1 the dust load classes are summarized with their corresponding values, as well as the total count of days in the data set in which each range is being measured. As there are no data from classes 7 and 8, they will not be included in the following models.

A.2 Model description

Taking all of these into account, the aim of this section is to create a model able to predict the dust intrusion class, taking the same ones as the MONARCH model does. The general structure of the model is the same as described in Section 3.3. However, the output layer is changed to a new one made up of 7 neurons

Dust Load Class	Dust Load Range $[mg/m^2]$	Class Total Data
Class 0	0 - 100	10019
Class 1	100 - 200	1727
Class 2	200 - 400	1589
Class 3	400 - 800	1244
Class 4	800 - 1200	302
Class 5	1200 - 1600	64
Class 6	1600 - 3200	31
Class 7	3200 - 6400	0
Class 8	> 6400	0

Table A.1: MONARCH dust load classes with the total number of them on the whole dust data set

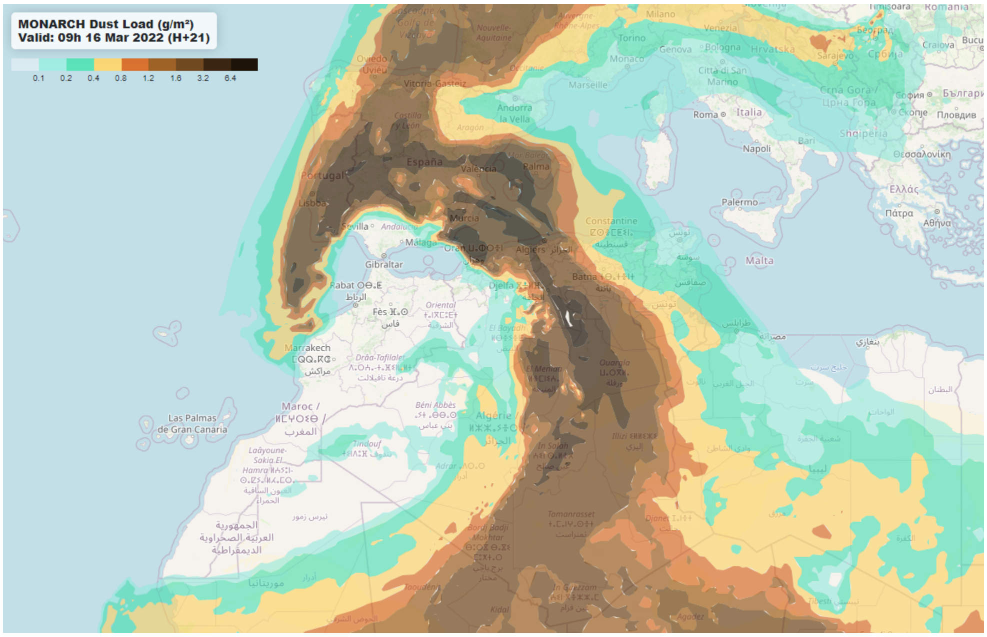


Figure A.1: Dust Load Forecast for 16-3-2022 from WMO Barcelona Supercomputing Center [OA22].

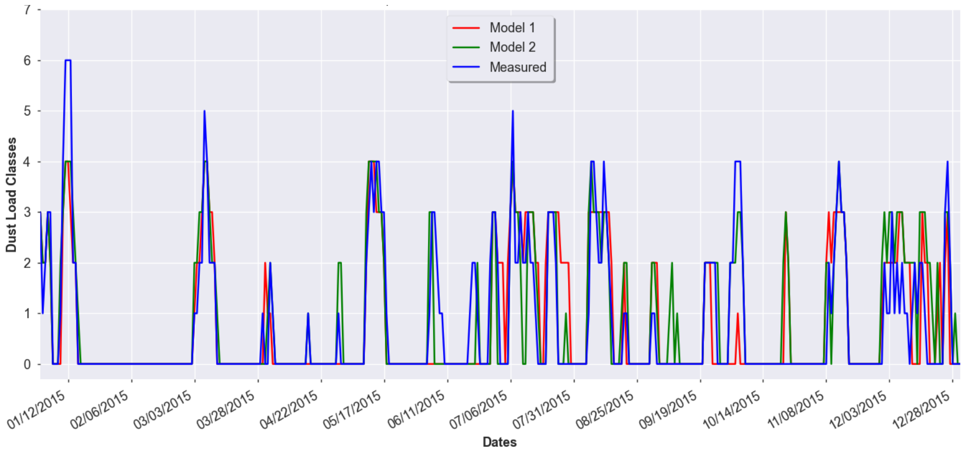


Figure A.2: Comparison of models performance on dust load class forecast for 2015.

with a different activation function called softmax. This new function generates a probability of each output class, and then, the one that maximizes the probability is taken as the predicted one.

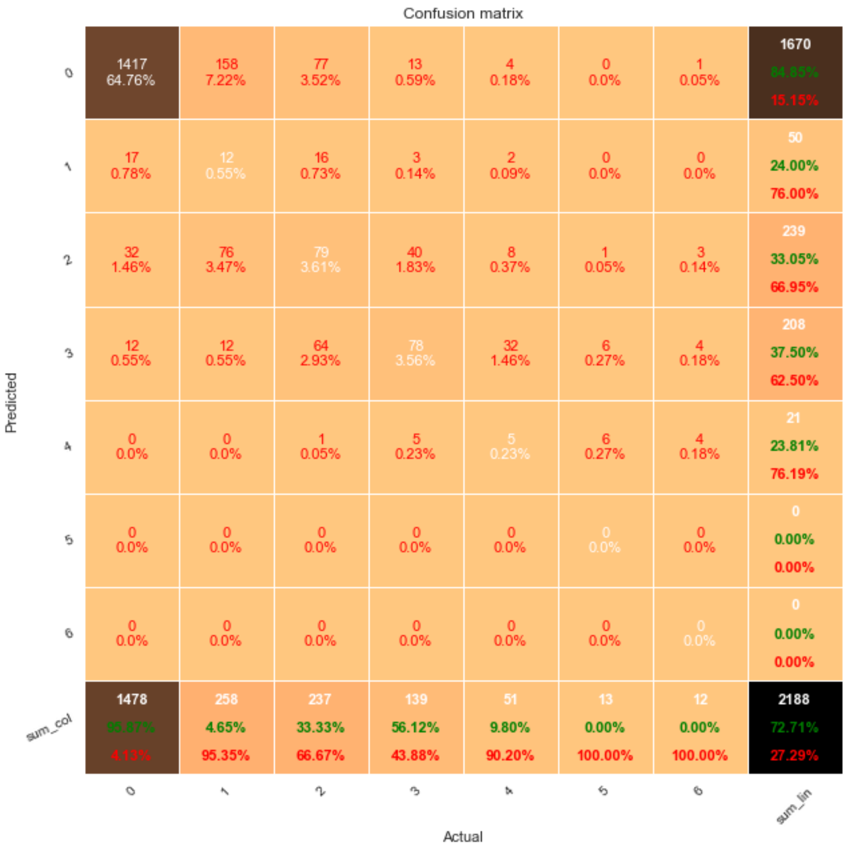
Having seen in Section 4.1 that the best results were obtained when a window size of 5 days lag was taken and trained with whole year data, these conclusions have been extrapolated to this new model. Apart from these parameters, the surface wind speed has been taken as input variable. In Figure A.2, the output of both models, the fully connected neural network (Model 1) and the convolutional neural network (Model 2), for the year 2015 is plotted. Due to the fact that the output in these models are classes, the best way to easily compare them and see their accurateness is to create the confusion matrices, with the structure explained in Section 2.3.2.

In Figures A.3 and A.4, the confusion matrices for models 1 and 2, are shown. Apart from their main characteristics, the percentages on them corresponds to: the ones shown inside the matrix (on the classes) refers to the portion of this class over the total number of classes; the ones shown in green corresponds to the correct predictions over the total measured (column) or the total predicted on the corresponding class (row); finally, the red percentages are the complementary of the ones in green, and has the same meaning but for the incorrect forecasts. By doing a quick inspection of them, the following characteristics stand out:

- Both models (1 and 2) have an acceptable global accuracy: 72.71 and 73.95 %, respectively.

- The class 0 predictions are quite accurate in both models: 95.87 and 94.18 %.

- Forecasts of class 2 and, specially, class 1 days are the least precise.



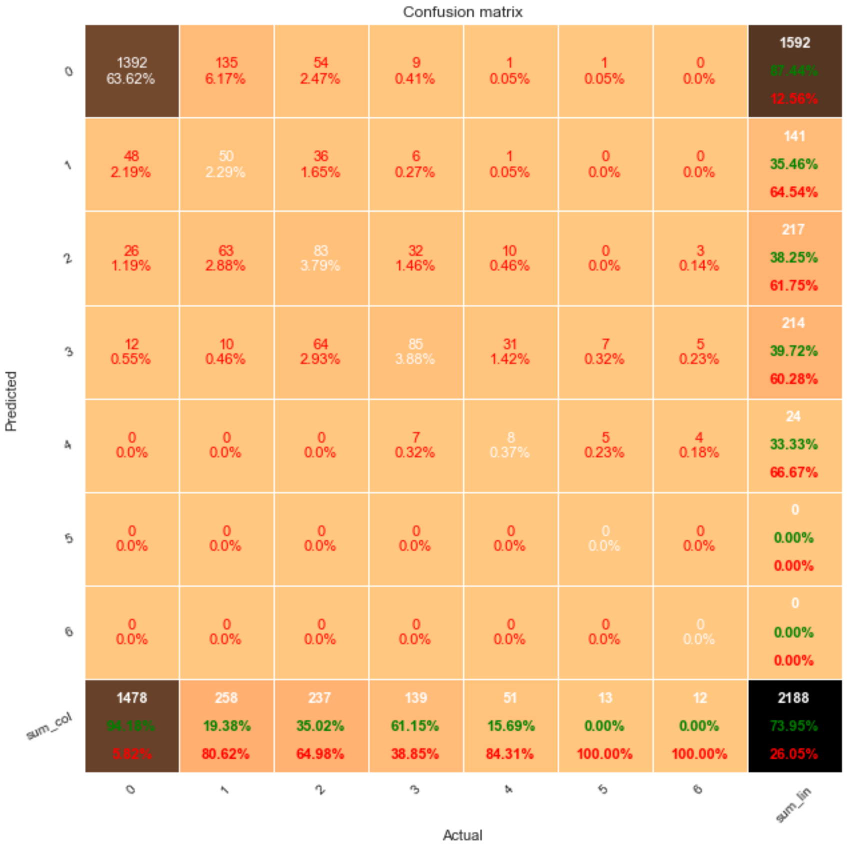
	0	1	2	3	4	5	6	sum_lin
0	1417 64.76%	158 7.22%	77 3.52%	13 0.59%	4 0.18%	0 0.0%	1 0.05%	1670 84.85% 15.15%
1	17 0.78%	12 0.55%	16 0.73%	3 0.14%	2 0.09%	0 0.0%	0 0.0%	50 24.00% 76.00%
2	32 1.46%	76 3.47%	79 3.61%	40 1.83%	8 0.37%	1 0.05%	3 0.14%	239 33.05% 66.95%
3	12 0.55%	12 0.55%	64 2.93%	78 3.56%	32 1.46%	6 0.27%	4 0.18%	208 37.50% 62.50%
4	0 0.0%	0 0.0%	1 0.05%	5 0.23%	5 0.23%	6 0.27%	4 0.18%	21 23.81% 76.19%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.00% 0.00%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.00% 0.00%
sum_col	1478 95.87% 4.13%	258 4.65% 95.35%	237 33.33% 66.67%	139 56.12% 43.88%	51 9.80% 90.20%	13 0.00% 100.00%	12 0.00% 100.00%	2188 72.71% 27.29%

Figure A.3: Analysis of model 1 performance on dust load class forecast on the test set.

- The class 3 forecasts accuracy is around 60 %, which is slightly lower than the global one.

- Classes 5 and 6, are not even predicted neither by model 1, nor by model 2, which could be due to the lack of training data from these classes.

Furthermore, despite the poor results obtained in some classes, the predictions of them are almost always around. This means that, say the measured value belongs to class 3, the model output may be between classes 2 and 4, in most cases. Additionally, and due to the huge amount of class 0 days, the model fails to generalize well on other less common classes. This problem of limited data of some classes could be overcomed by applying regularization techniques, which penalizes the algorithm when it makes certain incorrect predictions.



	0	1	2	3	4	5	6	sum_lin
0	1392 63.62%	135 6.17%	54 2.47%	9 0.41%	1 0.05%	1 0.05%	0 0.0%	1592 87.44% 12.56%
1	48 2.19%	50 2.29%	36 1.65%	6 0.27%	1 0.05%	0 0.0%	0 0.0%	141 35.46% 64.54%
2	26 1.19%	63 2.88%	83 3.79%	32 1.46%	10 0.46%	0 0.0%	3 0.14%	217 38.25% 61.75%
3	12 0.55%	10 0.46%	64 2.93%	85 3.88%	31 1.42%	7 0.32%	5 0.23%	214 39.72% 60.28%
4	0 0.0%	0 0.0%	0 0.0%	7 0.32%	8 0.37%	5 0.23%	4 0.18%	24 33.33% 66.67%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.00% 0.00%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.00% 0.00%
sum_col	1478 94.18% 5.82%	258 19.38% 80.62%	237 35.02% 64.98%	139 61.15% 38.85%	51 15.69% 84.31%	13 0.00% 100.00%	12 0.00% 100.00%	2188 73.95% 26.05%

Figure A.4: Review of model 2 performance on dust load class forecast on the test set.